Georefereren van teksten op basis van sociale media

Georeferencing Text Using Social Media

Olivier Van Laere

Promotoren: prof. dr. ir. B. Dhoedt, dr. S. Schockaert
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen: Computerwetenschappen

Vakgroep Informatietechnologie
Voorzitter: prof. dr. ir. D. De Zutter
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2012 - 2013

UNIVERSITEIT
GENT

ISBN 978-90-8578-586-6
NUR 980
Wettelijk depot: D/2013/10.500/19

Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur
Vakgroep Informatietechnologie

Promotoren: prof. dr. ir. Bart Dhoedt
dr. Steven Schockaert

Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur

Vakgroep Informatietechnologie
Gaston Crommenlaan 8 bus 201, B-9050 Gent, België

Tel.: +32 (0)9 33 14 900
Fax.: +32 (0)9 33 14 899

Cardiff University

School of Computer Science & Informatics
5 The Parade CF24 2AA Cardiff, UK

Tel: +44 (0)29 2087 4812
Fax: +44 (0)29 2087 4598

Proefschrift tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen:
Computerwetenschappen
Academiejaar 2012-2013

Dankwoord

28 december 2003, 23u34, nog net niet in het holst van de nacht, tijdens de kerstvakantie (in die tijd vooral gekend onder de noemer “blok”), kreeg ik volgende email:

> *Beste,*
>
> *Naar aanleiding van je thesispresentatie zou ik graag een afspraak maken in de inhaalweek om wat dieper in te gaan op de mogelijkheden tot doctoreren in onze groep. Ik denk dat het ook nuttig is wat meer uitleg te geven over een nieuw onderzoekscentrum rond breedbandtechnologie dat in 2004 wordt opgericht . . .*
>
> *Met vriendelijke groeten,*
> *Piet Demeester*

Het is een email die mijn levenswandel een richting instuurde tot op vandaag, nu meer dan 9 jaar later. Geprikkeld door de kans die geboden werd, ben ik ingegaan op de uitnodiging van prof. Piet Demeester en zat ik op maandag 5 januari 2004 om 10u tegenover hem in Urbis. Eigenlijk was er niet veel nodig om mij te overtuigen om te beginnen doctoreren na mijn studies informatica. Maar er was ook de lokroep van de “dark side”, het schakelprogramma naar burgerlijk ingenieur in de computerwetenschappen, een pad vol hindernissen die weinigen indertijd durfden in te slaan. De geschiedenis heeft uitgewezen dat de tweede optie toen de overhand heeft gehaald. Het schakelprogramma bracht me een ongelofelijk leerrijke ervaring door mijn Erasmusjaar (al eens electronica of signaalanalyse in het Spaans gevolgd zonder voorkennis? De snelcursus Spaans van UGent was duidelijk niet snel genoeg!) en een jaar vol uitdagingen in tweede proef met 6 vakken gespreid over 5 studiejaren en alweer een thesis. Die laatste werkte ik voor de tweede maal af bij IBCN. Het onbekende en het onzekere van een eindwerk maken is er in een dergelijke situatie van af en je wordt beter en zekerder in wat je doet, waardoor ik toen ook veel meer tijd in mijn eindwerk kon steken en helemaal gebeten was door het onderzoek.

In juli 2006 werd ik opgebeld door prof. Bart Dhoedt met de vraag of ik geen assistentenmandaat wou opnemen in plaats van te doctoreren via een beurs. Een handtekening plaatsen op een applicatieformulier in het rectoraat was alles wat nodig was en voor ik het wist was het gebeurd: je gaat door het leven als *AAP*.

Door de mogelijkheden die mij geboden werden om via mijn mandaat samen te werken met verschillende mensen in het kader van de vele onderwijsopdrachten,

en daarnaast de samenwerkingen met collega's (binnen en buiten IBCN), wens ik een groot aantal mensen te bedanken.

Eerst en vooral wens ik de mensen te bedanken die aan de basis liggen van het opstarten van mijn doctoraat. Jan Coppens, als immer gemotiveerde thesisbegeleider: was je niet geïnteresseerd in het sniffen van peer-to-peer netwerken via je thesisvoorstel, ik stond hier wellicht niet. Prof. Bart Dhoedt en prof. Filip De Turck, als toenmalige promotoren van mijn eindwerken. Uiteraard zou ik prof. Demeester, ten zeerste willen bedanken voor de kansen die hij voor mij en vele anderen heeft geschapen binnen de onderzoeksomgeving van IBCN. Zijn betrokkenheid met de moraal van de manschappen (zij het niet met taart, dan is het met IBCNdag) verdient zeker een vermelding. Dankjewel Piet. Daarnaast wens ik ook de Universiteit Gent te bedanken om mij een mandaat als assistent toe te kennen dat ik 6 jaar met plezier heb kunnen uitoefenen.

Ik ben de promotoren en dagelijkse begeleiders van dit doctoraatsproefschrift, prof. Bart Dhoedt en dr. Steven Schockaert, immens dankbaar voor hun rol in dit verhaal. Zonder hun steun was dit werk er gewoon niet gekomen.

Niemand van ons kan er om heen, de administratie. Zonder de deskundige begeleiding van het secretariaat van IBCN en de ondersteuning van de *finances* zou menig conferentieganger zelfs niet vertrekken. Ook de admins, ondertussen dusdanig gewijzigd in samenstelling ten opzichte van de start van mijn doctoraat, spelen een onmiskenbare rol in het goede verloop van al het onderzoek binnen de onderzoeksgroep.

Naast mijn onderzoek leverde het onderdeel *onderwijs* uit mijn taakomschrijving mij enorm veel voldoening. Velen onder U, zijn daar ofwel van dicht of van ver getuige van geweest via vakken zoals o.a. Informatica, Softwareontwikkeling of Ingenieursproject I en het begeleiden van thesissen.

Indien de "war stories" van mijn onderwijsactiviteiten nog niet bij U zijn geraakt, dan kan U volgens mij zeker terecht bij de vele collega's met wie ik samen in bureau 2.21 heb gezeten. De lijst 2.21'ers is ondertussen ook te lang geworden om nog accuraat te kunnen neerschrijven (het is wellicht mogelijk maar ik vergeet liever niemand). De *Friday drinks* lijken ondertussen wat uitgestorven, *Lachgas* speelt volgens mij geen volleybal meer, *Vet smaakt slecht* en de *Apero* lijken wel nooit te hebben bestaan, maar toch, voor zij die het hebben meegemaakt, dankjewel dat we dat allemaal samen hebben kunnen beleven.

Lang lang geleden, in de beginjaren van mijn doctoraat, lag de focus vooral op *context awareness*. Hierdoor kreeg ik de mogelijkheid om met Bart, Matthias en Samuel samen te werken op een Europees project. Naast de zeer goede herinneringen hierrond, blijft uiteraard het verhaal van de duiktabellen mij achtervolgen (zie Dankwoord Proefschrift Matthias Strobbe, 23 juni 2011). Helaas moet ik U meedelen dat door de volledige omschakeling naar het gebruik van decompressiecomputers de tabellen overbodig zijn geworden en de kennis rond de ware toedracht van hoe dat nu in elkaar zat, teloor is gegaan...

In een latere fase van mijn onderzoek, de fase die resulteerde in dit proefschrift, werd de focus gewijzigd naar *Geographic Information Retrieval* (GIR). Nieuw zijnde in de *IR* wereld werd ik deskundig bijgestaan door Steven Schockaert, eeu-

wige *partner in crime* tijdens onze studentenjaren maar ondertussen (en ik citeer een expert in het GIR domein tijdens een workshop) “disgustingly clever” onderzoeker aan Cardiff University. Dankjewel voor de suggestie om de richting van GIR in te slaan en het delen van je inzichten in de IR community. In het bijzonder bedankt voor de ontelbare discussies, brainstorms, traditiegetrouwe nachtelijke paper deadlines en zoveel meer.

A special word of thanks goes out to Martha Larson from TU Delft for providing me with the most motivational quote on doing research, ever:

> Wooooow, this is so cool... This is like... euh ... research going on, right here, right now!

With these words, she described me the night before the first MediaEval workshop in Pisa in 2010 while I was working like a madman into the night in the inner courtyard of the medieval convent. After the social event someone informed me (thank you for that Pavel) about a possible flaw in the methodology of my results which I was about to present 10 hours later, so I wanted to be sure that the results were clean, therefore spending hours during the (very short) night and morning verifying my remote data over a very unstable wifi connection (due to the thick convent walls). Of course, apart from Martha, I would like to thank the other organizers of the MediaEval benchmark and in particular the organizers of the Placing Task. Thank you Pavel, Vanessa, Pascal, Adam and Mohammad for bringing people together working in the same field. And since the 2011 edition, thank you Claudia for providing us with a grand research challenge after your presentation and the numerous discussions we had afterwards. Also for the follow-up work on this, Pascal and Sebastian, thank you for picking up on the opportunity to write a book chapter on geotagging.

With respect to my research stay at Cardiff University, I would like to express my gratitude towards prof. Christopher Jones. Chris, thank you for supporting my project and stay with your group and providing me with new challenges in the field of GIR. Obviously, neither dr. Jon or dr. Phil can be forgotten for the endless discussions (both related but mostly unrelated) to my research topics.

Gerelateerd aan UGent maar los van het feitelijke doctoraat is mijn loopbaan bij de Gentse Universitaire Duikclub (GUD). Individuele bedankingen voor de meest fantastische reizen van de afgelopen jaren, de feestjes, de duiken (weer of geen weer), de “zotte toeren”, de trainingen, de opleidingen en proeven,... zijn gewoon onbegonnen werk, hiervoor zou ik naar de volledige ledenlijst moeten verwijzen. In het bijzonder hoop ik hier onder U, de lezers, een talrijke delegatie van de GUD te mogen verwelkomen om eens een verdediging mee te maken die buiten de alomvertrouwde biotoop van de (mariene) biologie valt. Het komende anderhalf uur zal de rol tussen de vertrouwdheid met de vakterminologie tussen de bioloog en de informaticus even omgedraaid worden, maar ik wens u alvast toe dat u even goed kan volgen zoals ik reeds vissen en naaktslakken kan determineren.

6 jaar doctoreren stond voor mij ook gelijk aan 7 extra seizoenen reddingsdienst aan zee. Ondertussen heb ik nagerekend dat ik in de zomer van 2012 de

kaap van de 365 dagen van mijn leven tussen 10u30 en 18u36 naar onze Noordzee heb staan turen vanaf de Knokse en vooral Koksijdse stranden. Gedurende al die jaren heb ik mezelf blijven uitdagen om beter te worden en dingen bijgeleerd van mijn oversten, terwijl ik de laatste 7 jaar geprobeerd heb van die kennis als postoverste aan zoveel mogelijk van mijn redders door te geven. De lijst van reddingsploegen met wie ik gediend heb is te lang om hierin op te nemen, maar ik wens toch in het bijzonder mijn tweede redders van de laatste jaren te bedanken die de honneurs waarnamen als ik een dag verlof had (lees: een normale werkdag op IBCN beleefde). Een aantal van mijn beste vrienden heb ik daar leren kennen en ook al zien we elkaar nog weinig, bedankt dat jullie er steeds zijn, Lander, Dieter, Leen, Hanne, Karlien, Katrien en Kristel. Daarnaast wens ik in het bijzonder John "Zorro" VDB te bedanken om mij mateloos aan te steken met zijn duikverhalen, waardoor ik uiteindelijk zelf ben beginnen duiken.

Lotti en Annelies wens ik te bedanken voor de vele leuke gesprekken en inzichten die we uitgewisseld hebben bij het afronden van dit werk. Verder wens ik Elise en Pieter, Tineke en Steven, Marian en Philip, Tine en Steven, Elien en Karel, Lies en Youri, Marijn en Simon, Stefanie en Niels, Seraphine en Sam, Brigid en Bart en Tine te bedanken voor al de mooie dingen die we samen hebben meegemaakt in Gent en tot soms heel ver daarbuiten.

Een oprechte bedanking gaat uit naar mijn familie en in het bijzonder mijn ouders en broer voor de jarenlange steun, in goede en kwade dagen, het erin blijven geloven, voor de thuis die er altijd is en de kansen die werden mogelijk gemaakt, dankjewel.

De jaren aan zee hebben mij de mogelijkheid gegeven om een maand per jaar door te brengen bij mijn grootouders. Heeft mij (en hen) toegelaten een stuk mee te leven met 2 generaties verschil, met alle gevolgen van dien. Ik ben heel dankbaar voor al die tijd die ik met hen heb kunnen doorbrengen want die tijd is maar al te beperkt.

Nu sta ik hier weer, in die situatie waar ik al die jaren met zoveel passie voor heb gestaan: te turen naar een eindeloze zee van mogelijkheden waarvan je niet weet waar eerst gekeken. Je weet enkel dat je elk moment kan opgeroepen worden om de boot in te stappen en daarheen te gaan waar je ingezet kan worden. Zeg nu zelf, dat maakt het leven toch elke dag weer een beetje spannend?

"The greater the obstacle, the more glory in overcoming it."

– "Molière" (Jean-Baptiste Poquelin, 1622 - 1673), toneelschrijver.

Gent, 18 maart 2013
Olivier Van Laere

Table of Contents

List of Figures

List of Tables

List of Symbols and Acronyms

A

ACM	Association for Computing Machinery
AI	Artificial Intelligence
API	Application Programming Interface

C

CA	California
CT	Connecticut

D

DS	Dempster-Shafer

E

EM	Expectation-Maximization

F

FS	Feature Selection

G

GAC-GEO	Generic Agglomerative Clustering framework for geo-referenced datasets
GB	Gigabyte
GIR	Geographic Information Retrieval
GIS	Geographic Information System
GMM	Gaussian Mixture Model
GPS	Global Positioning System

I

IEEE	Institute of Electrical and Electronics Engineers
IG	Information Gain
IL	Illinois
IP	Internet Protocol
IR	Information Retrieval

K

KDE	Kernel Density Estimation
KL	Kullback-Leibler

L

LAU	Local Administrative Unit
LM	Language Model

M

MER	Median Error distance
MFU	Most frequently used
ML	Maximum Likelihood

N

NB	Naive Bayes
NER	Named Entity Recognition
NGA	National Geospatial-Intelligence Agency
NY	New York

P

PAM	Partitioning Around Medoids
PAV	Pool Adjacent Violators

R

ROC	Receiver Operating Characteristic

T

TBM	Transferable Belief Model
TF-IDF	Term frequency - inverse document frequency

U

UK	United Kingdom
URL	Uniform Resource Locator
US	United States of America
USGS	United States Geological Survey

V

VT	Vermont

W

WGS84	World Geodetic System
WOEID	Where On Earth ID

Samenvatting
– Summary in Dutch –

Webapplicaties en -diensten die gebruik maken van locatiegebaseerde informatie zijn vandaag de dag essentieel geworden in vele processen. Een verscheidenheid aan applicaties gebruikt dit soort gegevens, zoals routeplanning- en navigatiesoftware, zoekmachines die de resultaten aanpassen aan de locatie van de gebruiker, planningsoftware of locatiegebaseerde spellen die men kan spelen op een smartphone. In een poging om content, die te vinden is op het Web automatisch te voozien van geografische coördinaten, richten we ons in dit proefschrift op het voorspellen van de locatie waar Flickr foto's genomen zijn. Op Flickr zijn momenteel meer dan 200 miljoen foto's, geannoteerd met tags, te vinden die voorzien zijn van een geografische coördinaten. In dit proefschrift stellen we de hypothese voorop dat deze foto's een potentieel waardevolle bron van geografische informatie zijn. Bovendien stellen we dat, indien deze foto's effectief waardevolle en precieze geografische aanwijzingen bevatten en indien we dit kunnen modelleren, we deze modellen kunnen gebruiken om andere tekstuele inhoud geografisch te annoteren. Om locaties te voorspellen van tekst(fragmenten) wordt veelal beroep gedaan op het gebruik van een gazetteer. Een gazetteer kan in zijn meest eenvoudige vorm beschouwd worden als een lijst van geografische entiteiten waarvoor gedetailleerde informatie is voorzien, zoals bijvoorbeeld de naam, locatie en populatie van een bepaalde stad. Deze lijsten worden gebruikt om in een gegeven stuk tekst plaatsnamen te detecteren, waarna op basis van de gevonden entiteiten een locatie berekend wordt. Hoewel deze aanpak reeds heeft bewezen goede resultaten te leveren bij het voorspellen van de locatie van nieuwsartikelen of webpagina's, kan deze techniek niet zomaar gebruikt worden bij het localiseren van tekst(fragmenten) uit sociale media. Een aantal beperkingen treden op wanneer men dit zou proberen. Eerst en vooral is er het feit dat gazetteers, hoewel deze informatie bevatten over miljoenen entiteiten, beperkt zijn tot kennis die beschikbaar was bij het opstellen ervan. In sociale media verwijzen gebruikers vaak naar plaatsen aan de hand van namen die verschillen van de officiële, administratieve, naam. Ook is het mogelijk dat de namen die gebruikt worden in de volksmond wijzigen doorheen de tijd. Gazetteers zijn hier slechts in beperkte mate op voorzien en zullen dus niet in staat zijn deze geografische referenties te detecteren. Anderzijds is er het probleem dat, in het geval van het localiseren van Flickr foto's, er slechts weinig context-informatie beschikbaar is om de bedoelde betekenis van ambigue termen te achterhalen. Als derde en laatste probleem stellen we vast dat de dekking van gazetteers beperkt

is tot zaken die men terug kan vinden op stadsniveau: informatie over buurten in een stad, interessante plaatsen, lokale handelzaken, toeristische attracties, ... zijn veelal niet opgenomen in een gazetteer. Omwille van deze beperkingen is er nood aan een alternatieve manier om de locatie te achterhalen van Flickr foto's.

De afgelopen jaren is er reeds heel wat onderzoek verricht naar het localiseren van Flickr foto's. In het bijzonder is gebleken dat het gebruik van taalmodellen goed werkt voor dit doel. Omwille van de bemoedigende resultaten die men hiermee geboekt heeft, hebben wij het gebruik van taalmodellen overgenomen als vertrekpunt voor dit werk. Gedurende het onderzoek dat werd verricht in het kader van dit proefschrift zijn een aantal bijdragen geleverd in dit domein.

Ten eerste hebben we de verschillende componenten, die deel uitmaken van het proces om de locatie van een foto te voorspellen, verbeterd. Hierbij hebben we telkens een experimentele vergelijking gemaakt tussen de door ons voorgestelde technieken en de state-of-the-art methoden die beschreven zijn in de literatuur. Om deze evaluaties mogelijk te maken werd een schaalbaar raamwerk geïmplementeerd dat in staat is om taalmodellen te trainen die bestaan uit maximaal 20 000 klassen en die getraind worden op basis van tot 16 miljoen training foto's. Deze berekeningen kunnen worden uitgevoerd op één enkele computer die beschikt over 16 rekeneenheden en 16 GB geheugen. Via onze experimenten hebben we aangetoond dat het clusteren van de training data aan de hand van het k-medoids algoritme beter werkt, voor het localiseren van Flickr foto's, dan het clusteren aan de hand van alternatieve methoden die reeds toegepast zijn in de literatuur, zoals het gebruik van een vast geodetisch rooster of mean-shift clustering. Daarnaast hebben we aangetoond dat het belangrijk is om over een aangepast algoritme te beschikken dat in staat is om die tags te selecteren die relevant zijn voor het localiseren van Flickr foto's. Onze resultaten tonen aan dat standaard termselectiemethoden hier tekort schieten. Met betrekking tot het gebruik van taalmodellen hebben we twee nieuwe methoden voorgesteld om de prior probabiliteit te berekenen en hebben we aangetoond dat het belangrijk kan zijn om informatie die eigen is aan de gebruiker, zoals zijn thuislocatie, te gebruiken om betere voorspellingen te maken. De experimenten die we hebben uitgevoerd, maken gebruik van een standaard collectie testdocumenten die beschikbaar is om de experimenten te reproduceren. Daarnaast werden de resultaten vergeleken met alternatieve methoden beschreven in de state-of-the-art.

Verder hebben we nieuwe algoritmen beschreven die gebruikt worden voor termselectie. Bestaande term-selectie methoden zijn veelal gebaseerde op entropie (information gain), het aantal voorkomens van een tag, het gebruik door minstens een vooropgegeven aantal gebruikers of statistische afwijkingen (χ^2 en log-likelihood). Deze methoden negeren hierbij de spatiale informatie die beschikbaar is onder de vorm het voorkomen van een bepaalde tag op een bepaalde locatie. Om deze informatie op te nemen in het termselectieproces hebben we een aantal nieuwe algoritmen voorgesteld en geïmplementeerd gebruik makend van Kernel Density Estimation (KDE) en Ripley's K functie. Hiervoor hebben we twee spatiale smoothing algoritmen beschouwd. Een eerste methode maakt gebruik van de afwijking van de distributie van de voorkomens van een bepaalde term ten opzichte

van de algemene distributie van de termen. De tweede methode maakt gebruik van de entropiewaarde van de distributie van de voorkomens van de term om te meten in welke mate de voorkomens zich centreren rond bepaalde punten. Tevens voorzien we een use case voor het toepassen van aggresieve termselectie en tonen we aan dat de nieuw voorgestelde methoden in dergelijke situaties beter presteren dan de bestaande algoritmen.

Een aanzienlijke tekortkoming van de bestaande raamwerken voor het localiseren van tekst(fragmenten) is dat ze, onafhankelijk van de hoeveelheid informatie waarover ze beschikken, een exacte coördinaat zullen voorspellen voor de tekst. In situaties waarbij men de locatie moet voorspellen van een foto waaraan geen tags gekoppeld zijn komt dit eenvoudigweg neer op het gokken van een locatie op Aarde. Het is duidelijk dat in een dergelijke situatie een systeem geen zinvolle uitspraak kan doen over de locatie van de foto. Meer algemeen is het wenselijk om over een systeem te beschikken dat de granulariteit van zijn voorspellingen aanpast aan de beschikbare hoeveelheid informatie. Om dit te realizeren trainen we taalmodellen op verschillende niveau's en combineren we de informatie van deze niveau's aan de hand van Dempster en Shafer's *theory of evidence*. Hierdoor is ons systeem in staat om een voorspelling te maken op het fijnst mogelijk niveau waarvoor de beschikbare informatie dit verantwoordt. Gebruik makend van een dergelijke aanpak kunnen we in geval van twijfel tussen twee gebieden terugvallen op een minder gedetailleerd (generieker) gebied dat beide locaties omvat in plaats van verplicht te moeten kiezen voor een van beide alternatieven. Verschillende regels om de informatie tussen de niveau's te combineren werden experimenteel onderzocht alsook het gebruik van verschillende metrieken voor het instellen van de drempelwaarden die gebruikt worden om het meest gepaste niveau te bepalen.

Vervolgens hebben we aangetoond dat onze taalmodellen, die opgebouwd worden aan de hand van informatie van Flickr foto's, kunnen gebruikt worden om de locatie te voorspellen van documenten verschillend van Flickr foto's. De meeste onderzoeken die beschreven zijn in de literatuur beperken de evaluatie van hun taalmodellen tot hetzelfde type document dat werd gebruikt om de modellen te construeren. We hebben het potentieel geëvalueerd van taalmodellen die opgebouwd worden aan de hand van data van Flickr, Twitter en Wikipedia die gebruikt worden om de locatie te achterhalen van Wikipedia documenten, wetende dat deze documenten grondig verschillen in structuur van de tags van Flickr foto's of Twitter berichten. In ons werk hebben we een manier voorgesteld om te interpoleren tussen verschillende taalmodellen. Onze experimentele resultaten tonen aan dat taalmodellen getraind op basis van Flickr data significant beter presteren bij het localiseren van Wikipedia documenten dan een gazetteer. In het kader van de grootschalige evaluatie die we hebben uitgevoerd voor deze laatste bijdrage werd de schaalbaarheid van ons raamwerk verbeterd. Hiermee zijn we in staat om taalmodellen te construeren op basis van verschillende informatiebronnen en classificatie uit te voeren aan de hand van meer dan 125 000 klassen in combinatie met 1.5 miljoen features. Het systeem kan zijn modellen probleemloos opbouwen aan de hand 64 miljoen training documenten en dit alles op één enkel systeem met 16 rekeneenheden en 16 GB geheugen.

Het raamwerk om tekstuele inhoud te georefereren, dat werd voorgesteld in dit proefschrift, werd gebruikt om deel te nemen aan de 2010, 2011 en 2012 edities van de MediaEval Placing Task benchmark. Hiervoor werd onze inzending in 2010 bekroond met de *"quantum leap award"* omwille van het feit dat onze resultaten substantieel beter waren dan de inzendingen van de overige deelnemers. Ons raamwerk laat toe om, afhankelijk van de evaluatiedocumenten, tot meer dan 40% van de documenten binnen 1 km van de correcte locatie te voorspellen. Dit werd experimenteel aangetoond aan de hand van verschillende types van documenten, waaronder Flickr foto's, Twitter berichten, onderschriften van Getty Images foto's en Wikipedia documenten. In een aantal van onze evaluaties werd een vergelijking gemaakt met Yahoo! Placemaker als basis voor de vergelijking met een gazetteergebaseerde methode. In elk van deze vergelijkingen werden de resultaten van Yahoo! Placemaker substantieel overtroffen door ons systeem. Dit bevestigt dan ook onze hypothese dat er waardevolle geografische aanwijzingen gevonden kunnen worden in de tags van Flickr foto's die gebruikt kunnen worden om taalmodellen te trainen. Ten slotte hebben we experimenteel aangetoond dat, om de locaties te voorspellen van Wikipedia documenten, het gebruik van een taalmodel dat getraind is aan de hand van Flickr data beter werkt dan een model dat werd getraind op Wikipedia data zelf.

Summary

Web applications and services that use location-based information have become central in many of today's workflows. A variety of applications that use or consume this information exist today: mapping and navigation applications, search engines that optimize their results for the user's location, planning tools or location-based games on smartphones to name just a few. Geographic information has become big business, and the need for this type of information grows by the day. In an effort to automatically annotate content with geographical coordinates, in this dissertation, we focus on the task of georeferencing Flickr photos. Given the fact that there are over 200 million geotagged photos available on Flickr described by tags, we believe this data to be a potentially valuable source of geographical information. If sufficiently rich and accurate information is indeed (implicitly) contained in Flickr data, and if this information can be extracted, it can be exploited to automatically georeference other textual content that has no spatial grounding.

When it comes to georeferencing text, the use of a gazetteer, which is in essence a list containing geographical information about entities, is widely adopted. Gazetteers are used to scan the text for occurrences of place names (toponyms). This approach has proven to work well for georeferencing news articles or webpages. However, a number of issues arise when this method is applied in the context of social media. First, although gazetteers contain geographical information about millions of entities, these are limited to places that were known at the time the gazetteer was created. In social media, people tend to refer to places by means of names (known as vernacular place names) that differ from the actual, administrative, place names. Also, the vernacular place names that people use can change over time. A gazetteer generally does not contain this kind of information and especially does not adapt to changes over time. Secondly, in the case of georeferencing Flickr photos, limited context information is available to resolve any potential ambiguities between two toponyms. Thirdly, the coverage of gazetteers is mostly limited to a city level: information from specific neighbourhoods, landmarks, local businesses, etc. are not commonly found in it. For these reasons, an alternative approach is called for.

Over the past few years, researchers have started looking into georeferencing Flickr photos. It has been shown that using language models is well-suited for this task. In view of these encouraging results, language models are adopted as a starting point in this work. During the research for this PhD, we have made a number of contributions in this field.

First, we improved several of the components of the georeferencing process,

each time experimentally comparing our proposal against the state-of-the-art methods from the literature. To this end, we implemented a scalable georeferencing framework capable of estimating language models for up to 20 000 classes using up to 16 million training photos, which can be processed on a single 16-core computer with 16 GB of memory. We experimentally found that the k-medoids clustering algorithm performs better at this task that using a fixed geodesic grid or mean-shift clustering, methods that are generally applied in related work. We also showed the importance of feature selection methods tailored to this task. With respect to the language models, we proposed two new methods for estimating the prior probability and demonstrated the importance of including user specific information such as information from his home location. All evaluations are carried out using a standard benchmark test set while the results were compared to state-of-the-art frameworks.

Secondly, we introduced new algorithms for feature selection. Current methods are generally based on entropy based scores (information gain), the number of tag occurrences, the usage by a minimum number of different users or statistical deviations (such as χ^2). These methods, however, all ignore the spatial information that is confined in the relation between a tag occurrence and its corresponding location. Therefore, we proposed and implemented a number of new feature selection methods based on Kernel Density Estimation (KDE) and Ripley's K function that include the spatial component of the different occurrences of a tag. We studied two spatial smoothing algorithms. The first method uses the divergence between the distribution of the occurrences of a single tag and the overall distribution. A second method uses the entropy value of the distribution of the occurrences of a single tag to measure the extent to which they occur in clusters around certain points. We provide a use case that calls for aggressive feature selection, and show that our methods outperforms standard feature selection methods in such situations.

Thirdly, a shortcoming of the current georeferencing approaches is that, regardless of the amount of information available, most systems will return a precise location for a given textual description. This includes situations in which a system has no meaningful suggestion and is thus basically guessing (e.g. when estimating the location of a photo without any tags). In such a case, it would be better if a georeferencing system refrained from making any prediction at all. To this end, we proposed an evidence-based approach to multilevel georeferencing. To realize this, we train language models at different levels of granularity and combine the information contained in these levels by using Dempster and Shafer's theory of evidence. In this way, we are able to provide a location estimate for Flickr photos at the finest level of granularity that is warranted by the available evidence. In this approach, if the system is not able to disambiguate between two locations, for instance, it returns an estimate at a coarser level of granularity that contains both of the possible locations rather than guessing between the two. We evaluated the use of different combination rules to aggregate evidence from the different scales. In addition, we experimented with the use of different threshold criteria to determine the appropriate scale on which the location estimate should be done.

Fourthly, we demonstrated that our Flickr models can be used to georeference other content than Flickr photos. Most use cases in literature restrict themselves to georeferencing data from the same kind of data as the one used for training the language models. We evaluate the potential of language models trained using data from Flickr, Twitter and Wikipedia at the task of georeferencing Wikipedia documents, which are structured quite differently from the tags from Flickr photos. In our work, we proposed a way of interpolating between different language models. Our experimental results show that language models trained from Flickr significantly outperform gazetteer based methods at the task of georeferencing Wikipedia documents. To evaluate our methods on a large scale, we improved the scalability of our georeferencing framework to cope with generic data sources, over 125 000 classes for classification, 1.5 million features and over 64 million training documents on a single 16-core computer.

The georeferencing framework outlined in this PhD dissertation has been evaluated in the 2010, 2011 and 2012 editions of the MediaEval Placing Task benchmark, for which we received the *"quantum leap award"* in 2010 with a submission that substantially outperformed all other submissions to the task. Our framework allows to accurately assign a geographical coordinate to different sources of textual documents. This has experimentally been verified using tagged Flickr photos, Twitter messages, Getty Images photo captions and Wikipedia documents. Using our framework we can locate over 40% (depending on the test set) of the test documents within 1 km of their true location. In some of our evaluations, we have included a comparison to Yahoo! Placemaker as a baseline gazetteer approach, which was significantly outperformed in all of our tests. This confirms our hypothesis that there is valuable geographical information contained in Flickr tags that can be used to train language models. To conclude, we demonstrated in our experimental results that a language model trained only using Flickr data outperforms a model trained on Wikipedia data at the task of georeferencing Wikipedia documents itself.

1 Introduction

The only way to be truly satisfied is to do what you believe is great work. And the only way to do great work is to love what you do. If you haven't found it yet, keep looking. Don't settle. As with all matters of the heart, you'll know when you find it.

– Steven Paul Jobs (1955 - 2011)

1.1 Context

The time when only a small number of users actually contributes content to the Web is over. Social media applications such as Facebook, Flickr, Twitter, Foursquare or LinkedIn facilitate networking and sharing of information between hundreds of millions people. These applications have managed to lower the bar to enter the digital world in such a way that almost anyone can start using social media services with almost no prior knowledge. People are encouraged to freely share information about their thoughts, actions and whereabouts in many different forms. Most of the social media applications provide some (licensed) access to their (anonymized) data by means of an API. Although foundations for the protection of the privacy of Internet users raise valid reservations regarding this evolution, actual users seem to be little restrained in taking part in this information sharing society, as illustrated by the following numbers:

- 250 million photos are uploaded and over 2 billion posts are liked on Facebook every day. [1]

- 8.1 billion photos are shared on Flickr in total. [2]
- 175 million tweets are shared every day on Twitter. [3]
- 187 million professionals are actively networking using LinkedIn. [4]
- 20 million people use the location sharing service Foursquare. [5]

What makes this evolution even more interesting is that part of this data are now geotagged, i.e. they are associated with a geographical location. Geotags generally consist of geographical coordinates such as a (latitude, longitude) pair in WGS84 but this can also be a *where on earth ID* (WOEID) [6] or a textual reference to a certain place name. Due to the increasing popularity of smartphones and devices with integrated GPS systems, the amount of geotagged content will strongly increase in the future. Application developers successfully started exploiting this type of information, as can be witnessed by the recent trend of location-aware applications. For example, in the field of search and recommendation engines, location information has become essential to refine the search scope.

Location-based services have become big business, with an expected revenue of 10.3 billion dollar in 2015, compared to 2.8 billion dollar in revenues in 2010 [7], and the need for georeferenced content grows. This need is exactly what this dissertation addresses: we seek ways of automatically geotagging content that has no spatial anchoring. To this end, we specifically exploit geotagged textual data available in social media. In what follows, we present an overview of the initial approaches to georeferencing textual content (Section 1.2). Section 1.3 outlines the limitations of those methods when applied in the context of social media and describes the current evolution of using language models for this task. The main contributions of our research to the state-of-the-art are described in Section 1.4. Subsequently, an outline of this dissertation is presented in Section 1.5. We conclude this chapter with a list of publications that are the result of this PhD research in Section 1.6.

1.2 Geographic Information Retrieval (GIR)

A lot of webpages refer to geographic entities in one way or another. In the case of a homepage of a commercial company, there should be some information about the company's address, while on a blog or in a news article, there might be references to place names.

The most straightforward approach for finding geographical entities is scanning the text for toponyms (i.e. place names), often referred to as *toponym resolution*. This approach supposes that one has knowledge that allows identifying terms as toponyms. Luckily, there are tools that contain this kind of information: a

gazetteer is in essence a list or database containing millions of geographical entities along with details such as alternative names, population (in case of populated places) and geographical coordinates. Just to give an idea, Yahoo! GeoPlanet [1] and Geonames [2] contain about 6 million and over 8 million entities respectively.

Given access to a comprehensive gazetteer, a natural way to discover the geographic scope of a webpage consists of identifying place names and looking up their coordinates in the gazetteer. In practice, however, this method is complicated by the fact that many place names are highly ambiguous. A well known-example is "Springfield": at least 58 populated places with this name are listed in Geonames. Georeferencing methods using a gazetteer have to cope with this. In [8], gazetteers are used to estimate the locations of toponyms mentioned in text and a geographical focus is determined for each page. During this process, two different types of ambiguities are described: geo/geo, e.g. the previous example of "Springfield", or geo/non-geo, such as "Turkey" or "Bath", which are also common nouns in English. Heuristic strategies to resolve both type of ambiguities are proposed in [8].

A complementary approach to extracting geographic information from text is by resolving comma groups [9]. The general idea, similar to scanning for zip codes and certain lexical constructions, is that for some countries or regions, spatial clues come in comma-separated constructions like *Houston, Texas*. From this, an explicit indication is given that there is a relationship between the terms. If one is able to resolve one of the terms of the comma group, the spatial extent might be resolved more accurately or might help in disambiguating terms.

Another alternative is to use Named Entity Recognition (NER) algorithms to identify which named entities refer to places, and which refer to people, organisations, or other entities. A NER classifier is trained using examples of sequences of text, for example: "A woman *from New Zealand* and a German man have won the 35th Annual Empire State Building Run-Up *in New York*." In this example, the word "in" followed by "New York", suggests that "New York" refers to a place, similar to the sequence "from New Zealand". When it comes to detecting an entity that exists both as a place name and a another entity (for example "Downing Street" refers to a political entity in a sentence such as *"Downing Street insists Leveson saw all communication between Cameron and Brooks"*), NER can be used to detect this, although it is a non-trivial problem.

1.3 Problem statement

As the need for location based information grows, new ways are investigated of automatically generating georeferences. With respect to geotagging textual con-

[1] http://developer.yahoo.com/geo/geoplanet/
[2] http://www.geonames.org/

tent, the use of a gazetteer seems a straightforward way of finding geographical references. Gazetteers however exhibit two major limitations:

1. The data contained in a gazetteer is mostly manually selected and reviewed by domain experts and thus tends to be of high quality. However, manual moderation is a time-consuming and cumbersome task which makes gazetteers hard to maintain. This leads to an inherently limited and possibly outdated coverage of the data contained in gazetteers.

2. Another limitation of gazetteer based methods is that people often use vernacular names to describe places, names that tend to be missing in gazetteers. For instance, "The Big Apple" is used when referring to "New York City".

In the context of georeferencing Flickr photos, where only a limited number of tags are available per photo, using a gazeteer to find a precise location will not work well. If two place names occur, only little context information is available for a gazetteer to be able to disambiguate between the places. Second, if a photo is tagged with the name of geographical entities at a sub-city scale, such as names of neighbourhoods or local events, the coverage of the gazetteer will usually be too limited to include information about them. Also, if a toponym is misspelled we will most likely not be able to lookup any relevant information in a gazetteer. In order to georeference content from social media, a new approach is called for.

The advent of social media enables users to share experiences by means of photos, videos, comments,... along with information about their whereabouts. The main hypothesis of this PhD research is then the following: we assume that the geotagged textual data, originating from these social media sources, allow us to train statistical models that link terms to geographical locations. Subsequently, we assume that these models are potentially sufficiently accurate to automatically assign coordinates to other resources on the web, enabling applications such as GIR.

The correlations between objects and location used throughout this dissertation are tuples of the form $< x, y, z, t, U >$, in the sense of [10–12], where U represents a 'thing' which was present at location (x, y, z) at time t. In the aforementioned works U is referred to by some web object; e.g. a Flickr photo or Twitter post refers to the presence of a user at a particular location. Furthermore, in this work, the time component is ignored while a location x, y, z is generally represented using the WGS84 [13] latitude and longitude coordinates. Alternative definitions of "place", to the one used in this work, are described in literature, for instance by means of affordances [14] or to include the inherent vagueness of place boundaries [15].

Over the past few years, researchers have started looking into georeferencing Flickr photos. They have shown that using language models is the state-of-the-art for this task, and consequently we adopted this method as a starting point for the research. The tendency for particular tags to be clustered spatially, and hence

to provide strong evidence for the place at which a photo was taken, was studied in [16, 17]. Most existing georeferencing methods use a form of clustering in one way or another to convert the task to a classification problem. For instance, in [18] locations of unseen resources are determined using mean shift clustering, a non-parametric clustering technique from the field of image segmentation. To assign locations to new images, both visual (keypoints) and textual (tags) features have been used in [18]. Experiments were carried out on a sample of over 30 million images, using both Bayesian classifiers and linear support vector machines, with slightly better results for the latter. In [19], the idea is suggested that whenever a classifier determines a certain area where an image was most likely taken, the surrounding areas could be considered as well to improve the results. Their starting point is that typically not only the correct area will receive a high probability, but also the areas surrounding the correct area.

The interest of the research community for this problem resulted in the Placing Task, an evaluation framework focussing on this problem of georeferencing Flickr videos [20], as part of the MedialEval benchmarking initiative[3]. We actively participated in the 2010, 2011 and 2012 editions of this task.

1.4 Main research contributions

During the research for this PhD, we have made a number of contributions in this field. First, it is important to analyze and understand the contribution of the various individual components involved in our and others' language modelling (LM) approaches to georeferencing. To this end, we implemented a scalable, georeferencing framework capable of constructing language models for up to 20 000 classes using up to 16 million training photos, that can be processed on a single multi-core computer with 16 GB of memory. Using this framework:

- We evaluated three often used clustering algorithms for the task of constructing the set of classes (areas): k-medoids, grid-based and mean-shift clustering.

- We carried out a quantitative and qualitative evaluation of six feature selection algorithms: χ^2, maximum-χ^2, log-likelihood, information gain, most frequently used and geographical spread.

- Four different methods for estimating the prior probability for the language models were evaluated: a maximum likelihood prior, a uniform prior, a prior based on the home location of the user (in case of for example Flickr photos) and a prior based on Gaussian mixture models (GMM).

[3] http://www.multimediaeval.org/

- The influence of using more training data (up to 10 million training items) on the performance of the language models was analyzed.

For our evaluation, we used an available, standard benchmark set. These contributions are described in detail in [21].

Secondly, there is a need for new feature selection algorithms that take the spatial nature of the problem into account, which current term selection techniques ignored. Terms are generally selected based on entropy based scores (information gain), their number of occurrences or the usage by a different number of users. Methods exploiting statistical deviations, such as χ^2, are better suited at selecting relevant tags from the set of tags associated to the Flickr photos, but still ignore the spatial information that is confined in the relation between a tag occurrence and the corresponding location. During our research:

- We proposed, implemented and compared a number of feature selection methods based on Kernel Density Estimation (KDE) and Ripley's K function that include the spatial component of the different tag occurrences.
- We studied two spatial smoothing techniques:
  - by using the divergence between the distribution of the occurrences of a single tag and the overall distribution.
  - by using the entropy value of the distribution of the occurrences of a single tag to measure the extent to which they occur in clusters around certain points.

The results from these contributions are described [22].

Thirdly, an adaptive way of georeferencing is needed based on the evidence available to support decisions. Indeed, a shortcoming of the current georeferencing approaches is that, regardless the amount of information available, most systems will return a precise location for a given textual description, which might be significantly worse than not returning a location at all. For this reason, we argue for an approach that provides a location estimate within certain confidence thresholds. Such an approach leads to a number of new applications:

- The result of the georeferencing process can be improved: by taking evidence into account from different levels of granularity, better disambiguation is possible.
- Instead of returning specific coordinates, a region that has been determined based on the available knowledge, can be returned. If only little evidence is available, a larger area is returned to model the uncertainty.
- A heat map can be created that indicates the places that are more, or less, likely for a given photo.

Addressing this problem led to the following contributions:

- An evidence-based approach to multilevel georeferencing.
- The evaluation of different combination rules: Dempster-Shafer, Yager and Dubois-Prade.
- The evaluation of different threshold criteria: plausibility, belief and pignistic probability.

These contributions are published in [23].

Fourthly, as the main hypothesis of this dissertation states, we believe that once a model is trained using a given data set of geographical information, it can be used to geotag other sources of textual content. However, most use cases in literature restrict themselves to georeferencing data from the same source of data as the one used for training the language models. To demonstrate the generalizability of our approach:

- We present an evaluation of georeferencing Wikipedia pages by means of language models trained using data from Flickr, Twitter or Wikipedia, individually or combined, on a standard test set published in literature.
- We proposed a way of combining language models trained using different sources of data.
- We provide experimental results that show that language models trained using Flickr significantly outperform gazetteer based methods at georeferencing Wikipedia documents.

In order be able to evaluate our methods on a large scale, we improved the scalability of our georeferencing framework to cope with generic data sources, over 125 000 classes for classification, 1.5 million features and over 64 million training documents on a single multi-core computer. The Wikipedia dataset used for our evaluation is published online[4] while the research itself is described in [24].

A fifth and final aspect to which we contributed, is the use of user specific information. As social media are inherently coupled to user accounts, prior knowledge of a certain user might provide clues in hard cases that need disambiguation. For instance, the knowledge about the previous whereabouts of a Flickr user might eliminate many locations in the world when estimating the location of one of his newly uploaded photos. Also, if a user exhibits a certain specific tagging behaviour, this could be exploited as well. In the approach we outline in this dissertation, we investigate how we can use these two specific pieces of user information:

[4]Our pre-processing script, along with the original XML and processed test set are made available online at https://github.com/ovlaere/georeferencing_wikipedia

- We proposed a method to include information from the home location of the user in the prior probability of a language model. If this information is available, it significantlty improves the results.
- We evaluated the results of georeferencing Flickr photos using two different datasets: one in which a user only occurs in either the training or the test set, and a second in which a user occurs in both. This evaluation yielded interesting insights in the performance gain that can be achieved by exploiting user specific tagging behaviour.

The results of this analysis are published in [25] and [21].

The georeferencing framework proposed in this PhD research has been evaluated in the 2010, 2011 and 2012 editions of the MediaEval Placing Task benchmark [26–28]. In 2010 our results substantially outperformed all other submissions to the task, for which we were given the *"quantum leap award"*. In 2011, we were awarded a *"distinctive mention"* for our submission to the fifth run *"in which we embraced the spirit of the task and went above and beyond what was asked of us"*.

1.5 Outline of this dissertation

This dissertation is composed as a comprehensive set of publications written within the scope of this PhD. The selected publications provide an integral and consistent overview of the work performed. The different research contributions are detailed in Section 1.4 and the complete list of publications that resulted from this work follows in Section 1.6. Within this section we give an overview of the remainder of this dissertation and explain how the different chapters are linked together. Table 1.1 shows the research contributions that were targeted per chapter.

Table 1.1: An overview of the contributions per chapter in this dissertation.

	Ch.3	Ch.4	Ch.5	Ch.6	Ch.7
Analysis of language modeling for georeferencing	•				
Exploiting user specific information	•	•			
Feature selection	•	•	•		
Multilevel georeferencing		•		•	
Georeferencing using different sources of data					•

Chapter 2 provides an overview of related work in the field of georeferencing textual resources from social media in general and georeferencing Flickr photos in particular. In Chapter 3, we present a thorough analysis of the performance of

using language models at the task of georeferencing Flickr videos. A georeferencing framework that is the result from participating in the 2010 and 2011 editions of MediaEval's Placing Task is introduced. For each of the individual components (i.e. language modeling, feature selection, clustering, ...) different methods are implemented. The extensive experimental results allow us to analyze why certain methods work well on this task and show that a median error of just over 1 kilometer can be achieved on a standard benchmark test set. The analysis in this chapter shows us, among others, two important things. First, the results indicate that the choice of a good feature selection algorithm significantly influences the performance. Second, it does not make sense to assign a specific coordinate to an entity that has a large geographical scope, for instance a country like Canada. A granular approach imposes itself in this situation. These two findings will be investigated in detail in two later chapters (Chapter 4 and Chapter 6).

A common argument in favor of feature selection is the reduction of the computational complexity of the problem. With the current developments in the field of computational resources, this argument is nowadays less important. However, in Chapter 4 we provide experimental results that show a clear need for feature selection when it comes to georeferencing. Indeed, removing noise from the available features substantially outperforms the results of using all features. A second effect studied in this chapter is the gain that can be obtained by exploiting user specific tagging behaviour. Chapter 4 briefly investigates the effect of exploiting knowledge of user specific tags on the performance of georeferencing unseen photos from this user. Furthermore, results are presented that demonstrate that feature selection is necessary to remove noisy terms from the overall vocabulary, as the optimal results for the experiments carried out in this chapter are obtained with fewer than all features. Furthermore, the results in this chapter show the effect of the number of tags in relation to the error made in georeferencing photos, which clearly indicate that as soon as 4 tags are present, good location estimations can be made.

Current approaches to georeferencing rely on term selection techniques, such as TF-IDF, χ^2 or Information Gain (IG), which ignore the spatial nature of the domain. In Chapter 5, we implement the idea of spatial smoothing of term occurrences by using Kernel Density Estimation (KDE) to model each term as a two-dimensional probability distribution over the surface of the Earth. Experimental results are provided which demonstrate a considerable improvement over the standard term selection methods. As an alternative, a feature selection algorithm is presented that uses Ripley's K function. This latter approach yields results that are comparable to those of the KDE-based method but at a significantly reduced computational cost.

Next, Chapter 6 provides an overview of a multilevel approach to georeferencing Flickr photos. We present an adaptive technique that assigns locations to

photos at the right level of granularity, or, in some cases, even refrains from making any estimations regarding location at all. To this end, we consider the idea of training language models at different levels of granularity, and combining the evidence provided by these language models using Dempster and Shafer's theory of evidence. We provide experimental results which clearly confirm that the increased spatial awareness that is thus gained allows us to make better informed decisions, and moreover increases the overall accuracy of the individual language models.

Subsequently, in Chapter 7 we evaluate the performance of using language models obtained by training on data from Wikipedia, Flickr and Twitter, individually and in a combined way, at the task of georeferencing Wikipedia documents. In our experimental evaluation, we demonstrate that our language models, trained using data from social media, substantially outperform both classical gazetteer-based methods and language modelling approaches trained on Wikipedia alone. This supports the hypothesis that social media are an important source of geographic information, which is valuable beyond the scope of individual applications.

Finally, Chapter 8 provides the overall conclusions and future perspectives.

Each of the chapters 2 to 7 are based on a paper which has been published or is currently under review. While dr. Steven Schockaert and prof. Bart Dhoedt have contributed to these papers in their role as supervisors, the work which is described in them has been integrally carried out by myself. The only exception is the implementation of the KDE-based score methods and Figures 5.1, 5.2 and 5.3. The figures, implementation and the resulting score values that were used as the input for the experiments have been provided by dr. Jonathan Quinn from Cardiff University, UK. For the work described in Chapter 7 I have also been supported by valuable feedback from prof. Chris Jones and dr. Vlad Tanasescu during my research stay with them at Cardiff University.

1.6 Publications

The research results obtained during this PhD research have been published in scientific journals and presented at a series of international conferences. The following list provides an overview of the publications during my PhD research. At the beginning of this research, my focus was on context-awareness. This led to a number of publications (A1 articles [1-4], conference articles [1-8][12], national conference articles [1-2]) that are included in this overview related to this topic. As the core topic of this dissertation is my research in geographic information retrieval, the aforementioned papers are no longer referenced in the remainder of this thesis.

1.6.1 A1: Publications in international journals (listed in the Science Citation Index [5])

[1] M. Strobbe, **O. Van Laere**, S. Dauwe, B. Dhoedt, F. De Turck, P. Demeester, C. van Nimwegen, J. Vanattenhoven, *"Interest Based Selection of User Generated Content for Rich Communication Services"*. Journal of Network and Computer Applications 33(2):84-97, March 2010.

[2] M. Strobbe, **O. Van Laere**, B. Dhoedt, F. De Turck, P. Demeester, *"Hybrid Reasoning Technique for Improving Context-Aware Applications"*. Knowledge and Information Systems 31(3):581-616, June 2012.

[3] M. Strobbe, **O. Van Laere**, S. Dauwe, B. Dhoedt, F. De Turck, P. Demeester, K. Luyten, *"Integrating Location and Context Information for Novel Personalised Applications"*. IEEE Pervasive Computing 11(2):64-73, April-June 2012.

[4] M. Strobbe, **O. Van Laere**, B. Dhoedt, F. De Turck, P. Demeester, *"Automatic Generation of User Profiles Based on Bookmark Clustering"*. Submitted to Word Wide Web: Internet and Web Information Systems.

[5] **O. Van Laere**, S. Schockaert, B. Dhoedt, *"Georeferencing Flickr photos using language models at different levels of granularity: an evidence based approach"*. Journal of Web Semantics 16(1):17-31, November 2012.

[6] **O. Van Laere**, S. Schockaert, V. Tanasescu, B. Dhoedt, C. Jones, *"Georeferencing Wikipedia documents using data from social media sources"*. Submitted to ACM Transactions on Information Systems, ACM, November 2012.

[7] **O. Van Laere**, S. Schockaert, B. Dhoedt, *"Georeferencing Flickr resources based on textual meta-data."*. Accepted for publication in Information Sciences, Elsevier, February 2013.

[8] **O. Van Laere**, J. Quinn, S. Schockaert, B. Dhoedt, *"Spatially-aware Term Selection for Flickr Photo Geotagging"*. Accepted for publication in IEEE Transactions on Knowledge and Data Engineering, IEEE, February 2013.

1.6.2 B1: Book chapters

[1] P. Kelm, V. Murdock, S. Schmiedeke, S. Schockaert, P. Serdyukov and **O. Van Laere**, *"Geotagging in social networks"*. Chapter in the book

[5] The publications listed are recognized as 'A1 publications', according to the following definition used by Ghent University: A1 publications are articles listed in the Science Citation Index, the Social Science Citation Index or the Arts and Humanities Citation Index of the ISI Web of Science, restricted to contributions listed as article, review, letter, note or proceedings paper.

"Social Media Retrieval" (N. Ramzan, R. van Zwol, J. Lee, K. Clüver, X. Hua, eds.), Computer Communications and Networks series, Springer, ISBN: 978-1-4471-4554-7, November 2012.

1.6.3 Publications in other international conferences

[1] M. Strobbe, J. Hollez, G. De Jans, **O. Van Laere**, J. Nelis, F. De Turck, B. Dhoedt, P. Demeester, N. Janssens, T. Pollet, *"Design of CASP: an Open Enabling Platform for Context Aware Office and City Services"*. Proceedings of the 4th International Workshop on Managing Ubiquitous Communications and Services (MUCS 2007), Munich, Germany, May 2007.

[2] M. Strobbe, **O. Van Laere**, S. Dauwe, F. De Turck, B. Dhoedt, P. Demeester, *"Efficient Management of User Interests for Personalized Communication Services"*. Proceedings of the 5th International IEEE Workshop on Management of Ubiquitous Communications and Services (MUCS 2008), Salvador, Brazil, April 2008.

[3] **O. Van Laere**, M. Strobbe, S. Dauwe, B. Dhoedt, F. De Turck, P. Demeester, O. Verde, F. Hlsken, *"Interest Based Selection of User Generated Content for Rich Multimedia Services"*. Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008), Klagenfurt, Austria, May 2008.

[4] **O. Van Laere**, M. Strobbe, P. Leroux, B. Dhoedt, F. De Turck, P. Demeester, *"Enabling Platform for Mobile Content Generation Based on 2D Barcodes"*. Proceedings of ICOMP 2008, the 2008 International Conference on Internet Computing: 209-214, Las Vegas, Nevada, USA, July 2008.

[5] M. Strobbe, **O. Van Laere**, B. Volckaert, F. De Turck, B. Dhoedt, P. Demeester, *"Context Based Selection of User Generated Content"*. Proceedings of SWWS 2008, the 2008 International Conference on Semantic Web & Web Services: 100-106, Las Vegas, Nevada, USA, July 2008.

[6] M. Strobbe, **O. Van Laere**, B. Bogaerts, S. Dauwe, B. Dhoedt, F. De Turck, P. Demeester, *"Tag Based Generation of User Profiles"*. Proceedings of ICOMP 2009, the 2009 International Conference on Internet Computing: 125-131, Las Vegas, Nevada, USA, July 2009.

[7] S. Dauwe, M. Strobbe, **O. Van Laere**, F. De Turck, B. Dhoedt, P. Demeester, *"Location-based Service Enabling Platform for Cultural Heritage Environments"*. Proceedings of ICWN 2009, the 2009 International Conference on Wireless Networks, Las Vegas, Nevada, USA, July 2009.

[8] **O. Van Laere**, M. Strobbe, K. Michiels, S. Schockaert, S. Dauwe, J. Vanattenhoven, C. van Nimwegen, P. Dhondt, T. Verbelen, B. Dhoedt, F. De Turck, P. Demeester, *"Enriching Networked Applications and Services through User Generated Content"*. Proceedings of the 48th FITCE congress, Prague, Czech Republic, September 2009.

[9] **O. Van Laere**, S. Schockaert, B. Dhoedt, *"Towards Automated Georeferencing of Flickr Photos"*. Proceedings of 6th Workshop on Geographic Information Retrieval (GIR), Zürich, Switzerland, February 2010.

[10] **O. Van Laere**, S. Schockaert, B. Dhoedt, *"Combining Multi-Resolution Evidence for Georeferencing Flickr Images"*. Proceedings of 4th International Conference on Scalable Uncertainty Management (SUM): 347-360, Toulouse, France, September 2010.

[11] **O. Van Laere**, S. Schockaert, B. Dhoedt, *"Ghent University at the 2010 Placing Task"* **(quantum leap award)**. Working notes of the 2010 MediaEval Workshop, Pisa, Italy, October 2010.

[12] J. Vanattenhoven, C. van Nimwegen, M. Strobbe, **O. Van Laere**, B. Dhoedt, *"Enriching Audio-Visual Chat with Conversation-Based Image Retrieval and Display"*. Proceedings of the International Conference on Multimedia (MM 2010), Firenze, Italy, October 2010.

[13] **O. Van Laere**, S. Schockaert, B. Dhoedt, *"Finding locations of Flickr resources using language models and similarity search"*. Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR), Trento, Italy, April 2011.

[14] **O. Van Laere**, S. Schockaert, B. Dhoedt, *"Ghent University at the 2011 Placing Task"* **(distinctive mention)**. Working notes of the 2011 MediaEval Workshop, Pisa, Italy, October 2011.

[15] C. De Rouck, **O. Van Laere**, S. Schockaert, B. Dhoedt, *"Georeferencing Wikipedia pages using language models from Flickr"*. Proceedings of the Terra Cognita 11 Workshop, Bonn, Germany, October 2011.

[16] S. Van Canneyt, S. Schockaert, **O. Van Laere**, B. Dhoedt, *"Time-dependent recommendation of tourist attractions using Flickr"*. Proceedings of the 23rd Benelux Conference on Artificial Intelligence (BNAIC), Ghent, Belgium, November 2011.

[17] **O. Van Laere**, J. Quinn, F. Langbein, S. Schockaert, B. Dhoedt, *"Ghent and Cardiff University at the 2012 Placing Task"*. Working notes of the 2012 MediaEval Workshop, Pisa, Italy, October 2012.

[18] S. Van Canneyt, S. Schockaert, **O. Van Laere**, B. Dhoedt, *"Using social media to find places of interest: A case study"* **(best paper award)**. Proceedings of the First ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, Redondo Beach, California, USA, November 2012.

[19] S. Van Canneyt, S. Schockaert, **O. Van Laere**, B. Dhoedt, *"Detecting Places Of Interest using Social Media"*. Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence, Macau, China, December 2012.

1.6.4 Publications in national journals and conferences

[1] M. Strobbe, **O. Van Laere**, F. De Turck, B. Dhoedt, *"Automatic Learning of User Interests for Personalized Communication Services"*. Published in proceedings of the 9th UGent-FirW PhD Symposium, Gent, Belgium, December 2008.

[2] **O. Van Laere**, M. Strobbe, F. De Turck, B. Dhoedt, *"Managing and Using Context Aware Information"*. Published in proceedings of the 10th UGent-FirW PhD Symposium, Gent, Belgium, December 2009.

References

[1] *Facebook by the Numbers*. Available from: http://mashable.com/2011/10/21/facebook-infographic/ [cited November 19th, 2012].

[2] *Flickr: Explore!* Available from: http://www.flickr.com/explore [cited November 19th, 2012].

[3] *Twitter 2012 - Infographic*. Available from: http://infographiclabs.com/news/twitter-2012/ [cited November 19th, 2012].

[4] *LinkedIn Announces Third Quarter 2012 Financial Results*. Available from: http://press.linkedin.com/News-Releases/147/LinkedIn-Announces-Third-Quarter-2012-Financial-Results [cited November 19th, 2012].

[5] *Foursquare Nears 20 Million Users And Crowley Talks About His Co-founder's Recent Departure*. Available from: http://articles.businessinsider.com/2012-03-10/tech/31142426_1_foursquare-sxsw-dennis-crowley [cited November 19th, 2012].

[6] *Yahoo! Where on Earth IDentifier*. Available from: http://developer.yahoo.com/geo/geoplanet/guide/concepts.html [cited November 19th, 2012].

[7] *Location-Based Services: Market Forecast, 2011-2015*. Available from: http://www.pyramidresearch.com/store/Report-Location-Based-Services.htm [cited November 19th, 2012].

[8] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. *Web-a-where: geotagging web content*. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 273–280, 2004. Available from: http://doi.acm.org/10.1145/1008992.1009040, doi:http://doi.acm.org/10.1145/1008992.1009040.

[9] M. D. Lieberman, H. Samet, and J. Sankaranayananan. *Geotagging: using proximity, sibling, and prominence clues to understand comma groups*. In Proceedings of the 6th Workshop on Geographic Information Retrieval, pages 6:1–6:8, 2010. Available from: http://doi.acm.org/10.1145/1722080.1722088, doi:http://doi.acm.org/10.1145/1722080.1722088.

[10] M. F. Goodchild, M. J. Egenhofer, K. K. Kemp, D. M. Mark, and E. Sheppard. *Introduction to the Varenius project*. International Journal of Geographical Information Science, 13(8):731–745, 1999.

[11] M. F. Goodchild. *A Geographer Looks at Spatial Information Theory*. In Proceedings of the International Conference on Spatial Information Theory, pages 1–13. Springer-Verlag, 2001.

[12] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind. *Geographic Information Systems and Science*. John Wiley & Sons, April 2005.

[13] *World Geodetic System 1984*. Available from: http://earth-info.nga.mil/GandG/publications/tr8350.2/tr8350_2.html [cited November 19th, 2012].

[14] T. Jordan, M. Raubal, B. Gartrell, and M. J. Egenhofer. *An Affordance-Based Model of Place in GIS*.

[15] S. Schockaert and M. De Cock. *Neighborhood restrictions in geographic IR*. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 167–174, 2007.

[16] T. Rattenbury, N. Good, and M. Naaman. *Towards automatic extraction of event and place semantics from flickr tags*. In Proceedings of the 30th Annual International ACM SIGIR Conference, pages 103–110, 2007.

[17] T. Rattenbury and M. Naaman. *Methods for extracting place semantics from Flickr tags*. ACM Transactions on the Web, 3(1):1–30, 2009.

[18] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. *Mapping the world's photos*. In Proceedings of the 18th International Conference on World Wide Web, pages 761–770, 2009.

[19] P. Serdyukov, V. Murdock, and R. van Zwol. *Placing flickr photos on a map*. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 484–491, 2009. Available from: http://doi.acm.org/10.1145/1571941.1572025, doi:http://doi.acm.org/10.1145/1571941.1572025.

[20] A. Rae and P. Kelm. *Working Notes for the Placing Task at MediaEval2012*. In Working Notes of the MediaEval Workshop. CEUR-WS.org, ISSN 1613-0073, online http://ceur-ws.org/Vol-927/mediaeval2012_submission_-6.pdf, 2012.

[21] O. Van Laere, S. Schockaert, and B. Dhoedt. *Georeferencing Flickr resources based on textual meta-data*. Accepted for publication in Information Sciences, Elsevier, February 2013.

[22] O. Van Laere, J. Quinn, S. Schockaert, and B. Dhoedt. *Spatially-aware Term Selection for Flickr Photo Geotagging*. Accepted for publication in IEEE Transactions on Knowledge and Data Engineering, IEEE, February 2013.

[23] O. Van Laere, S. Schockaert, and B. Dhoedt. *Georeferencing Flickr photos using language models at different levels of granularity: An evidence based approach*. Web Semantics: Science, Services and Agents on the World Wide Web, 2012.

[24] O. Van Laere, S. Schockaert, V. Tanasescu, B. Dhoedt, and C. Jones. *Georeferencing Wikipedia documents using data from social media sources*. submitted, 2012.

[25] O. Van Laere, S. Schockaert, and B. Dhoedt. *Finding locations of Flickr resources using language models and similarity search*. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval, pages 48:1–48:8, 2011.

[26] O. Van Laere, S. Schockaert, and B. Dhoedt. *Ghent university at the 2010 Placing Task*. In Working Notes of the MediaEval Workshop, 2010.

[27] O. Van Laere, S. Schockaert, and B. Dhoedt. *Ghent university at the 2011 Placing Task*. In Working Notes of the MediaEval Workshop, 2011.

[28] O. Van Laere, S. Schockaert, J. A. Quinn, F. C. Langbein, and B. Dhoedt. *Ghent and Cardiff University at the 2012 Placing Task*. In Working Notes of the MediaEval Workshop, 2012.

2

Extracting geographic information from textual data in Social Networks

Related work for the main contributions outlined in chapters 3 to 7 is highlighted in this chapter. As the latter chapters also contain a related work section, the same publications are often referred.

⋆ ⋆ ⋆

P. Kelm, V. Murdock, S. Schmiedeke, S. Schockaert, P. Serdyukov and O. Van Laere

The contents of this chapter is published as part of the book chapter *"Geotagging in social networks"* in the book "Social Media Retrieval" (N. Ramzan, R. van Zwol, J. Lee, K. Clüver, X. Hua, eds.), Computer Communications and Networks series, Springer, ISBN: 978-1-4471-4554-7, November 2012. Originally submitted, March 2012.

2.1 Introduction

The current Web 2.0 and its social media enable hundreds of millions users to actively participate in the creation of content. Apart from the millions of photos and videos that are uploaded every day, a huge amount of textual data is created

in the form of tweets, tags, status updates, news and blog articles, among others. Three major categories of textual data can be considered, based on the length of a typical message:

Articles such as stories on news websites, blog posts or Wikipedia pages largely correspond to the classical notion of a text document, using full sentences which are structured in paragraphs and sections. Such documents typically discuss one or few related topics in some level of detail.

Microposts such as Twitter and SMS messages have inherent length restrictions and are therefore mostly limited to a few words. They make wide use of abbreviations and often use short phrases instead of full sentences. Similarly, user comments and Facebook status updates, while not limited in size per se, seldom contain more than a few phrases. In addition to their length, microposts are also characterized by the use of terms that are not found in natural language, including hashtags, emoticons, and (shortened) URLs.

Tags are individual terms or keywords which are assigned to some resource, e.g. a photo on Flickr [1], an URL on social bookmarking sites such as Delicious [2], or an artist on Last.fm [3]. Tags can be used to describe the associated resource, either to allow others to find it or to add structure to one's own collections, although tags are used in practice for other purposes as well (e.g. describing actions, such as *toread* in the case of bookmarks).

Textual data often provide cues to its geographic scope, which we can use, for instance, to find out to which user communities a given blog post is likely to be relevant, to find out where a Twitter user is likely located, or to find out where a given photo was taken. However, the mechanisms that are needed to map textual data to geographic locations vary substantially, depending on the nature of the text. In what follows, we provide an overview of the state of the art approaches to extracting geographic information from the three aforementioned classes of textual data.

2.2 Finding the geographical scope of articles

One of the most natural ideas to map textual data to geographic locations is to identify place names (toponyms) in the text, and to look up where the corresponding places are located. To this end, a gazetteer can be used. Section 1.2 already briefly discussed the important challenge of resolving ambiguities.

Another challenge with using place names is that apart from administrative place names, people frequently use a variety of vernacular place names. The location of these places can often not be found in gazetteers, and they may even have inherently vague boundaries. For most cities, for instance, it is not clear where

downtown is located, exactly. Similarly, regions such as *Eastern Europe* or *the Mediterranean* do not have clear-cut boundaries either. For this reason, a number of authors have looked at acquiring knowledge about vernacular place names from the web, in an automated or semi-automated fashion [4–6]. To cope with the vague nature of region boundaries, these methods represent the spatial extent of a vernacular place name as a probability distribution or a fuzzy set; see [7] for a discussion on the links and differences between both representations. Most approaches use some form of social media to collect a set of coordinates that are believed to lie within the region of interest (e.g. using georeferenced Flickr photos that are tagged with the name of that region), and then estimate a density from the resulting point set.

While most gazetteer based methods to georeferencing have been introduced for general web pages, a number of authors have recently looked at georeferencing social media. In Fink et al. [8], for example, a more or less standard toponym resolution strategy is used to determine the geographical focus of blogs. On the other hand, Wing and Baldridge [9] propose a supervised learning approach to georeferencing Wikipedia articles. In particular, they place a grid over the surface of the Earth and represent grid cells and the document to be georeferenced as probability distributions of terms. The Kullback-Leibler divergence is then used to find the grid cell which is most similar to the document. Their method led to a median prediction error of 11.8 km when estimating the location of Wikipedia articles, without the use of any gazetteers or any attempts at explicitly disambiguating place names. As we will see further, this approach is closer in spirit to methods that have been proposed to georeference tagged resources, such as Flickr photos.

2.3 Finding the geographical scope of microposts

Taken in isolation, microposts are much harder to georeference than full-length texts. For instance, in [9] it was found that when moving from georeferencing Wikipedia pages to georeferencing Twitter messages (tweets), the median error increased from 11.8 km to 479 km. Due to their short length, microposts often do not provide enough context for accurately disambiguating place names. Moreover, the aversion of full sentences renders techniques such as named entity recognition ineffective. However, microposts are often not posted in isolation. Previous messages from the same user can be exploited as context information.

For example, Cheng et al. [10] propose a method to determine the city in which a Twitter user is located (among a pre-selected set of cities). Each city is modeled as a probabilistic language model, which can be used to estimate the probability that the users tweets were written by a resident of that city. While this baseline model only found the correct city for 10% of the users, substantial improvements were found when using a term selection method to filter out all terms that are

not location-relevant, leading to a 49.8% accuracy. Along similar lines, Kinsella et al. [11] train language models over geotagged Twitter messages, and rely on Kullback-Leibler divergence to compare the models of locations with the models of tweets. The results that are reported show that around 65% of the tweets can thus be located within the correct city (among a pre-selected set of cities) and around 20% even within the correct neighbourhood (in this case, within the spatial scope of New York only). To assess the effectiveness of a gazetteer based method, it was found that passing tweets to Yahoo! Placemaker only classifies 1.5% of the tweets within the correct neighbourhood. This provides further support for the hypothesis that gazetteer based methods are generally ineffective to deal with microposts.

Another aspect of microposts that can help georeferencing lies in the fact that profiles of the users of such application are often (partially) available. For example, the location field of Facebook users reveals their location, when this information has been provided. Interestingly, missing values of this field can often be accurately estimated by looking at the location fields of people in the social network of the user. For example, Backstrom et al. [12] show that for users with at least 16 friends whose location field is available, the correct location can be estimated within 25 miles in 69.1% of the cases, as opposed to 57.2% when using IP address georeferencing.

2.4 Finding the geographical scope of tagged resources

Where the short length of microposts already poses difficulties for traditional gazetteer based georeferencing methods, this is even more prominent in the case of tagged resources, such as Flickr photos. Since resources are essentially described as a bag of terms, any form of linguistic processing (e.g. named entity recognition) is impossible. On the other hand, due to the descriptive nature of tags, the collection of all georeferenced Flickr resources provides a potentially invaluable source of geographic information. As already mentioned, Flickr photos could be used to find information about the spatial extent of vernacular regions. In such cases, we start from a place name and are interested in the corresponding location. Conversely, some authors have looked at the problem of choosing the best term to describe a given region, a choice which is strongly affected by the geographic scale. Indeed, depending on the chosen scale, the same location may best be described by *France*, *Paris*, or *Eiffel tower*. Rattenbury and Naaman [13], for instance, use burst analysis techniques to find terms that occur unusually often in a given region at a given scale. In the case of the tag *Paris*, a burst of occurrences will be witnessed at world scale, but not within the city of Paris itself.

Several authors have looked at the problem of estimating the location of a Flickr photo, and a dedicated benchmark initiative called the Placing Task[1] has

[1] http://www.multimediaeval.org/mediaeval2011/placing2011/

even been introduced to allow for a fair comparison of different methods. In [14], Serdyukov et al. propose to train a probabilistic language model for each cell of a grid-representation of the Earth, and assign a photo to the cell whose model is most likely to have generated its tags. Particular emphasis is put to the influence of smoothing, showing that spatially aware forms of smoothing may lead to small, but statistically significant improvements. Along similar lines, [15] proposes a two-step approach to find the location of a Flickr photo. In the first step, again language models are used to find the area which is most likely to contain the location where a given photo was taken, although a k-medoids clustering is used instead of a grid. In the second step, a form of similarity search is used to find the photo from the training set that is most similar to the photo to be georeferenced, among all photos that are known to be located within the area that was selected in the first step. Their results show that neither of these two steps alone is sufficient for accurate georeferencing. Intuitively, the first step is needed as a form of implicit disambiguation of ambiguous tags, while the second step is needed to escape from the limited granularity of classification based approaches. In [16], Crandall et al. combine textual, visual and temporal features to georeference Flickr photos. First, mean shift clustering is used to find important locations from the training set, after which linear Support Vector Machines (SVMs) are trained for each of these locations. Their results show that, depending on the considered scale, combining visual and textual features results in a significant improvement over using only textual features.

As in the case of microposts, using an effective technique to select spatially relevant terms can substantially improve the results. While standard methods such as χ^2 feature selection or selecting the most frequently occurrent tags are sometimes used, these methods are outperformed using a term selection method proposed by Hauff et al. in [17]. Their *geographic spread* term selection is based on the idea that spatially relevant tags are those tags that occur only around a few clusters of locations, while still favouring tags that occur often.

Training location-specific language models from georeferenced Flickr photos can have uses beyond the task of georeferencing other Flickr resources. For example, De Rouck et al. [18] show how Wikipedia pages (about places) can be georeferenced using language models trained on Flickr photos. They found that for 15.4% of the pages coordinates were found that are within 1 km of their true location, as opposed to 4.2% in the case of Yahoo! Placemaker. This suggests that implicit geographic information can indeed be extracted from the Social Web, in this case Flickr data, which can be used to complement or even replace gazetteers. As another example of the use of location-specific language models, several authors have looked at extending topic models [19] with a spatial component. For example, Sizov et al. in [20] propose such a method, called GeoFolk, and show its benefits on tasks such as tag recommendation. In [21] a geographically informed

topic model is used to analyze the geographic influence on the popularity of different topics. Somewhat related, [22] uses topic models trained from georeferenced Twitter messages with the aim of analyzing lexical variation across different parts of the US.

References

[1] *Flickr*. Available from: http://www.flickr.com [cited November 18th, 2011].

[2] *Delicious*. Available from: http://www.delicious.com/ [cited November 18th, 2011].

[3] *Last.fm*. Available from: http://www.last.fm/ [cited November 18th, 2011].

[4] F. A. Twaroch, C. B. Jones, and A. I. Abdelmoty. *Acquisition of a vernacular gazetteer from web sources*. In Proceedings of the 1st International Workshop on Location and the Web, pages 61–64, 2008.

[5] S. Schockaert and M. De Cock. *Neighborhood restrictions in geographic IR*. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 167–174, 2007.

[6] L. Hollenstein. *Capturing vernacular geography from georeferenced tags*. Master's thesis, University of Zurich, 2008.

[7] S. Schockaert. *Vague regions in Geographic Information Retrieval*. SIGSPATIAL Special, 3:24–28, July 2011.

[8] C. Fink, C. Piatko, J. Mayfield, D. Chou, T. Finin, and J. Martineau. *The geolocation of web logs from textual clues*. In Proceedings of the 2009 International Conference on Computational Science and Engineering, pages 1088–1092, 2009.

[9] B. Wing and J. Baldridge. *Simple supervised document geolocation with geodesic grids*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 955–964, 2011.

[10] Z. Cheng, J. Caverlee, and K. Lee. *You are where you tweet: a content-based approach to geo-locating Twitter users*. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pages 759–768, 2010.

[11] S. Kinsella, V. Murdock, and N. O'Hare. *"I'm eating a sandwich in Glasgow": modeling locations with tweets*. In Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents, pages 61–68, 2011.

[12] L. Backstrom, E. Sun, and C. Marlow. *Find me if you can: improving geographical prediction with social and spatial proximity*. In Proceedings of the 19th International Conference on World Wide Web, pages 61–70, 2010.

[13] T. Rattenbury and M. Naaman. *Methods for extracting place semantics from Flickr tags*. ACM Transactions on the Web, 3(1):1–30, 2009.

[14] P. Serdyukov, V. Murdock, and R. van Zwol. *Placing flickr photos on a map*. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 484–491, 2009.

[15] O. Van Laere, S. Schockaert, and B. Dhoedt. *Finding locations of flickr resources using language models and similarity search*. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval, pages 48:1–48:8, 2011.

[16] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. *Mapping the world's photos*. In Proceedings of the 18th International Conference on World Wide Web, pages 761–770, 2009.

[17] C. Hauff and G.-J. Houben. *WISTUD at MediaEval 2011: Placing task*. In Working Notes of the MediaEval Workshop, 2011.

[18] C. De Rouck, O. Van Laere, S. Schockaert, and B. Dhoedt. *Georeferencing Wikipedia pages using language models from Flickr*. In Proceedings of the Terra Cognita 2011 Workshop, pages 3–10, 2011.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan. *Latent Dirichlet allocation*. Journal of Machine Learning Research, 3:993–1022, 2003.

[20] S. Sizov. *GeoFolk: latent spatial semantics in web 2.0 social media*. In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, pages 281–290, 2010.

[21] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. *Geographical topic discovery and comparison*. In Proceedings of the 20th International Conference on World Wide Web, pages 247–256, 2011.

[22] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. *A latent variable model for geographic lexical variation*. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1277–1287, 2010.

3

Georeferencing Flickr resources based on textual meta-data

In this chapter, we present a thorough analysis of the performance of using language models for the task of georeferencing Flickr photos. To this end, we implemented a scalable georeferencing framework capable of esimating language models for up to 20 000 classes using up to 16 million training photos, which can be processed on a single 16-core computer with 16 GB of memory. This framework is the result of participating in the 2010 and 2011 editions of MediaEval's Placing Task. We analyze and optimize the results for each of the different components involved in the georeferencing process and propose new methods to further improve this. All of the evaluation is carried out using a standard benchmark test set while the results are compared to state-of-the-art frameworks.

⋆ ⋆ ⋆

Olivier Van Laere, Steven Schockaert and Bart Dhoedt.

Accepted for publication in Information Sciences, Elsevier, February 2013. Originally submitted, March 2012. Major revision submitted, November 2012.

Abstract The task of automatically estimating the location of web resources is of central importance in location-based services on the Web. Much attention has been focused on Flickr photos and videos, for which it was found that language modeling approaches are particularly suitable. In particular, state-of-the art systems for

georeferencing Flickr photos tend to cluster the locations on Earth in a relatively small set of disjoint regions, apply feature selection to identify location-relevant tags, then use a form of text classification to identify which area is most likely to contain the true location of the resource, and finally attempt to find an appropriate location within the identified area. In this paper, we present a systematic discussion of each of the aforementioned components, based on the lessons we have learned from participating in the 2010 and 2011 editions of MediaEval's Placing Task. Extensive experimental results allow us to analyze why certain methods work well on this task and show that a median error of just over 1 kilometer can be achieved on a standard benchmark test set.

3.1 Introduction

With the rising popularity of smartphones and tablet computers, location plays an ever increasing role on the web. Many applictions, including search engines, try to adapt the services they offer to the current location of the user. This requires that resources (e.g. web pages in the case of search engines) be associated with a geographic scope. Such geographic information can be obtained in various ways. One way of learning information about places is to encourage users to explicitly share information about their whereabouts with their friends and contacts. This is the case with Foursquare[1], on which users can compete with each other for points they earn for each "check-in" at a certain place, or Twitter[2] where the user's current location can be attached to the tweet. Secondly, a gazetteer can be consulted as a source of geographical information. Gazetteers (for example GeoNames[3] or Yahoo! Geoplanet[4]) are essentially lists or indexes containing information about a large number of known places, described by different features such as geographical coordinates and semantic types. Creating and maintaining such a gazetteer is mostly expert driven and a cumbersome and time-consuming task. Gazetteers clearly provide a valuable source of geo-information, if one is able to disambiguate between the possibly multiple entities with the same name. For instance, if one needed details on an entity described as "Paris", a gazetteer would normally contain at least two entities: one for Paris, France and one for Paris, Texas. In absence of any additional information, it is hard to disambiguate between these two entities, although in this example using a "default sense" heuristic (based on for instance the population count) would in most cases return the correct meaning.

In our work, we focus on yet another way of gathering geographical information. As the amount of user-contributed textual data on the Web is growing

[1] http://www.foursquare.com
[2] http://www.twitter.com
[3] http://www.geonames.org
[4] http://developer.yahoo.com/geo/geoplanet/

every day (e.g. by means of status updates on social networks, comments, reviews, ratings, blog posts, tagged photo and video uploads), and as many of those contributions also include geographical coordinates, there is a vast amount of textual information available for automated mining of geographical knowledge. More specifically, in this paper, we show how such automatically obtained geographic knowledge allows us to estimate geographical coordinates for Flickr photos and videos, using only the textual information from their Flickr tags. To this end, we train a classifier from the tags of Flickr photos with known coordinates (i.e. the location where the photo was taken), which is capable of selecting the area in which a previously unseen photo or video has most likely been taken. In a subsequent step, our system tries to find a precise location within that area, by identifying the photos from the training data that are most similar to the photo or video we want to localise.

Several approaches to this problem of georeferencing Flickr resources have already been proposed in the literature [27, 28, 32, 34–36] . To facilitate the comparison of different solutions, the Placing Task has been introduced in 2010 as part of the MediaEval[5] evaluation campaign. This task requires participants to georeference Flickr videos based on the associated tags, visual features, and user profile information. Both in 2010 and 2011, our system came out as the best performing one. The research goal of this paper is to analyze the results of our system and to perform an in-depth evaluation of the contributions of each of the different steps in our approach to the overall results.

The remainder of this paper is organized as follows: Section 3.2 summarizes related work. A general overview of our approach to extracting implicit geographical information from Flickr is presented in Section 3.3, as well as the description of the techniques used to employ this textual information in estimating the location of Flickr photos and videos. Next, Section 3.4 provides an in-depth analysis of different approaches of each individual component of the georeferencing process along with experimental results. Finally, Section 3.5 states the conclusions and discusses future work.

3.2 Related work

3.2.1 Finding locations of resources

The task of deriving geographic coordinates for photos has recently gained in popularity; see e.g. [26]. [44] published a survey on recent research and applications on the topic of georeferencing resources. Most existing approaches are based on clustering, in one way or another, to convert the task into a classification problem. For instance, in [6] locations of unseen resources are determined using mean shift

[5] http://www.multimediaeval.org

clustering, a non-parametric clustering technique from the field of image segmentation. The advantage of this clustering method is that an optimal number of clusters is determined automatically, requiring only an estimate of the scale of interest. Specifically, to find locations, the difference is calculated between the density of photos at a given location and a weighted mean of the densities in the area surrounding that location. To assign locations to new images, both visual (keypoints) and textual (tags) features were used. Experiments were carried out on a sample of over 30 million images, using both Bayesian classifiers and linear support vector machines, with slightly better results for the latter. Two different resolutions were considered corresponding to approximately 100 km (finding the correct metropolitan area) and 100 m (finding the correct landmark). It was found that visual features, when combined with textual features, substantially improve accuracy in the case of landmarks. A similar conclusion follows from the multimodal approach demonstrated and evaluated in [11]. In contrast, [24] discusses a method using only visual information. A novel high-level representation for videos, called bag-of-scenes, is proposed. In this approach, each component of the representation has a self-contained semantics that can be directly related to a specific place of interest. Experiments were conducted in the context of the MediaEval 2011 Placing Task, using the same dataset that we will use in this paper. In [15], another approach is presented which is based purely on visual features. For each new photo, the 120 most similar photos with known coordinates are determined. This weighted set of 120 locations is then interpreted as an estimate of a probability distribution, whose mode is determined using mean-shift clustering. The resulting value is used as a prediction of the image's location. Around 16% of the resources in the test set can be estimated within 200 km of their actual location.

The idea that when georeferencing images, the spatial distribution of the classes (areas) could be utilized to improve accuracy has been suggested in [32]. Their starting point is that typically not only the correct area will receive a high probability, but also the areas surrounding the correct area. Indeed, the expected distribution of tags in these areas will typically be quite similar. Hence, if some area a receives a high score, and all of the areas surrounding a also receive a relatively high score, we can be more confident in a being approximately correct than when all the areas surrounding a receive a low score. Motivated by this intuition, [32] proposes a location-aware form of smoothing when estimating probabilistic language models.

In addition to georeferencing Flickr photos, several authors have recently focused on finding the location of other web resources such as Twitter posts or Wikipedia pages. For instance, in [3], a probabilistic framework based on maximum likelihood estimation was used to estimate the location of users based on the content of their tweets. In particular, a generative probabilistic model proposed in [2] is used to determine words with a geographic scope within a tweet, and a form

of neighborhood smoothing is employed to refine the estimations. For 51% of the users, a location was obtained that is within a 100 mile radius of their true location. Next, [40] looked into georeferencing Wikipedia articles as well as Twitter posts. After laying out a grid over the Earth's surface (in a way similar to [32]), for each grid cell a generative language model is estimated. To assign a test item to a grid cell, its Kullback-Leibler divergence with the language models of each of the cells is calculated. In [7], we have shown how Wikipedia pages can be georeferenced using language models that are trained from Flickr, taking the view that the relative sparsity of georeferenced Wikipedia pages does not allow for sufficiently accurate language models to be trained, especially at finer levels of granularity.

Interestingly, some recent language modeling approaches have combined the idea of topic models with location-dependent language models. For instance, [9] proposes geographic topic models with the aim of simultaneously capturing linguistic variation across different regions and different topics.

3.2.2 Using locations of resources

When available, the coordinates of a photo may be used in various ways. In [1], for instance, coordinates of tagged photos are used to find representative textual descriptions of different areas of the world. These descriptions are then put on a map to assist users in finding images that were taken in a given location of interest. Their approach is based on spatially clustering a set of geotagged Flickr images, using k-means, and then relying on (an adaptation of) tf-idf weighting to find the most prominent tags of a given area. Similarly, [23] looks at the problem of suggesting useful tags, based on available coordinates. The relevance of a given tag is measured in terms of the number of users that have used it to describe photos located within a certain radius of the current photo's coordinates. A refinement of this method only looks at tags that occur with visually similar photos, which is shown to improve the quality of the proposed tags. Along similar lines, our method could be used to suggest coordinates when users are tagging their photos and videos, automating the process that is now carried out manually using Suggestify[6], a web application that enables people to suggest a location for ungeotagged Flickr photos of someone else. This could contribute to making a larger fraction of the photos and videos on Flickr associated with an explicit location[7]. As a related use case, we can consider the problem of making search engines aware of spatial constraints in users' queries. For example, to allow users to specify a geographic scope for their query, Google introduced an option to search *nearby*[8] in February 2010. Implementing such a method involves a correct interpretation of the spatial

[6] http://suggestify.appspot.com/

[7] http://www.flickr.com/map/ shows that around 178M photos are geotagged of over 6.97 billion photos (http://www.flickr.com/explore) on Flickr. Accessed on March 14th, 2012.

[8] http://googleblog.blogspot.com/2010/02/refine-your-searches-by-location.html

constraint (e.g. based on a gazetteer in combination with location information obtained from the user's IP address for disambiguation) and a mechanism to identify the geographic scope of a website [18]. This latter problem could be solved using a combination of different methods. Web pages containing explicit mentions of addresses could be localised using standard techniques for geocoding (e.g. by comma group resolution [22]). In general, however, the textual content of the web page needs to be used as evidence. While traditionally gazetteer-based methods have been used to this end, initial results have shown that our model for georeferencing based on language models trained from Flickr can successfully be used to georeference resources such as Wikipedia pages [7].

Some authors have looked at using geographic information to help diversify image retrieval results [19, 25]. Finally, in [16], GeoSR is presented as a way of measuring the semantic relatedness of Wikipedia articles based on their geographic context, allowing users to explore information in Wikipedia that is relevant to a particular location. In [41], one would like to discover points of interests based on geotagged photos by applying a form of spectral clustering. The problem with this approach is that there is no unified way for determining the appropriate parameters for the clustering algorithm. For that purpose, a self-tuning clustering approach is proposed.

To conclude the discussion of related work, we describe a number of techniques that also treat the problem of extracting knowledge about toponyms from Flickr, but for the goal of learning geographic knowledge per se, e.g. as a method of enriching existing gazetteers. In our approach, in addition to toponyms, various other types of tags may provide useful evidence. For example, the tag "pepsi" has no relevance when compiling or enriching gazetteers, but, since it will occur more frequently in some countries or states than in others, it may be helpful to disambiguate the meaning of other terms.

Geotagged photos are useful from a geographic perspective, to better understand how people refer to places, and overcome the limitations and/or costs of existing mapping techniques [12]. For instance, by analyzing the tags of georeferenced photos, [17] found that the city toponym was by far the most essential reference type for specific locations. Moreover, evidence is provided suggesting that the average user has a rather distinct idea of specific places, their location and extent. Despite this tagging behaviour, the conclusion was that the data available in the Flickr database meets the requirements to generate spatial footprints at a sub-city level. Finding such footprints for non-administrative regions (i.e. regions without officially defined boundaries) using georeferenced resources has also been adressed in [31] and [39]. Another problem of interest is the automated discovery of which names (or tags) correspond to places. Especially for vernacular place names, which typically do not appear in gazetteers, collaborative tagging-based systems may be a rich source of information. In [28], methods based on burst-

analysis are proposed for extracting place names from Flickr. Finally, note that to some extent, ontologies, and in particular ontologies of places may be derived from Flickr tags [30]. The approach differs substantially from the one presented in this work, as the authors do not use geographic coordinates for deriving the ontologies; these are induced from the Flickr tags vocabulary using a subsumption-based model.

3.3 Georeferencing framework

3.3.1 Overview

In this paper we present our approach to georeferencing resources from the Web purely based on textual meta-data. Given an unseen resource x described by a certain set of tags $\mathcal{T}$, we estimate a location based on the information contained in $\mathcal{T}$. In particular, we consider the scenario of estimating the location (i.e. in actual latitude/longitude coordinates) of Flickr photos, based on the tags associated with them. This approach is purely text based and no visual or other features are used in the process, although existing approaches described in literature do leverage these features, as described in Section 3.2.

A common approach to georeferencing is by resolving toponyms (place names) in the given text with the help of gazetteers or named entity recognition (NER). Although this may seem straightforward, it is complicated in practice due to the ambiguity of toponyms. For full-text documents, named entity taggers can be used to detect the words in a phrase that represent place names, while their coordinates can be resolved from a gazetteer. In the case of Flickr tags however, linguistic context and capitalization is missing, hence heuristics need to be used to determine whether names such as “turkey” or “nice” refer to places or to the common words in English.

To avoid explicitly disambiguating tags, we interpret the problem of georeferencing as a classification problem, by partitioning the locations on Earth into a finite number of areas, of which the most likely area for a given resource, represented as its set of tags $\mathcal{T}$, is determined. This method avoids seeking specifically for toponyms and the need of any form of (explicit) disambiguation. A first drawback, however, is that the result is an *area*, consisting of multiple photos and their locations, rather than a single pair of coordinates. Another drawback is that the partitioning of the training data into a finite set of areas superimposes a certain factor of scale to the results: when the partitioning results in a relatively small number of areas, say 500, they are likely to cover a larger area of the world’s surface. Depending on the textual information available, such a partitioning can be too coarse for one resource whereas it is too fine-grained for another resource. Take the following example: consider a photo with only one tag *elbulli*, referring

to a restaurant in Spain. It is very unlikely that starting from 500 areas, one would be able to pinpoint the location of the restaurant within 1 kilometer of its actual location. On the other hand, for a photo annotated with the tags *germany, europe*, one would rather think of a larger area consisting of some of the 500 areas, actually requiring a coarser scale for this kind of resource. There clearly is no single scale that will perform best for all photos we would like to georeference. In our approach we present, in Section 3.3.8, two different methods for converting the resulting area into a pair of coordinates, resolving the first issue. The similarity based area refinement we propose addresses the second issue in particular. It allows using a coarser scale while still being able to accurately estimate locations by finding similar items within this coarse clustering.

The general architecture of the georeferencing framework we propose is outlined below:

1 Starting from a (preprocessed) geotagged training set, i.e. a dataset that contains the *true location* of the resources (where *true location* is to be considered as the location provided by the owner of the resource), a clustering algorithm is applied to cluster the locations of the resources into a finite set of disjoint areas $\mathcal{A}$.

2 Next, by applying feature selection, a vocabulary V consisting of discriminative tags is compiled, i.e. tags that are likely to be indicative of geographic location.

3 In a subsequent step, we train a language model. Given a unseen resource x, identified by its set of tags $\mathcal{T}$ after feature selection, a classifier will rank the areas $\mathcal{A}$ at a given scale and determine the area a that is considered to be the most likely area to contain the resource x.

4 To convert this area a into an actual location estimate, we search training items contained in this area that are most similar based on their tags. The location of these training items is then used to derive a location estimate.

We now discuss each of these steps in more detail.

3.3.2 Data preprocessing

The training sets we use consist of meta-data from Flickr photos. For each photo that is uploaded to its website, Flickr maintains several types of meta-data, which can be obtained via a publicly available API. In this paper, three types of meta-data will be relevant: descriptive tags that have been provided by the photo owners, the user's location (as provided by the user in her profile as free text, e.g. "Ghent, Belgium"), and information about where the photos were taken. The location information includes a geographical coordinate (latitude and longitude), and information

about the accuracy of the location, encoded as a number between 1 (world-level) and 16 (street-level). Starting from a raw dataset, a number of preliminary filtering steps are carried out on this data:

1 Photos that do not contain any tags or have invalid coordinates are removed from the collection.

2 In order to retain only those photos that provide meaningful information w.r.t. within city or sub-city scale location, only photos whose location accuracy is at least 12 (viz. city level accuracy) are retained.

3 Users on Flickr can upload content in bulk, i.e. uploading multiple photos with the same information at once. In order to reduce the impact of these bulk uploads, as pointed out in [32], for photos containing the same upload date, an identical tag set and the same coordinates, only a single instance is retained.

The photos that remains after these filtering steps are used for obtaining clusters of locations, and for estimating language models.

3.3.3 Clustering the training data

In order to interpret the problem of georeferencing resources as a classification problem, we cluster the locations of the training data into sets of disjoint areas $\mathcal{A}$ over which language models can be trained.

Different approaches have been described in literature. In [6], a mean shift procedure is used to find highly photographed locations based on the density of photos. The authors found that this procedure was effective in determining these places at different scales (a metropolitan scale of 100 km and a landmark-level scale of 100 m). In contrast to most clustering approaches, mean shift does not require the number of clusters to be predetermined, but rather relies on a scale parameter to choose the number of clusters implicitly. In [32] a fixed grid overlay is placed over to the Earth. In this work, the authors considered varying grid sizes (and thus scales) comparing to location cells of roughly 1, 5, 10, 50 and 100 kilometers long over their sides. In [19], k-means clustering is used to identify famous locations in collections of geo-tagged photos from Flickr. In our previous work [36] we also used k-medoids (partitioning around medoids) clustering to obtain areas of interest. An alternative to clustering would be to use boundaries of administrative divisions such as cities, provinces, and countries. However, such boundaries are not freely available for every country, and usually no information about areas at the sub-city scale is available.

In what follows, we provide an overview of a number of techniques for obtaining a clustered representation of locations. An experimental comparison of these techniques will be provided in Section 3.4.2.

3.3.3.1 k-medoids clustering

Partitioning Around Medoids (PAM) or k-medoids is a clustering technique closely related to the well-known k-means algorithm; the algorithm partitions the data into groups of data points while the objective is to minimize the squared error, which is the sum of the distances between each individual point in a cluster and the cluster center (the medoid). The k-medoids algorithm is more robust to noise and outliers than k-means. Distances are calculated using the geodesic (great-circle) distance measure. The algorithm is an iterative process. Also, increasing the number of data points results in a quadratically increasing computing cost ($O(n^2)$). We therefore apply sampling during the optimization of each individual cluster. Per cluster, a maximum of 512 data points are swapped with the medoid point m in every iteration. An example clustering of our main training set using this algorithm (k = 1000) is shown in Figure 3.1. As can be seen, metropolitan areas on both the Northeast and the West coast of the US are covered by a large number of smaller clusters, in contrast to little clusters covering large parts of northern South-America. This shows that the granularity of the clusters is based on the amount of information available in these regions.

The algorithm is defined as follows:

Randomly select k points to be the initial medoids
repeat
  Construct clusters c_i by assigning each data point to the closest medoid in terms of geodesic distance
  for each cluster c_i **do**
    for each regular datapoint o_j in the cluster c_i **do**
      Temporarily swap o_j and the medoid
      Calculate the new cluster cost $cost$ with this configuration
      if $cost$ is the lowest cost seen so far **then**
        Store a reference o_{best} to o_j
      end if
      Restore the original medoid and o_j as a regular data point
    end for
    Make o_{best} the new medoid and the original medoid a regular data point
    Remove all regular data points from the cluster c_i
  end for
until No more changes occur to the set of medoids

3.3.3.2 Grid based clustering

A second possibility is to use a grid. Intuitively, the idea is to lay a grid of square cells over the surface of the Earth. This clustering method is straightforward and computationally inexpensive ($O(n)$). A single run over all data points is sufficient

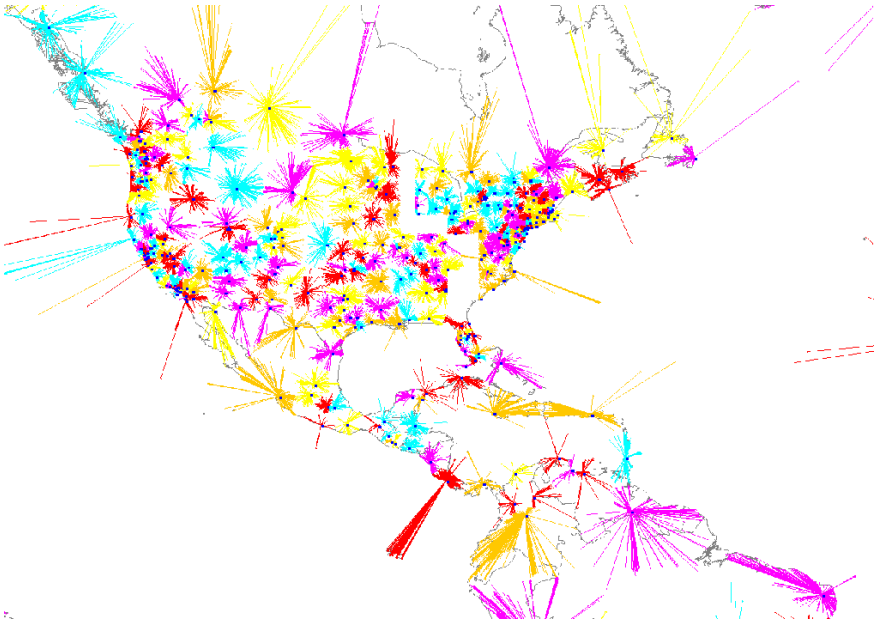


Figure 3.1: Sample clustering of a part of the main training set using Partition Around Medoids, k = 1000.

to assign them to their corresponding cluster based on the geographical coordinates of the points. When clustering the data, only cells that actually contain at least one image are considered as a cluster.

Note that, when a cell size of 1 degree in latitude and longitude is considered for each of the sides of the grid cells, this roughly corresponds to a side of 111 km in latitude and 111 km in longitude near the equator. However, the length of the longitude side converges to 0 km at the geographic poles, making it impossible to map equally sized cells when using only one parameter value to simultaneously define the length of both sides of the grid cells.

An example clustering of our main training set using this algorithm is shown in Figure 3.2. In this example, grid cells are considered using a cell size of 4.375 degrees of latitude and longitude, as this value resulted in a configuration of 1001 clusters, facilitating a comparison with Figure 3.1.

3.3.3.3 Mean shift clustering

A third and final clustering algorithm we discuss is mean shift clustering [5]. As opposed to k-medoids and grid-based clustering, which require specifying the desired number of clusters beforehand, mean shift clustering requires a parameter h that is considered the *scale of observation*. The number of resulting clusters

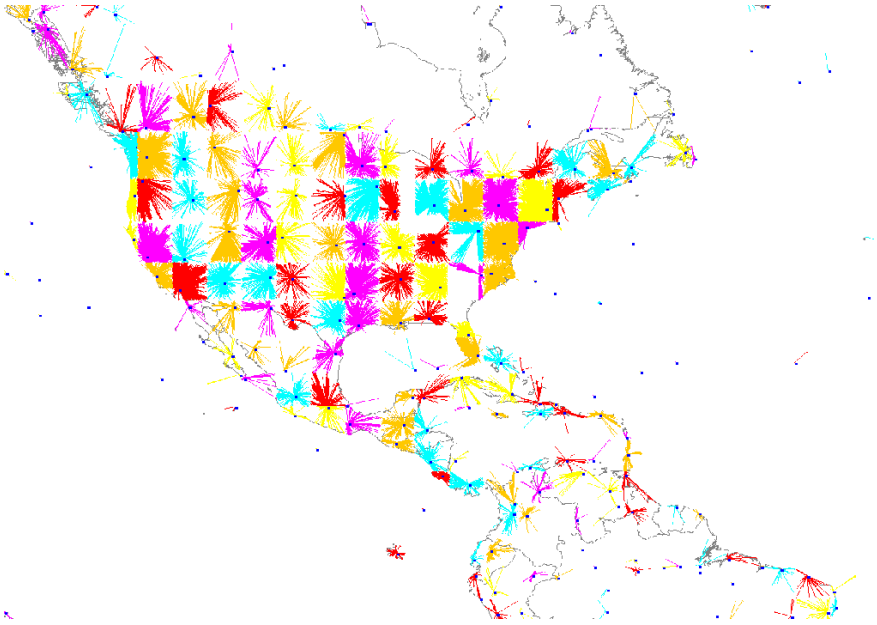


Figure 3.2: Sample clustering of a part of the main training set using the grid clustering approach. The side of each cell are 4.375 degrees latitude and longitude, resulting in 1001 clusters.

emerges from the choice of this scale factor. Mean shift clustering is again an iterative process.

Mean shift clustering is again an iterative process. During each iteration, a data point x, represented as an n-dimensional vector, *shifts* towards a mean location

$$x_{i+1} = x_i + m_h(x_i)$$

in which the *mean shift value* $m_h(x)$ is computed as follows

$$m_h(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) \cdot x}{\sum_{i=1}^{n} K_h(x - x_i)}$$

where the mean is computed using a kernel function $K_h(x)$. The kernel function can either be a uniform kernel

$$K_h(z) = \begin{cases} 1 & \text{if} \quad ||z|| \leq h \\ 0 & \text{if} \quad ||z|| > h \end{cases}$$

or a Gaussian kernel

$$K_h(z) = e^{\frac{-||z||^2}{2 \cdot h^2}}$$

with $||z||^2$ defined as $\sum_{i=1}^{n} |z_i|^2$ for an n-dimensional vector $\vec{z} = (z_1, z_2, \ldots, z_n)$.

For the approach outlined in this paper, we need clusterings at different levels of granularity. The reason is that depending on the nature of the training data, coarser or finer grained clusterings will lead to an optimal performance. Initial experiments have revealed, however, that changing the scale parameter does not substantially reduce the overall number of clusters. Figure 3.3 illustrates this: in this example, there exist a number of small clusters located close to the West and East coasts of Northern America. These clusters are outside the influence range (defined by the scale parameter h) of other clusters. One possible solution could be to increase the scale parameter, but due to these isolated clusters this parameter needs to be increased substantially, again resulting in a coarse clustering.

To cope with this effect, we consider a variant of the mean shift algorithm. Once the mean shift procedure finishes producing a set of clusters $\mathcal{A}$, the following additional steps are taken:

Initialize a set of data points $\mathcal{P}$ to the empty set
for each cluster a in the set of clusters $\mathcal{A}$ **do**
  if $|a| < t$ **then**
    Add all of the data points p of a to $\mathcal{P}$
    Remove a from $\mathcal{A}$
  end if
end for
for each data point p in $\mathcal{P}$ **do**
  Assign p to the closest cluster a in $\mathcal{A}$ where closest is defined as the minimum geodesic distance between p and the medoid of a

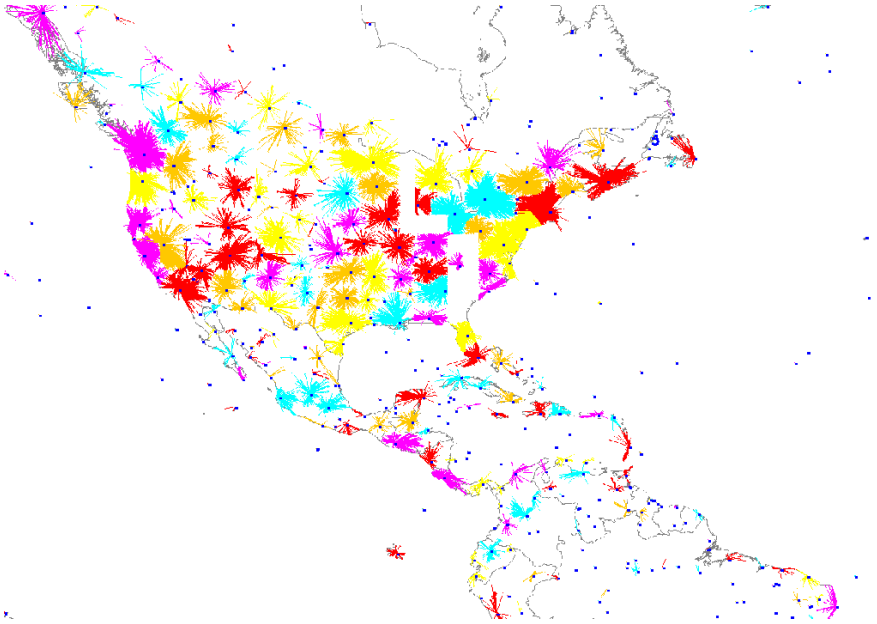


Figure 3.3: Sample clustering of a part of the main training set using the mean shift algorithm, h = 150, resulting in 2349 clusters in total.

end for

In order to avoid introducing additional parameters we keep the value of t fixed at 10 throughout the experiments; we thus only use the scale parameter h to change the number of clusters. Figure 3.4 illustrates the clustering obtained when merging smaller clusters with their closest neighbors.

The difference between Figures 3.3 and 3.4 is clear: the small clusters close to the coasts of the North American contintent are merged with other clusters. In general, most of the isolated clusters are merged: of the 2349 original clusters the standard mean shift algorithm produced, 1384 clusters with less than 10 photos were merged with with their nearest neighbours. Of these 1384 clusters, more than 1100 clusters contained even less than 5 photos. It is important for our approach that each cluster represents a certain minimal amount of information if one wants to train reliable language models based on that information.

3.3.4 Feature selection

In order to train a language model for a specific scale, a set of tags (vocabulary V) is needed, consisting of tags that are likely to be indicative for the geographic location. A comparative study on feature selection techniques used in text classification in general can be found in [42]. To the best of our knowledge, no similar

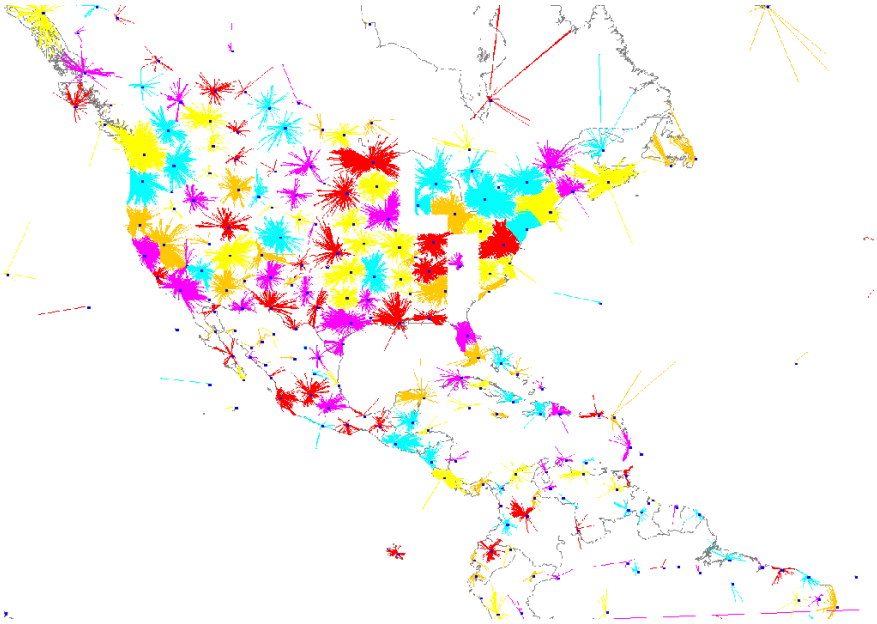


Figure 3.4: Sample clustering of a part of the main training set using the mean shift algorithm with merges, h = 150, $t = 10$, resulting in 965 clusters in total.

comparison has been carried out in literature focused on the effect of different feature selection approaches in georeferencing. These six feature selection methods will be evaluated in Section 3.4.3.

3.3.4.1 χ^2

Let $\mathcal{A}$ be the set of areas that is obtained after clustering the data into k clusters. Then for each area a in $\mathcal{A}$ and each tag t assigned to photos in a, the χ^2 statistic is given by:

$$\chi^2(a,t) = \frac{(O_{ta} - E_{ta})^2}{E_{ta}} + \frac{(O_{t\bar{a}} - E_{t\bar{a}})^2}{E_{t\bar{a}}} + \frac{(O_{\bar{t}a} - E_{\bar{t}a})^2}{E_{\bar{t}a}} + \frac{(O_{\bar{t}\bar{a}} - E_{\bar{t}\bar{a}})^2}{E_{\bar{t}\bar{a}}} \tag{3.1}$$

where O_{ta} is the number of photos in area a where tag t occurs, $O_{t\bar{a}}$ is the number of photos outside area a where tag t occurs, $O_{\bar{t}a}$ is the number of photos in area a where tag t does not occur, and $O_{\bar{t}\bar{a}}$ is the number of photos outside area a where tag t does not occur. Furthermore, E_{ta} is the number of occurrences of tag t in photos of area a that could be expected if occurrence of t were independent of the location in area a, i.e. $E_{ta} = N \cdot P(t) \cdot P(a)$ with N the total number of photos, $P(t)$ the probability that a photo contains tag t and $P(a)$ the probability

that a photo is located in area a, the latter two probabilities being estimated using maximum likelihood:

$$P(t) = \frac{\sum_{a \in \mathcal{A}} O_{ta}}{\sum_{t' \in V} \sum_{a \in \mathcal{A}} O_{ta}} \tag{3.2}$$

$$P(a) = \frac{|a|}{N} \tag{3.3}$$

Similarly, $E_{t\bar{a}} = N \cdot P(t) \cdot (1 - P(a))$, $E_{\bar{t}a} = N \cdot (1 - P(t)) \cdot P(a)$ and $E_{\bar{t}\bar{a}} = N \cdot (1 - P(t)) \cdot (1 - P(a))$.

The most relevant features for a given area can then be selected by choosing the features with the highest value for the χ^2 statistic. To select a vocabulary V containing the v most discriminative features, we need to aggregate the rankings obtained for every area a into a single ranking. This is accomplished by first selecting the best tag from each of the rankings, then the tags at position 2, etc.

3.3.4.2 Maximum χ^2

Maximum χ^2 (max χ^2) is similar to χ^2 except that when constructing the overall ranking, each tag is ranked according to its highest χ^2 value over all areas a. In other words, not only the ranking imposed by the χ^2 statistic plays a role here, but also the actual value. In principle, even the highest ranked tag for a given area may not be selected if its χ^2 value is too low (e.g. because the area corresponds to a small cluster where none of the photos bears any tags that are descriptive of the location).

3.3.4.3 Log-Likelihood

As an alternative to the χ^2 statistic, we consider Dunning's *log-likelihood* statistic [8]. For each term t and area $a \in \mathcal{A}$, the log-likelihood is given by:

$$\begin{aligned} G^2(a,t) = \quad & 2(O_{ta} \log O_{ta} + O_{t\bar{a}} \log O_{t\bar{a}} + O_{\bar{t}a} \log O_{\bar{t}a} + O_{\bar{t}\bar{a}} \log O_{\bar{t}\bar{a}} \\ & + N \log N \\ & - (O_{ta} + O_{t\bar{a}}) log(O_{ta} + O_{t\bar{a}}) - (O_{ta} + O_{\bar{t}a}) log(O_{ta} + O_{\bar{t}a}) \\ & - (O_{t\bar{a}} + O_{\bar{t}\bar{a}}) log(O_{t\bar{a}} + O_{\bar{t}\bar{a}}) - (O_{\bar{t}a} + O_{\bar{t}\bar{a}}) log(O_{\bar{t}a} + O_{\bar{t}\bar{a}})) \end{aligned} \tag{3.4}$$

where $O_{ta}, O_{t\bar{a}}, O_{\bar{t}a}$ and $O_{\bar{t}\bar{a}}$ are defined as in Section 3.3.4.1 and N is the total number of photos in the training data. Similarly, the most relevant features for a given area can then be selected by choosing the features with the highest value for the G^2 statistic. In order to obtain the vocabulary V, the same method as described in Section 3.3.4.1 is used.

3.3.4.4 Information Gain

Whereas the χ^2 based methods are rooted in statistics, information gain uses information theory to select informative terms. It measures the change in entropy when learning about the presence or absence of the tag. The information gain of the tag t is defined as:

$$\begin{aligned} G(t) = \quad & -\sum_{a \in \mathcal{A}} P(a) \log P(a) \\ & +P(t) \sum_{a \in \mathcal{A}} P(a|t) \log P(a|t) \\ & +P(\bar{t}) \sum_{a \in \mathcal{A}} P(a|\bar{t}) \log P(a|\bar{t}) \end{aligned}$$

with $P(a)$, the probability for area a, is estimated by Equation (3.3). Similarly, the probability for $P(t)$ is estimated by dividing the number of occurrences of the tag by the total number of tag occurrences (Equation (3.2)). $P(a|t)$ is estimated as the number of tag occurrences of tag t in area a, divided by the total number of occurrences of tag t:

$$P(a|t) = \frac{O_{ta}}{\sum_{a' \in \mathcal{A}} O_{ta}}$$

$P(\bar{t})$ and $P(a|\bar{t})$ are defined likewise, using the number of occurrences of all tags but t.

Note that information gain immediately produces a single ranking, in contrast to the χ^2 statistic, which produces a ranking per area. Hence, we can simply choose the vocabulary V by selecting the v tags with the highest information gain.

3.3.4.5 Most frequently used (MFU)

A particularly simple term selection technique that is sometimes used consists of selecting the terms that occur in the largest number of documents. Despite the simplicity of the method, it often performs remarkably well in practice [42].

3.3.4.6 Geographical spread (geospread)

As a sixth and final feature selection method, we describe a *geographic spread filtering* feature selection method presented in [13] and applied in [14]. In this work, a score is proposed that captures to what extent the occurrences of a tag are clustered around a small number of locations. The geographical spread score is calculated as follows:

Place a grid over the world map with each cell having sides of 1 degree latitude and longitude
for each unique tag t in the training data **do**
  for each i, j **do**
    For cell $c_{i,j}$, determine $|t_{i,j}|$, the number of training items containing the tag t

if $|t_{i,j}| > 0$ **then**
for each $c_{i',j'} \in \{c_{i-1,j}, c_{i+1,j}, c_{i,j-1}, c_{i,j+1}\}$, the neighbouring cells of $c_{i,j}$, **do**
Determine $|t_{i',j'}|$
if $|t_{i',j'}| > 0$ and $c_{i,j}$ and $c_{i',j'}$ are not already connected **then**
Connect cells $c_{i,j}$ and $c_{i',j'}$
end if
end for
end if
end for
$count =$ number of remaining connected components
$score(t) = count/max(|t_{i,j}|)$, with $max(|t_{i,j}|)$ over all original cells $c_{i,j}$.
end for

In the algorithm, merging of neighbouring cells is necessary in order to avoid penalizing geographic terms that cover a wider area.

Given the definition of the geographical spread score, a clear distinction should come to light between terms that are quite location bound on the one hand, and very general tags on the other hand. The smaller the resulting score for a tag t, the more specific its geographic scope and thus the more it is coupled to a specific location. We will refer to this method as *geospread*.

3.3.4.7 Qualitative evaluation of the feature selection methods

Table 3.1 presents an overview of the 10 highest ranking features according to each of the term selection algorithms discussed in this section. The features selected by χ^2, *max* χ^2 and *log-likelihood* depend on a specific clustering of the training data (in this case, $k = 2500$), while the other methods construct a ranking over all the training data, independent a specific clustering.

Considering the features selected by χ^2, we observe that the list only consists of toponyms: a country, cities and regions are mentioned, as well as a name of a Russian conference center: *igromir*. Note that the top ranking tags in this example are a random sample of the best ranking tags for each of the 2500 areas used for creating the ranking. However, when analyzing the first 2500 terms (and beyond) of the entire feature ranking, the behaviour witnessed persists.

The *max* χ^2 method returns a seemingly similar ranking, but this time, the geographical entities are, all but two (*bahiabrazil* and *bolodecasamento*), referring to islands. The top ranking tag, *bolodecasamento*, is in fact non-geographical related and represents the Portuguese concept of a "wedding cake". This term immediately propagated to the top of the ranking because it occurred only once in the training data, within a given area containing only a single photo with this tag. By chance, the regular χ^2 method could have also ranked this term at the

Table 3.1: Overview of the top 10 terms according to different feature selection methods applied to the training data.

	χ^2	max χ^2	Log-likelihood
1	gijón	bolodecasamento	roma
2	lhaviyani	seychelles	hsinchu
3	montauk	vanuatu	medellin
4	wolfsburg	elhierro	korea
5	igromir	bahiabrazil	nara
6	saintebaume	lanyu	valdaosta
7	hartford	galapagos	alps
8	bulgaria	isleofman	nef
9	rochester	bermuda	snowymountains
10	mendoza	madagascar	stalbans
	Information Gain	Most frequently used	Geospread
1	california	geotagged	kaohsing
2	australia	2008	haninge
3	france	2009	greatermanchester
4	italy	california	hsinchu
5	japan	2007	antwerpen
6	canada	nikon	nikone3700
7	germany	beach	algarve
8	scotland	nature	sinpu
9	spain	canon	hsinpu
10	taiwan	travel	oxfordshire

top, instead of at position 2372, if it started processing the areas with the area specifically containing this tag. This behaviour can be explained by the use of the χ^2 measure (3.1) in general, which awards such a very specific case with a maximum score.
In general, the ranking favors tags that frequently occur in a single cluster (cfr. the islands) and rarely outside it over discriminative terms for certain areas that also occur elsewhere: e.g. *andorra*, ranked in position 347, has 314 occurrences in a single cluster, whereas *canada* is ranked in position 67 463 while it occurs 29 141 times, albeit spread out over many clusters.

The list of features obtained by *log-likelihood* contains words that describe admimistrative entities such as cities or countries, while the tags *alps*, *valdaosta* and *snowymountains* describe mountains or valleys. The tag *nef* refers to the raw file format for photos taken with Nikon cameras. It is included because it occured 295 times in the training data, of which 210 occurrences are by the same user in the same area. Methods to combat such problems would be to only use tags that have been used by a sufficiently large number of users, or only consider one occurrence per tag per user, for feature selecting purposes. In practice, however,

such methods tend to worsen results in the Placing Task setting, as training and test data may contain resources by the same user. In such a case user-specific tags are often helpful.

Inspecting the table further, we observe that information gain (IG), provides a list of country names, whereas "most frequently used" (MFU) returns a list of tags that rarely contain any reference to a place in particular (except for *california*). However, while tags like *beach* or *nature* are not toponyms, they might help in disambiguating cases where one needs to decide if a photo was taken near the sea or in the city.

Finally, the geospread measure presents a list of terms that it considers to have a very specific spatial scope. All but one tag in the list can indeed be easily located on a map. After analyzing the training data, the occurrence of *nikone3700* at position 6 out of more 1.13M in the list (details of the dataset can be found in Section 3.4.1) can be explained by the fact that a single user tagged 443 photos with the model of his camera in the same surroundings (the *greatermanchester* area, a tag also occuring in the top 3). As the geospread measure favors terms with a small geographical footprint, this term popped up as it can be tied to a very small region.

A quantitative evaluation of the different methods presented here follows in Section 3.4.3.

3.3.5 Language models

Given a previously unseen image x, we now attempt to determine in which area x was most likely taken. In this paper, we use a (multinomial) Naive Bayes classifier, which has the advantage of being simple, efficient, and robust. Initial results in [32] have shown good results for this classifier. Specifically, we assume that an image x is represented as its set of tags $\mathcal{T}$. Using Bayes' rule, we know that the probability $P(a|x)$ that image x was taken in area a is given by

$$P(a|x) = \frac{P(a) \cdot P(x|a)}{P(x)}$$

Using the assumption that the probability $P(x)$ of observing the tags associated with image x is fixed among all areas a, we find

$$P(a|x) \propto P(a) \cdot P(x|a)$$

Characteristic of Naive Bayes is the assumption that all features are independent. Translated to our context, this means that the presence of a given tag does not influence the presence or absence of other tags. Writing $P(t|a)$ for the probability of a tag t being associated to an image in area a, we find

$$P(a|x) \propto P(a) \cdot \prod_{t \in \mathcal{T}} P(t|a) \tag{3.5}$$

After moving to log-space to avoid numerical underflow, this leads to identifying the area a^* where x was most likely taken by:

$$a^* = \arg\max_{a \in \mathcal{A}} (\log P(a) + \sum_{t \in \mathcal{T}} \log P(t|a))$$

In this final equation, the prior probability $P(a)$ and the probability $P(t|a)$ remain to be estimated. In general, the maximum likelihood estimation can be used to obtain a good estimate of the prior probability but alternative approaches that include available meta-data are also possible, as we will show in Section 3.3.6. When estimating $P(t|a)$, a form of smoothing is needed to avoid a zero probability when a certain tag t does not occur in area a. We discuss different forms of smoothing in Section 3.3.7.

3.3.6 Estimating the prior probability

In this section, we discuss four possible ways of estimating the prior probability for the language models. An experimental comparison of these methods will be provided in Section 3.4.4.1.

3.3.6.1 Maximum likelihood and uniform prior

A common way of estimating the prior probability for the language models is using the maximum likelihood estimation:

$$P(a) = \frac{|a|}{N} \tag{3.6}$$

in which $|a|$ represents the number of training items contained in area a, and N represents the total number of training items as before.

A second, rather simple, way of estimating the prior probability might be to assign a uniform probability to all areas in $\mathcal{A}$.

$$P(a) = \frac{1}{|\mathcal{A}|} \tag{3.7}$$

One could also think of incorporating information from recent uploads from the same user, but this was not considered in this paper.

3.3.6.2 Using home location information from the user

The owner of a Flickr photo can provide a textual description of his home location in his profile, which can be retrieved using the public API. Most of the photo owners on Flickr have actually provided such a description, although it is not always precise or accurate. For example, in the best cases, the description looks like “San

Francisco, CA, United States" or "Cava de' Tirreni, Italia", pointing unambiguously to a known city whereas in the worst cases, the users describe their home location as "to infinity and beyond" or "homeless, US".

When the location information is present, we geocode this information (as provided by the user) using the Google Geocoding API[9] to convert the textual description to coordinates by extracting the *location* information returned from the Google API. For the example of "Cava de' Tirreni, Italia", this returns

```
"location" : {
   "lat" : 40.70205550,
   "lng" : 14.7065740
},
```

which indeed corresponds to the center of the town.

Although this example yields an interesting source of location information, this is however not the case for a large part of the descriptions. It might be clear that informal descriptions provided by the users present the Geocoding API with an unresolvable task. As will be explained in Section 3.4.1, in the case of the datasets considered in our experiments, the home location could be geocoded for 65% to 85% of the photos.

For an unseen resource x, when available, the information about the home location of the owner, $loc_{home}(x)$, can be used to estimate the prior probability as follows:

$$P(a) \propto (\frac{1}{d(m_a, loc_{home}(x)) + 0.001})^w \tag{3.8}$$

where $m_a(x)$ is the medoid of the area a and where $d(x, y)$ is the geodesic distance between the locations of points x and y. The parameter w allows to vary the influence of the home location on the prior probability. In the denominator a fixed value of 0.001 is introduced to avoid division by zero in the case that $loc_{home}(x)$ and m_a coincide. Using the home location of a user in this way corresponds to an assumption that all things being equal, locations within a reasonable distance from a user's home are more likely than locations at the other side of the world, even though we cannot exclude the latter case altogether.

3.3.6.3 Gaussian mixture models

Another way of using the home location when estimating the prior probability is to use a Gaussian mixture model (GMM) [29]. A Gaussian mixture model is a parametric probability density function represented as a weighted sum of Gaussian densities:

[9] http://code.google.com/apis/maps/documentation/geocoding/

$$P(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \tag{3.9}$$

$$\lambda = \bigcup_{i=1}^{M} \{w_i, \mu_i, \Sigma_i\}$$

where $x \in \mathbb{R}$ represents some numerical feature, $w_i = 1, \ldots, M$ are the mixture weights and $g(x|\mu_i, \Sigma_i), i = 1, \ldots, M$ are the component Gaussian densities. The mixture weights are required to sum up to 1: $\sum_{i=1}^{M} w_i = 1$. In our case, the Gaussian mixture model is used to estimate the prior probability of an area a, given the distance between a and the home location of the user. The feature x then corresponds to a distance.

The underlying idea is that there may be several types of relations between the home location of the user and the location of the photo:

1 With a certain probability w_1, the photo is taken nearby the house of the owner, in which case the prior probability of an area quickly decreases as the distance from the home location of the user increases.

2 With a probability w_2, the photo was taken on a day trip by the user.

3 With a probability w_3, the photo was taken on a holiday.

Using a Gaussian mixture model, we can jointly describe these scenarios, using the probabilities w_1, w_2 and w_3 as the mixture weights, and using one Gaussian to describe each scenario. Of course, neither the mixture weights nor the parameters of the Gaussians are known a priori. However, they can be estimated from the training data using the expectation-maximization (EM) procedure [29].

3.3.7 Smoothing methods

To avoid a zero probability when an unseen resource x contains a tag that does not occur with any of the photos from area a in the training data, smoothing is needed when estimating $p(t|a)$ in (3.5). Let O_{ta} be the occurrence count of tag t in area a. The total tag occurrence count O_a of area a is then defined as follows:

$$|O_a| = \sum_{t \in V} O_{ta} \tag{3.10}$$

where V is the vocabulary that was obtained after feature selection, as explained in Section 3.3.4.

One possible smoothing method is Bayesian smoothing with Dirichlet priors, in which case we have ($\mu > 0$):

$$P(t|a) = \frac{O_{ta} + \mu\ P(t|V)}{|O_a| + \mu} \tag{3.11}$$

where the probabilistic model of the vocabulary $P(t|V)$ is defined using maximum likelihood:

$$P(t|V) = \frac{\sum_{a \in \mathcal{A}} O_{ta}}{\sum_{t' \in V} \sum_{a \in \mathcal{A}} O_{ta}} \tag{3.12}$$

Another possibility is to use Jelinek-Mercer smoothing, in which case (3.11) becomes ($\lambda \in [0, 1]$):

$$P(t|a) = \lambda \frac{O_{ta}}{|O_a|} + (1 - \lambda)\ P(t|V) \tag{3.13}$$

with $P(t|V)$ defined as in (3.12). For more details on these smoothing methods for language models, we refer to [43]. The performance of both smoothing methods will be experimentally assessed in Section 3.4.4.1.

3.3.8 Finding a location within the chosen area

The previous steps result in the selection of an area a among those in $\mathcal{A}$ where the photo (or video) x has been taken (recorded). The final step that remains is converting this area a into an actual location, i.e. resolve the latitude and longitude coordinates for the resource x. We discuss two ways of accomplishing this: by determining the medoid of the area a, and by performing similarity search. Both methods are evaluated in Section 3.4.2.

3.3.8.1 Medoid based location estimation

The most straightforward way of converting an area a into actual coordinates is to choose the location of the medoid m_a, defined as:

$$m_a = \arg\min_{x \in \mathcal{A}_k} \sum_{y \in \mathcal{A}_k} d(x, y) \tag{3.14}$$

where $d(x, y)$ is the geodesic distance between the locations of photos x and y.

Clearly, the location estimates that are obtained in this way will mainly be useful when a sufficiently fine-grained clustering is used.

3.3.8.2 Similarity based location estimation

As an alternative, we explore the idea of using the location of the most similar resources from the training set that are known to be located in the chosen area a. Specifically, let $y_1, ..., y_n$ be the n most similar photos from our training set. We then propose to estimate the location of x as a weighted center-of-gravity of the locations of $y_1, ..., y_n$:

$$loc(x) = \frac{1}{n} \sum_{i=1}^{n} sim(x, y_i)^{\alpha} \cdot loc(y_i) \tag{3.15}$$

where the parameter $\alpha \in]0, +\infty[$ determines how strongly the result is influenced by the most similar photos only. The similarity $sim(x, y_i)$ between resources x and y_i was quantified using the Jaccard measure:

$$s_{jacc}(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

where we identify a resource with its set of tags *without feature selection*, to make full use of all the originally associated tags. In principle, Jaccard similarity may be combined with other types of similarity, e.g. based on visual features.

In (3.15), locations are assumed to be represented as Cartesian (x, y, z) coordinates rather than as (lat, lon) pairs. In practice, we thus need to convert the (lat_i, lon_i) coordinates of each photo y_i to its Cartesian coordinates.

3.4 Experimental results

In this section, we present a ground-truth based evaluation of each of the individual components of our georeferencing framework presented before. In general, after running an experiment using a given configuration, we will obtain an estimated location for each of the test items. We then analyze the results using two metrics:

1 **Acc@X**: number of location estimates within X km of the actual location, as defined by the ground truth, divided by the total number of items in the test set. The accuracy is determined for the following values of X: 1 km, 5 km, 10 km, 50 km, 100 km, 1000 km and 10 000 km.

2 **Median error distance (MER)**: median over all test items of the distance between the estimated and the true location.

The first metric was used in the evaluation of the Placing Task initiative, and provides a detailed view on the performance of a given method. However, in most cases, we also use the second metric, as it summarizes the performance of a method as a single value. A median error distance of for example 5 km (which is equal to

an $Acc@5$ of 50%), would indicate that half of the test set could be georeferenced with an error distance smaller than 5 kilometers.

The methodology of the experiments is as follows:

1 Using a baseline configuration, we will examine the performance of the different clustering approaches presented in Section 3.3.3. At the same time, we evaluate both area refinement approaches discussed in Section 3.3.8.

2 Next, using the best outcome of the initial experiment, we investigate the influence of the different feature selection algorithms, outlined in Section 3.3.4, on the results of the georeferencing use case.

3 Again adopting the feature selection method yielding the best result, we analyze the impact of applying different forms of smoothing (Section 3.3.7) and different ways of calculating the prior probability (Section 3.3.6).

At the end of this multi-step, greedy, way of experimentation we provide an overview of these different experiments and their potential improvements. Finally, in Section 3.4.6, we discuss the influence of adding more training data.

Before elaborating on the individual experiments, we provide a clear overview of the datasets used in this paper.

3.4.1 Datasets

For all experiments in this paper, the collection of test items is the same. This collection consists of the development and test data provided for the 2011 edition of the Placing Task, which is available with the Task organizers. The data consists of Flickr videos and their meta-data (which is represented in the same way as Flickr photos). Bearing in mind that some experiments need the home location of the owner of the videos, we filtered out those videos for which this information could not be retrieved. The final test set therefore contained the data for 13 390 Flickr videos.

With respect to the training data, we have used the dataset that was available to the Placing Task participants for most of the experiments carried out in this paper. This dataset constists of 3 185 343 georeferenced Flickr photos and their meta-data. As mentioned in Section 3.3.2, we preprocessed this dataset by removing photos with invalid coordinates, with missing tag information and items originating from a batch upload. On this dataset, no particular accuracy filtering was imposed, i.e. the accuracy level of the photos varies from 1 to 16, where 1 corresponds to accurate at world-level, 12 at city-block level and 16 at street-level[10]. This resulted in a dataset of 2 096 712 Flickr photos covering more or less the

[10]For details on the Flickr accuracy values, please refer to http://www.flickr.com/services/api/flickr.photos.search.html

entire world; it is referred to as the training set throughout the remainder of this paper, unless specified otherwise (viz. in the case of the experiments in Section 3.4.6). Figure 3.5 presents a geographical mapping of this dataset.

As some experiments require additional data, we crawled Flickr for data in April 2011 using the public API. The goal of the crawl was to fetch data about as many geotagged photos as possible. We were able to retrieve the meta-data of 105 118 157 photos being, at that time, over 70% of all geotagged photos. Again, we preprocessed the data obtained by removing the photos containing no valid coordinates or containing no tags, and we removed the bulk uploads. This resulted in a collection of 43 711 679 photos. Among these photos, we extracted those that reported an accuracy level of 16, which corresponds to a street level accuracy. This final step resulted in a set of 17 169 341 photos. This dataset was split into 16M and 1 169 341 photos. From the latter set, we randomly selected 10 000 photos whose owners have no other photos in the training set. This set of 10K photos is used as the *development set*, and will be used to optimize the parameters for the prior and smoothing techniques, independent of the actual test set. Of the remaining 16M photos, training sets of the first 1M, 2M, ..., 10M photos are extracted to provide the necessary training data for the experiments in Section 3.4.6.

Table 3.2 provides information on the different datasets and the number photos in each set, as well as information on the mean number of tags associated to the photos and the standard deviation of the number of tags.

3.4.2 Clustering and area refinement

The goal of this first experiment is to find out which clustering approach performs best and what is the optimal number of clusters, by comparing the results of the different clustering algorithms discussed in Section 3.3.3. At the same time, we compare both area refinement methods described in Section 3.3.8. The setup of this experiment is as follows:

- We use the training set consisting of 2 096 712 training items and 13 390 test items respectively.
- We cluster the training dataset into a predefined number of clusters k, varying from 500 to 20000 clusters.
- For the clustering algorithms that do not allow to fix the number of clusters beforehand (i.e. grid clustering and mean shift clustering), we set their respective parameters such that we can obtain a number of clusters that is more or less comparable to the predefined value for the PAM algorithm.
- In order to eliminate any side-effects introduced by the choice of the feature selection method, the *most frequently used* feature selection method (as

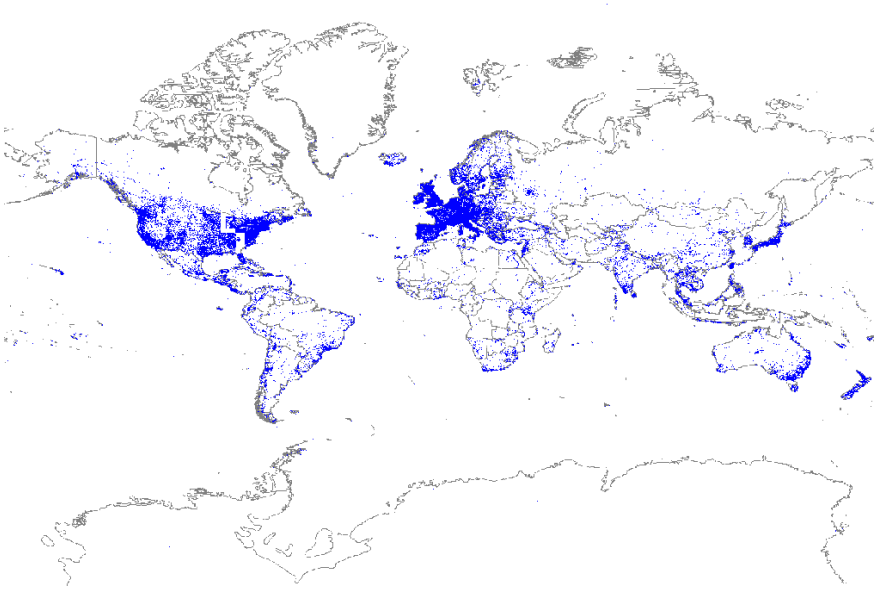


Figure 3.5: A plot of the photo data, after preprocessing, in the main training dataset from the Placing Task.

Table 3.2: Statistics of the considered datasets. Apart from the number of photos N in each of the datasets, the mean number of tags $\mu(|\mathcal{T}|)$ associated with each data item and the standard deviation $\sigma(|\mathcal{T}|)$ of this value are reported.

Dataset	N	$\mu(\lvert\mathcal{T}\rvert)$	$\sigma(\lvert\mathcal{T}\rvert)$	Type
General experiments				
Training set	2 096 712	7.801	7.491	photos
Test set	13 390	9.514	8.348	videos
Parameter optimization				
Development set	10 000	8.515	8.614	photos
Training experiments				
Training set 1M	1 000 000	8.745	8.463	photos
Training set 2M	2 000 000	8.746	8.462	photos
Training set 3M	3 000 000	8.747	8.456	photos
Training set 4M	4 000 000	8.747	8.457	photos
Training set 5M	5 000 000	8.748	8.461	photos
Training set 6M	6 000 000	8.749	8.463	photos
Training set 7M	7 000 000	8.749	8.465	photos
Training set 8M	8 000 000	8.750	8.464	photos
Training set 9M	9 000 000	8.750	8.465	photos
Training set 10M	10 000 000	8.751	8.466	photos

introduced in Section 3.3.4.5) is used. This method is independent of the underlying clustering.

- The baseline language model is applied with the maximum likelihood prior (3.6) and Bayesian smoothing with Dirichlet priors (3.11), $\mu = 1750$.

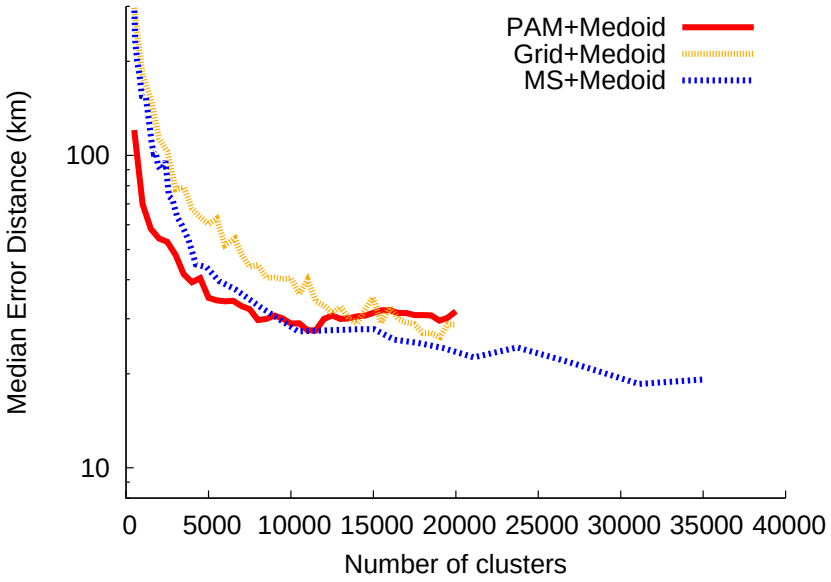
Figures 3.6(a) and 3.6(b) present the results of the experiment for a fixed number of features, $v = 45000$, using a log scale on the Y-axis. Two interesting conclusions can be drawn from this data. First, not surprisingly, using similarity search (Figure 3.6(b)) to convert an area to a precise location clearly performs better than returning the medoid of the areas (Figure 3.6(a)), especially when the number of clusters is small. Second, mean shift clustering is most effective to reduce the median error over the test set when the number of areas used is large (both Figures 3.6(a) and 3.6(b)). One should note that the results of this experiment are somewhat misleading: when using more clusters, more memory is required. Thus, when using a smaller number of clusters, we could include more features. Therefore, in a second experiment, we keep the amount of memory used fixed and choose the maximum number of features feasible for each clustering. When looking at Figures 3.7(a) and 3.7(b), containing the results when using 16 GB of memory and maximizing the number of features per number of clusters, we see that similarity search (Figure 3.7(b)) again outperforms the medoid based location conversion (Figure 3.7(a)). However, here PAM outperforms both other clustering algorithms substantially, and this at very low values for the number of clusters (Figure 3.7(b)). The optimal value is $k = 3000$ (and comparable results are found at $k = \{2500, 3000, \ldots, 4500\}$) with a median error distance of 10.89 km. Table 3.3 gives an idea of the total number of features we can include at different clustering scales.

The conclusion of this experiment is two-fold:

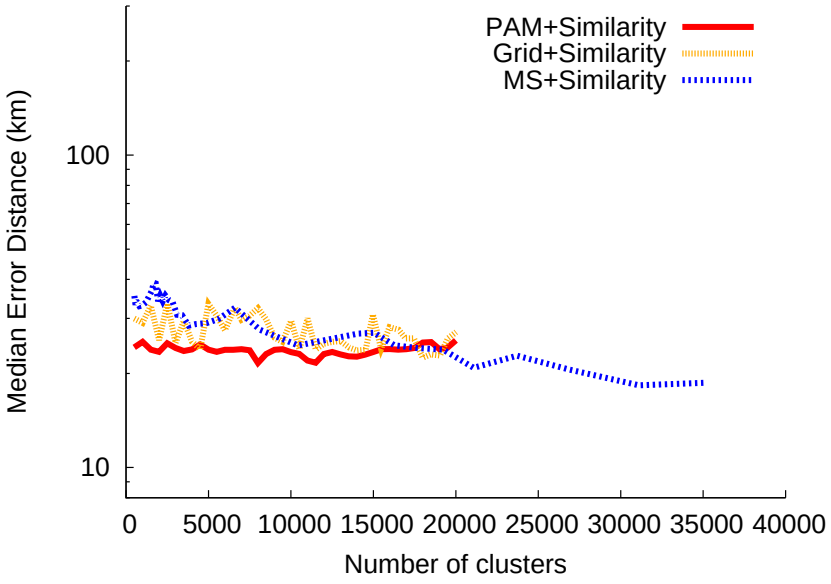
1 In order to convert an area to a precise location, a similarity based conversion (3.15) clearly outperforms a medoid based conversion.

2 In configurations that only allow a small number of features to be retained, mean shift clustering delivers the best performance. As soon as a sufficiently large number of features can be used, PAM outperforms both grid based and mean shift clustering algorithms, although we were unable to compare the algorithms in cases where a large number of clusters can be constructed using all features.

For the remainder of the paper, we will only consider PAM based clusterings combined with similarity based area refinement. To give an idea of the physical dimensions of the clusters generated by PAM, we included an overview of the (average) cluster size in kilometers (*size*) and standard deviation of the cluster size (σ) for a number of different values of k in Table 3.4. The average size of a cluster

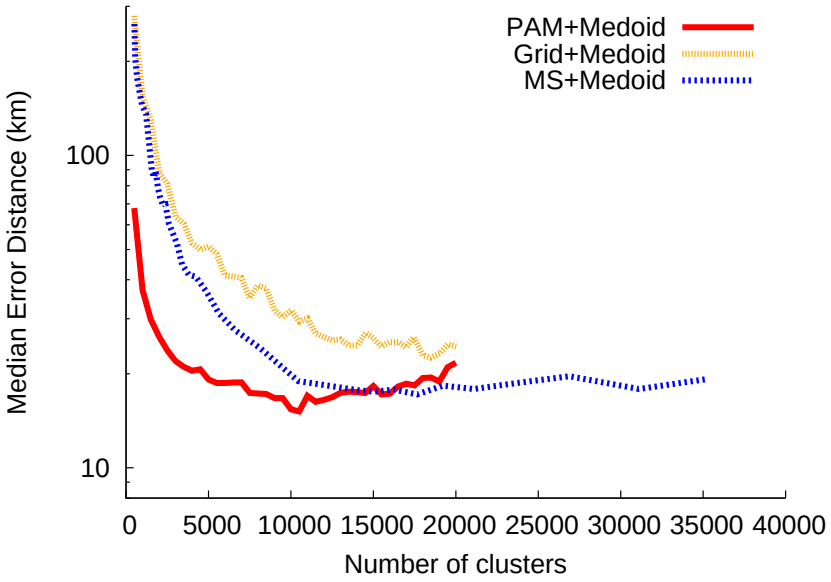


(a) Medoid based area refinement

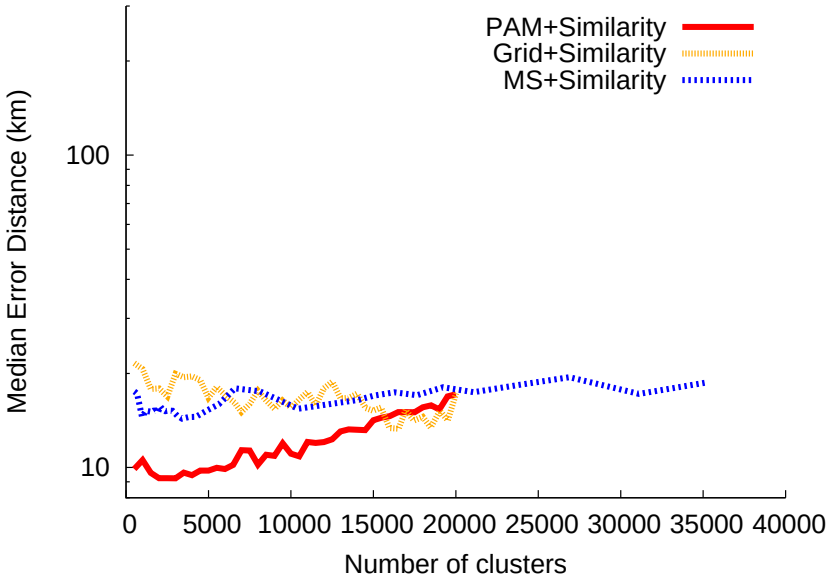


(b) Similarity based area refinement

Figure 3.6: Comparing the median error distance for 3 different clustering methods using a fixed number of features, v = 45 000.



(a) Medoid based area refinement



(b) Similarity based area refinement

Figure 3.7: Comparing the resulting median error distance of 3 different clustering methods using a fixed amount of memory (16 GB).

Table 3.3: Number of features $|V|$ that can be retained when using k clusters in the fixed memory configuration of our framework (16 GB of memory).

k	$\|V\|$		k	$\|V\|$
500	1 500 000		5500	275 000
1000	1 500 000		6000	250 000
1500	1 000 000		6500	225 000
2000	750 000		7000	200 000
2500	625 000		7500	200 000
3000	525 000		8000	200 000
3500	450 000		8500	175 000
4000	400 000		9000	175 000
4500	350 000		9500	150 000
5000	300 000		10000	150 000

Table 3.4: Statistics regarding the physical dimensions of clusters generated by the PAM algorithm.

k	size (km)	σ (km)
500	100.00	92.04
5000	20.76	20.21
10000	12.94	12.89
15000	9.47	9.50
20000	7.56	7.68

is defined as the sum of the distances between each datapoint and the medoid, divided by the number of datapoints.

3.4.3 Quantitative evaluation of the feature selection methods

In a second series of experiments, we evaluate the feature selection methods described in Section 3.3.4. Because of the outcome of the clustering experiment (Section 3.4.2), the clusterings are created using the PAM algorithm and similarity search will be used to convert the selected areas to a precise location, while the number of features is determined with respect to a fixed amount of memory (i.e. use as many features as possible for the experiment, given 16 GB of memory. For details, see Table 3.3). Also, as Figure 3.7 showed the optimal results to be obtained for a lower number of clusters, we will vary the cluster size from 500 to 10000.

Figure 3.8 depicts the results from this experiment. It is clear that for a large number of choices for the number of clusters, the *geospread* method outperforms all others and also results in the best performance overall, when the number of clusters is 2500. Somewhat surprisingly, the *most frequently used* method be-

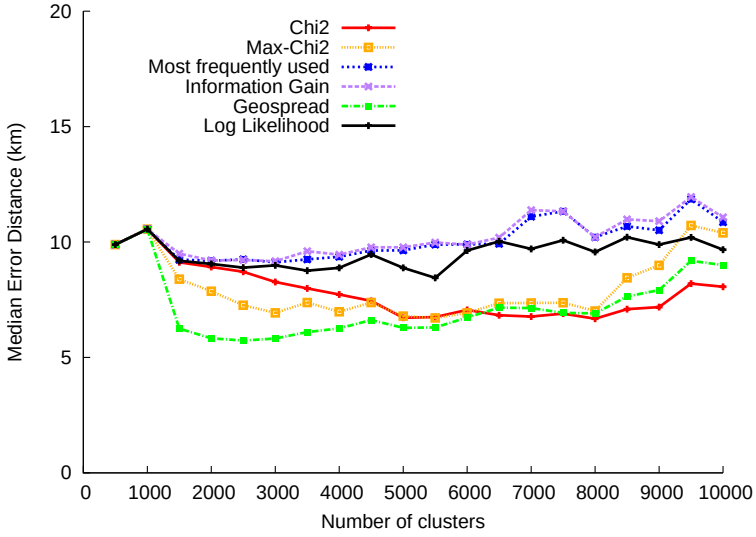


Figure 3.8: Median error distance over the test collection when estimating locations using different feature selection methods.

haves similarly to the Information Gain (IG) approach. Both perform substantially worse than the other methods. Note that all three aforementioned techniques are independent of the number of clusters k used, in constrast to χ^2, max χ^2 and *log-likelihood*. Also, χ^2 is surpassed in performance by the max χ^2 variant when the number of clusters is sufficiently small. Overall, the χ^2 based methods yield better results than *IG* or *most frequently used*. The *log-likelihood* measure mainly differs from χ^2 in the treatment of terms with only few occurrences, which leads to worse results in this scenario. The overall results deteriorate for an increasing number of areas k, while the best results, with the exception of χ^2 and *log-likelihood*, can be found around $k = 2500$.

We conclude by noticing that the *geospread* feature selection technique achieves a median error distance for the test set of 5.75 km. Applying a good feature selection technique thus improves the best results from the first experiment (9.23 km) by over 35%. Henceforth, we will apply *geospread* feature selection.

3.4.4 Language models

In the following experiment, we investigate two possible improvements to the baseline language modeling step. First, we investigate how different smoothing methods influence the results. In a subsequent experiment, we hope to find out which of the different implementations of the prior probability $P(a)$ outlined in Section 3.3.6 performs best.

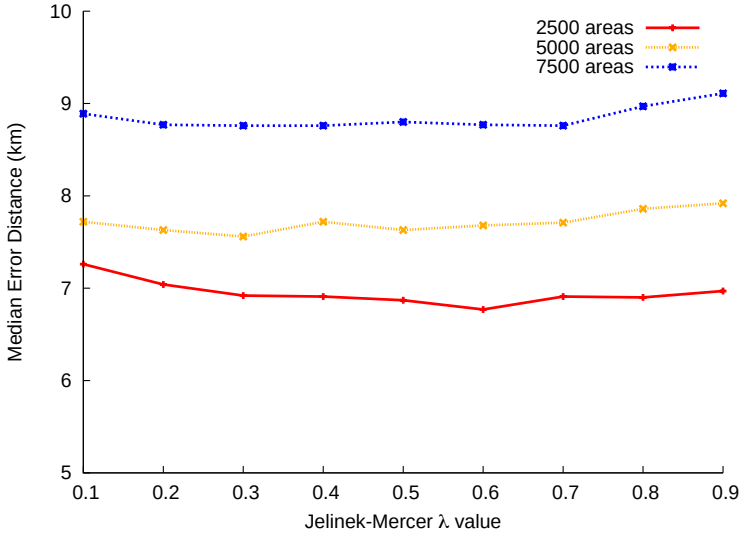


Figure 3.9: Median error distance over the development *set when estimating locations with 2500, 5000 and 7500 clusters using different λ values for the Jelinek-Mercer smoothing method.*

3.4.4.1 Smoothing methods

Before we start, let us outline the configuration used for the smoothing experiment. When optimizing the parameters, the regular test set of 13 390 test items is replaced by the *development set* introduced in Section 3.4.1, containing 10 000 previously unseen test photos. This avoids taking advantage of information in the regular test set when determining optimal parameter values.

Figures 3.9 and 3.10 present the median error distance of the evaluation over the *development set*. When using Jelinek-Mercer smoothing, we can see that varying parameter λ only has a limited impact on the results. We also observe that for each individual clustering scale k, the optimal parameter value differs. In these results, these values are 0.6, 0.3 and 0.3 for k equal to 2500, 5000 and 7500 respectively.

The results in Figure 3.10 reveal that the choice of the parameter μ has a stronger influence on the performance of Dirichlet smoothing. The main conclusion that we can draw from these results is that the optimal value for μ decreases when the number of clusters increases. Indeed, when the number of clusters increases, there are fewer tag occurrences per cluster, so intuitively we need a smaller value of μ for the same amount of smoothing.

Overall, when comparing the results from the Jelinek-Mercer smoothing method and Bayesian smoothing with Dirichlet priors, we see that the results are quite

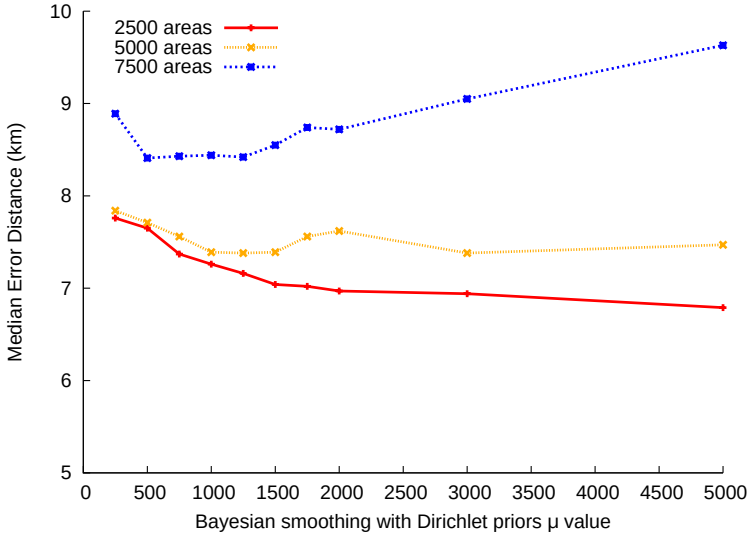


Figure 3.10: Median error distance over the development *set when estimating locations with 2500, 5000 and 7500 clusters using different μ values for the Bayesian smoothing method with Dirichlet priors.*

similar. In the best cases, just under 7 km of median error is measured, with a slightly better result for the Bayesian smoothing with Dirichlet priors. For $\lambda = 0.6$, Jelinek-Mercer smoothing produces a median error distance of 6.77 km, whereas Bayesian smoothing with Dirichlet priors results in 6.74 km at $\mu = 5000$. These findings confirm experimental results in other areas of information retrieval [33, 43], and to earlier work on georeferencing Flickr photos [32].

As our goal is to improve the overall performance of the framework, we will adopt the Bayesian smoothing method with Dirichlet priors for the remainder of our experiments, using optimized parameter values μ for each individual clustering level. These optimal parameter values are reported in Table 3.5.

3.4.4.2 Prior probability

Next, we determine the most suitable way of estimating the prior probability. In particular, we are interested in the results of the georeferencing process when using a maximum likelihood prior (ML), a uniform prior, the prior in (3.8), a prior based on Gaussian mixture models (GMM) with 1 to 5 component densities (GMM1 to GMM5), a combination of the ML and *Home* prior (ML+Home) and a combination of the ML and GMM1-5 priors (ML+GMMx).

Note that for this experiment, the regular test set (13 390 items) was used.

Table 3.6 presents the results of several of these configurations. The results

Table 3.5: Optimal μ values for Bayesian smoothing with Dirichlet priors for different values of clusters k, obtained after evaluation of a separate development set.

k	μ		k	μ
500	15000		5500	1000
1000	15000		6000	3000
1500	15000		6500	3000
2000	10000		7000	1500
2500	12500		7500	750
3000	5000		8000	750
3500	12500		8500	1000
4000	3000		9000	1000
4500	3000		9500	1000
5000	1750		10000	500

Table 3.6: Median error distance over the test collection when estimating locations with 500, 2500, 5000 and 7500 clusters, using different priors in the language models.

k	Uniform	ML	Home	ML+Home	GMM4	ML+GMM4
500	9.21	8.74	5.92	5.79	5.73	**5.61**
2500	5.38	5.34	**3.33**	3.34	3.96	3.65
5000	6.31	6.28	**2.92**	3.12	4.19	3.80
7500	7.23	6.75	**3.10**	3.21	4.88	3.63

of GMM1 to GMM3 are not presented, as these are all situated between the ML results and the GMM4 results. The results of GMM4 and GMM5 are identical and we therefore omitted GMM5 from this table. In the case of the *Home* prior, we set the parameter $w = 0.65$, a value that was experimentally found to be optimal. A discussion on this parameter value will follow shortly hereafter.

The optimal result can be found at $k = 5000$ when using a *Home* prior, resulting a median error distance of 2.92 km. The improvement over the baseline ML prior is clearly noticeable. When combined with the ML prior, the results of the Gaussian mixture model based prior are further improved. Interesting to note is that even though combining ML with the mixture models improves the overall performance, combining the Home prior with ML does not lead to a similar result. We can conclude that the *Home* prior, as defined in (3.8), is the best choice for optimizing the performance of our language modeling approach.

We investigated the robustness of the parameter w, controlling the influence of the distance between the suggested area and the home location of the photo owner. Figure 3.11 shows the results, confirming that our default parameter choice of $w = 0.65$ (based on initial experiments) turned out to be more or less in the middle of a range of good results. The figure also confirms that the infuence of the parameter w is rather limited, except for a small number of areas (e.g. $k = 500$).

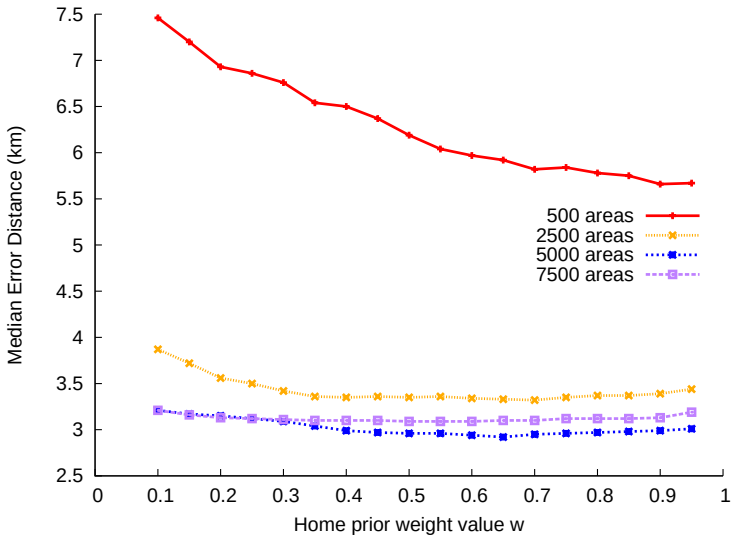


Figure 3.11: Median error distance over the test collection when estimating locations with 500, 2500, 5000 and 7500 clusters, using different weight values w for the home prior in the language models.

3.4.5 Summarizing improvements and results

Table 3.7 summarizes the result of optimizing the various components of the georeferencing framework and presents detailed accuracies for each of the configurations. Each transition to a better configuration is statistically significant[11] with a p-value $< 2.2 \times 10^{-16}$. The first substantial improvement is witnessed when using a similarity based area refinement instead of returning the location of the medoid of an area (Section 3.3.8.2). Although accuracies improve overall, the difference is most pronounced at smaller error distances. When the *geospread* method is used instead of choosing the most frequently occurring tags, the median error distance is further reduced. Finally, using Dirichlet smoothing with optimized values of the parameter μ and taking the home location of the photo owner into account if available, yields another significant improvement in accuracies and median error, which further decreases from 5.75 km to 2.92 km.

The optimal configuration presented here is an improved version of the baseline system that we used in the Placing Task benchmark that already outperformed other systems. The results presented in this paper show further improvements over the alternative approaches. Table 3.8 compares the optimal configuration of this paper to all the participants of the 2011 Placing Task.

[11] To evaluate the statistical significance, we used the sign test as the Wilcoxon signed-rank test is unreliable in this situation due to its sensitivity to outliers.

Table 3.7: Summarizing the results of optimal configurations of the framework in terms of accuracy at certain error distances and median error distance (in km) over the test collection of 13 390 items.

Configuration	Acc@1	Acc@10	Acc@100	Acc@1000	MER
clustering	22.15	46.3	59.24	69.02	15.16
+ similarity search	34.59	50.61	60.69	69.81	9.23
+ geospread	35.05	53.91	65.15	72.65	5.75
+ smoothing + home prior	38.21	65.58	83.24	92.05	2.92

Table 3.8: Comparison of the optimal configuration of this paper and the submissions to the 2011 Placing Task, evaluated over the 5347 test videos for 2011.

	1 km	10 km	100 km	1000 km	10000 km
Li et al. [21]	0.21%	1.12%	2.71%	12.16%	79.45%
Krippner et al. [20]	9.86%	21.49%	29.79%	43.26%	84.16%
Ferres et al. [10]	14.61%	42.66%	56.65%	68.64%	94.93%
Choi et al. [4]	20.00%	38.20%	52.60%	66.30%	94.20%
Hauff et al. [13]	17.20%	50.76%	**70.77**%	82.61%	97.21%
Van Laere et al. [37]	**24.20**%	**51.49**%	63.27%	**85.62**%	**97.85**%
This work	**25.04**%	**53.53**%	**75.16**%	**87.21**%	**99.01**%

3.4.6 The influence of training data

For this final experiment, we use the optimal configuration of the framework discovered so far. We start with a training set of 1M photos, and gradually increase the size of the training set in steps of 1M photos, establishing a trade-off between the amount of training data used by the system and the results it achieves with it. The difference in results between each pair of configurations is statistically significant with a p-value $< 2.2 \times 10^{-16}$.

Figure 3.12 presents the results of this experiment in terms of median error distance. Similar to the conclusion in Section 3.4.5, the best result is achieved at a scale of $k = 2500$ areas. In this case, making use of the full 10M training items results in a median error distance of 1.06 km. It is interesting to note that the coarsest scale $k = 500$ performs equally well, with a median error of only 1.08 km. More generally, we can notice that adding more training data has a larger effect when the number of clusters is smaller. Due to the large amount of training data available, the similarity search within an area performs very well.

It is important to understand why a two-step approach to georeferencing is necessary. Using only a (global) search for similar images, we will soon run into trouble. If there is no training photo available that has a tag set that is almost equal to the one we are looking for, there is no way for the similarity search to differentiate among the tags (some tags provide strong geographical clues), treating them

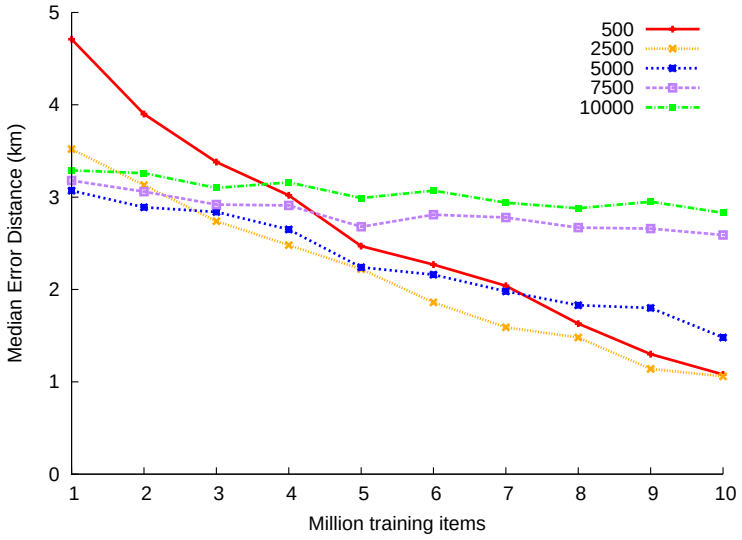


Figure 3.12: Median error distance of the 13 390 test items when estimating their locations at the 500, 2500, 5000, 7500 and 10 000 scales using an optimally tuned framework and a varying amount of training data.

Table 3.9: Detailed results in terms of accuracy at certain error distances and median error distance (in km) for the optimal results when using 10M training items, using the optimal configuration of the framework.

	Acc@1	Acc@10	Acc@100	Acc@1000	MER
1M	36.33	62.23	84.41	92.46	3.52
2M	38.40	63.81	84.71	**93.23**	3.13
3M	40.13	64.16	84.85	92.74	2.74
4M	41.78	65.29	84.93	92.78	2.48
5M	43.31	66.00	84.81	93.02	2.22
6M	45.18	66.42	85.22	92.86	1.86
7M	45.97	67.48	**85.32**	93.09	1.59
8M	46.80	67.50	84.73	92.67	1.48
9M	48.71	**69.37**	85.29	93.07	1.14
10M	**49.63**	68.96	85.08	93.22	**1.06**

all equally important. By starting the similarity search from the area that was obtained after classification, which implicitely resolves ambiguity among terms, this problem will likely be resolved in many cases.

As the amount of training data increases, it becomes more likely that a training photo will be present that largely resembles the tag set we are looking for, improving the effectiveness of a (global) similarity search. This effect is clearly visible in Figure 3.12 for the configuration using 500 clusters.

Also, as can be concluded from Table 3.9, a larger amount of training data enables the framework to improve the location estimations within the sub 10 kilometer range. If a developer is satisfied with an error distance of for example maximum 100 kilometer for an application, the results are largely independent of the amount of training data used.

3.5 Conclusions and future work

Converting the problem of georeferencing Flickr resources based on textual metadata into a classification problem is a popular approach in literature. After this initial classification step, a similarity search is performed in the area identified by the classifier. After a thorough experimental evaluation of this approach, we conclude the following:

- To achieve good results at sub-city scales (i.e. less than 10 kilometer of error distance), a similarity search component is essential.
- Information about the (home) location of the user is useful evidence for georeferencing Flickr resources.
- Among the clustering algorithms we have tested, k-medoids clustering performs best, due to its tendency to produce smaller scale clusters in areas of the world for which more training data is available.
- Applying a feature selection technique that is able to exploit the geographical aspect of the underlying data outperforms traditional methods.
- If we increase the amount of training data, the optimal number of clusters decreases due to an improved similarity search. Also, using more training data substanially improves accuracy in locating items within 10 km from their true location, while the results at an error margin of 100 km or 1000 km remain rather constant.

We see a number of opportunities for future work. Current approaches to georeferencing train models on the same type of data as the resources for which a location needs to be found. We believe that the language models trained from Flickr

can be successfully used to estimate locations for other types of textual resources, without the need for a gazetteer. Initial experiments in [7] show promising results to this end. Second, as has been demonstrated in the experiments in Section 3.4.3, using an appropriate feature selection method is essential. Although the geographical spread filtering method introduced in [13] is a good example of a method that takes the spatial distribution of the tags into account, we believe that there is still scope for improvement in this aspect. Next, in our current approach, all features are weighted equally in the similarity search step. It is clear that not all available features associated with a Flickr photo have an equal importance. Research should be carried out to find similarity measures that better reflect this than the Jaccard measure. Further, there may be other sources of information that could provide additional evidence for georeferencing Flickr resources. For example, intuitively it seems clear that in one way or another, gazetteers may help to improve the results, although a good way for disambiguating tags would be needed. Another idea is to use the timestamp of a photo in combination with some visual features to find out during what moment of the day a photo was taken (e.g. night, midday, or in between) may help us to narrow the possible locations down to a number of time zones. Finally, current georeferencing approaches focus on returning a specific location for each query, although this is not meaningful in all cases. If the only tag available for a photo is "France", it makes more sense to return the boundaries of the country instead of a pre-defined geographical coordinate in the city centre of Paris. As a partial solution to this problem, [38] introduces a method to automatically identify what is the most appropriate level of granularity at which a photo should be localized.

Acknowledgements The authors would like to thank Claudia Hauff for her help on the implementation of the geographical spread feature selection technique she presented at the MediaEval2011 workshop. We would also like to thank the organizers of the MediaEval workshop and the Placing Task organizers in particular for providing us with the Flickr dataset.

References

[1] S. Ahern, M. Naaman, R. Nair, J. H.-I. Yang, World explorer: visualizing aggregate data from unstructured text in geo-referenced collections, in: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, 2007, pp. 1–10.

[2] L. Backstrom, J. Kleinberg, R. Kumar, J. Novak, Spatial variation in search engine queries, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 357–366.

[3] Z. Cheng, J. Caverlee, K. Lee, You are where you tweet: a content-based approach to geo-locating Twitter users, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010, pp. 759–768.

[4] J. Choi, H. Lei, G. Friedland, The 2011 icsi video location estimation system, in: Working Notes of the MediaEval Workshop, Pisa, Italy, September 1-2, 2011, CEUR-WS.org, ISSN 1613-0073, online http://ceur-ws.org/Vol-807/Choi_ICSI_Placing_me11wn.pdf, 2011.

[5] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 603–619.

[6] D. J. Crandall, L. Backstrom, D. Huttenlocher, J. Kleinberg, Mapping the world's photos, in: Proceedings of the 18th International Conference on World Wide Web, 2009, pp. 761–770.

[7] C. De Rouck, O. Van Laere, S. Schockaert, B. Dhoedt, Georeferencing Wikipedia pages using language models from Flickr, in: Proceedings of the Terra Cognita 2011 Workshop, 2011, pp. 3–10.

[8] T. Dunning, Accurate methods for the statistics of surprise and coincidence, Comput. Linguist. 19 (1) (1993) 61–74.

[9] J. Eisenstein, B. O'Connor, N. A. Smith, E. P. Xing, A latent variable model for geographic lexical variation, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1277–1287.

[10] D. Ferres, H. Rodriguez, Talp at mediaeval 2011 placing task: Georeferencing flickr videos with geographical knowledge and information retrieval, in: Working Notes of the MediaEval Workshop, Pisa, Italy, September 1-2, 2011, CEUR-WS.org, ISSN 1613-0073, online http://ceur-ws.org/Vol-807/Ferres_UPC_Placing_me11wn.pdf, 2011.

[11] G. Friedland, J. Choi, A. Janin, Video2gps: a demo of multimodal location estimation on flickr videos, in: Proceedings of the 19th ACM international conference on Multimedia, 2011, pp. 833–834.

[12] M. Goodchild, Citizens as sensors: the world of volunteered geography, GeoJournal 69 (2007) 211–221.

[13] C. Hauff, G.-J. Houben, WISTUD at MediaEval 2011: Placing Task, in: Working Notes of the MediaEval Workshop, Pisa, Italy, September 1-2, 2011, CEUR-WS.org, ISSN 1613-0073, online http://ceur-ws.org/Vol-807/Hauff_WISTUD_Placing_me11wn.pdf.

[14] C. Hauff, G.-J. Houben, Placing images on the world map: a microblog-based enrichment approach, in: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 691–700.

[15] J. H. Hays, A. A. Efros, IM2GPS: Estimating geographic information from a single image, in: Proceedings of the 21st IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[16] B. Hecht, M. Raubal, GeoSR: Geographically explore semantic relations in world knowledge, in: L. Bernard, A. Friis-Christensen, H. Pundt (eds.), 11th AGILE International Conference on Geographic Information Science, 2008, pp. 95–114.

[17] L. Hollenstein, Capturing vernacular geography from georeferenced tags, Master's thesis, University of Zurich (2008).

[18] C. B. Jones, A. I. Abdelmoty, D. Finch, G. Fu, S. Vaid, The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing, in: Proceedings of the Third International Conference on Geographic Information Science, 2004, pp. 125–139.

[19] L. Kennedy, M. Naaman, Generating diverse and representative image search results for landmarks, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 297–306.

[20] F. Krippner, G. Meier, J. Hartmann, R. Knauf, Placing Media Items Using the Xtrieval Framework, in: Working Notes of the MediaEval Workshop, Pisa, Italy, September 1-2, 2011, CEUR-WS.org, ISSN 1613-0073, online http://ceur-ws.org/Vol-807/Krippner_CUT_Placing_me11wn.pdf.

[21] L. T. Li, J. Almeida, R. da S. Torres, Recod working notes for placing task mediaeval 2011, in: Working Notes of the MediaEval Workshop, Pisa, Italy,

September 1-2, 2011, CEUR-WS.org, ISSN 1613-0073, online http://ceur-ws.org/Vol-807/Li_UNICAMP_Placing_me11wn.pdf, 2011.

[22] M. D. Lieberman, H. Samet, J. Sankaranayananan, Geotagging: using proximity, sibling, and prominence clues to understand comma groups, in: Proceedings of the 6th Workshop on Geographic Information Retrieval, 2010, pp. 6:1–6:8.

[23] E. Moxley, J. Kleban, B. Manjunath, Spirittagger: a geo-aware tag suggestion tool mined from Flickr, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, 2008, pp. 24–30.

[24] O. A. B. Penatti, L. T. Li, J. Almeida, R. da S. Torres, A visual approach for video geocoding using bag-of-scenes, in: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, 2012, pp. 53:1–53:8.

[25] A. Popescu, I. Kanellos, Creating visual summaries for geographic regions, in: IR+SN Workshop (at ECIR), 2009.

[26] A. Rae, V. Murdock, P. Serdyukov, P. Kelm, Working Notes for the Placing Task at MediaEval2011, in: Working Notes of the MediaEval Workshop, Pisa, Italy, September 1-2, 2011, CEUR-WS.org, ISSN 1613-0073, online http://ceur-ws.org/Vol-807/Rae_Placing_me11overview.pdf.

[27] T. Rattenbury, N. Good, M. Naaman, Towards automatic extraction of event and place semantics from flickr tags, in: Proceedings of the 30th Annual International ACM SIGIR Conference, 2007, pp. 103–110.

[28] T. Rattenbury, M. Naaman, Methods for extracting place semantics from Flickr tags, ACM Transactions on the Web 3 (1) (2009) 1–30.

[29] D. Reynold, Gaussian mixture models, Tech. rep., MIT Lincoln Laboratory (2008).

[30] P. Schmitz, Inducing ontology from Flickr tags, in: Proceedings of the Collaborative Web Tagging Workshop, 2006, pp. 210–214.

[31] S. Schockaert, M. De Cock, Neighborhood restrictions in geographic IR, in: Proceedings of the 30th Annual International ACM SIGIR Conference, 2007, pp. 167–174.

[32] P. Serdyukov, V. Murdock, R. van Zwol, Placing Flickr photos on a map, in: Proceedings of the 32nd Annual International ACM SIGIR Conference, 2009, pp. 484–491.

[33] M. D. Smucker, J. Allan, An investigation of Dirichlet prior smoothing's performance advantage, Tech. Rep. IR-445, University of Massachusetts (2005).

[34] O. Van Laere, S. Schockaert, B. Dhoedt, Combining multi-resolution evidence for georeferencing Flickr images, in: Proceedings of the 4th International Conference on Scalable Uncertainty Management, 2010, pp. 347–360.

[35] O. Van Laere, S. Schockaert, B. Dhoedt, Towards automated georeferencing of flickr photos, in: Proceedings of the 6th Workshop on Geographic Information Retrieval, 2010, pp. 5:1–5:7.

[36] O. Van Laere, S. Schockaert, B. Dhoedt, Finding locations of Flickr resources using language models and similarity search, in: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, 2011, pp. 48:1–48:8.

[37] O. Van Laere, S. Schockaert, B. Dhoedt, Ghent university at the 2011 Placing Task, in: Working Notes of the MediaEval Workshop, 2011.

[38] O. Van Laere, S. Schockaert, B. Dhoedt, Georeferencing flickr photos using language models at different levels of granularity: An evidence based approach, Web Semantics: Science, Services and Agents on the World Wide Web.

[39] F. Wilske, Approximation of neighborhood boundaries using collaborative tagging systems, in: Proceedings of the GI-Days, 2008, pp. 179–187.

[40] B. Wing, J. Baldridge, Simple supervised document geolocation with geodesic grids, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 955–964.

[41] Y. Yang, Z. Gong, L. H. U, Identifying points of interest by self-tuning clustering, in: Proceedings of the 34th International ACM SIGIR Conference, 2011, pp. 883–892.

[42] Y. Yang, J. O. Pedersen, A comparative study on feature selection in text categorization, in: Proceedings of the 14th International Conference on Machine Learning, 1997, pp. 412–420.

[43] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to Ad Hoc information retrieval, in: Proceedings of the 24th Annual International ACM SIGIR Conference, 2001, pp. 334–342.

[44] Y.-T. Zheng, Z.-J. Zha, T.-S. Chua, Research and applications on georeferenced multimedia: a survey, Multimedia Tools and Applications 51 (2011) 77–98.

4

Finding locations of Flickr resources using language models and similarity search

This chapter was originally published as our first conference paper that discusses our two step approach to the problem of georeferencing. In this dissertation, Chapter 3 already provides an in depth-discussion of this georeferencing process, deprecating a number of ideas in this chapter. However, this paper is included as it provides valuable insights in the effect of user specific tagging on the overall georeferencing process. Results are presented that demonstrate that feature selection is necessary to remove noisy terms from the overall vocabulary, as the optimal results for the experiments carried out in this chapter are obtained with less than all features. Furthermore, the results in this chapter show the effect of the number of tags in relation to the error made in georeferencing photos, which clearly indicate that as soon as 4 tags are present, good location estimations can be made.

⋆ ⋆ ⋆

Olivier Van Laere, Steven Schockaert, Bart Dhoedt

Published in the Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR), Trento, Italy, April 2011.

Abstract We present a two-step approach to estimate where a given photo or video was taken, using only the tags that a user has assigned to it. In the first step, a language modeling approach is adopted to find the area which most likely contains the geographic location of the resource. In the subsequent second step, a precise location is determined within the area that was found to be most plausible. The main idea of this step is to compare the multimedia object under consideration with resources from the training set, for which the exact coordinates are known, and which were taken in that area. Our final estimation is then determined as a function of the coordinates of the most similar among these resources. Experimental results show this two-step approach to improve substantially over either language models or similarity search alone.

4.1 Introduction

Web 2.0 systems such as Flickr bring structure in collections of shared multimedia objects by taking advantage of both structured and unstructured forms of metadata. Unstructured metadata is mainly available in the form of tags, i.e. short (but otherwise unconstrained) textual descriptions that are provided by users, although in the case of Flickr, only owners can add tags. Such tags help users to organize the resources they find interesting or to otherwise facilitate retrieval of such resources (by themselves or by others) in the future [1]. In the case of photos and videos, most of the structured metadata is provided automatically by the camera, without any involvement of the user. These types of metadata usually include the type of camera, the settings that were used (e.g. aperture, focal distance, etc.) and the time and date. In a limited number of cases, cameras also provide geographic coordinates, using a built-in or externally connected GPS device. Flickr additionally offers the possibility of manually indicating on a map where a photo was taken.

The availability of location metadata is important for at least two reasons. First, it allows users to easily retrieve photos or videos that were taken at a particular location, e.g. by explicitly supporting spatial constraints in queries [2], or by displaying the resources on a map which users can explore [3]. Second, by analyzing the correlation between geographic location and the occurrence of certain tags, we may discover geographic knowledge beyond what is usually described in gazetteers [4, 5]. As a result of these considerations, and given that only a small fraction of Flickr resources are currently geo-annotated, there has been a recent interest in techniques that could automatically estimate the geographic location of photos and videos [6]. More generally, there seems to be a trend towards leveraging user-contributed, unstructured information to structured, semantic annotations, e.g. automatically completing Wikipedia infoboxes [7] or building ontologies from user tags [8].

Several kinds of information are available to estimate the geographic location

of a photo or video: visual features, user profiles, and tags. Visual features may be useful to recognize certain types of landmarks, or to differentiate photo or videos that were taken e.g. at the beach from resources taken in a city center. In general, however, visual information alone is not likely to be sufficient for determining a specific location. Similarly, user profiles may be useful to introduce a bias (e.g. users are more likely to take photos closer to the place where they live), but they do not contain sufficient information to pinpoint where a photo or video was taken. In this paper, we exclusively focus on the third type of available information, viz. the tags associated with a resource. Indeed, before the value of visual features or user profiles for this task can be assessed, in our opinion, a more thorough understanding is needed of the kind of geographic information that can be extracted from tags.

To estimate the location of a multimedia object based on its tags, three natural strategies present themselves. First, we may use gazetteers to find the locations of those tags that correspond to toponyms. Although intuitive, this strategy has proven to be particularly challenging in practice, among others due to the fact that no capitalization occurs in tags, making it difficult to identify the toponyms (e.g. *nice* vs. *Nice*), as well as due to the high ambiguity of toponyms and the limited amount of context information that is available for disambiguation. Second, we may interpret the problem of georeferencing as a classification problem, by partitioning the locations on earth into a finite number of areas. Standard language modeling approaches can then be used to determine the most likely area for a given resource, represented as its set of tags. This method eliminates the problem of determining which tags are toponyms, or any form of (explicit) disambiguation. A drawback, however, is that it results in an entire area, rather than a precise coordinate. The more areas in the partition, the more fine-grained our conclusion will be, but, the higher the chances of classification error become. Third, we may resort to similarity search, and estimate the location of a given resource as a weighted average of the locations of the most similar objects in our training set, e.g. using a form of similarity that is based on the overlap between tag sets. In this case, we do obtain precise coordinates, but the performance of the method may be limited by the fact that it treats spatially relevant tags in the same way as others. For instance, a resource tagged with *paris,bridge* will be considered as similar to a resource tagged with *london,bridge* as to a resource tagged with *paris,cathedral*. In this paper, we propose to combine the best of the latter two strategies: first use a classifier to find the most likely area in which a photo or video was taken, and then use similarity search to find the most likely location within that area.

We have participated in the Placing Task of the 2010 MediaEval benchmarking initiative [6] using a system based on this two-step approach. Our system came out best, localizing about 44% of the videos in the test collection within 1km of their true location. In this paper, we present the details of our system, and we

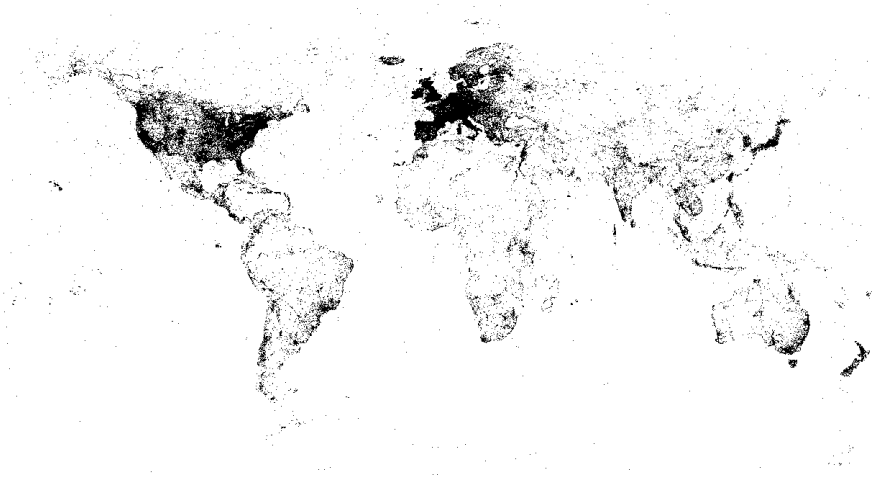


Figure 4.1: Plot of all the photos in the training set

analyze which aspects are responsible for its performance, focusing on two crucial points. First, we stress the importance of combining classification (e.g. using language models) with interpolation (e.g. using similarity search), revealing that neither method alone is capable of producing equally good results. Second, we analyze the influence of user-specific tags. In particular, in case of the Placing Task, it turns out that most of the users that own a video from the test collection also own one or more photos in the training data: among the 4576 test videos with at least one tag, 923 different users appear of whom 873 own at least one photo in the training set. We analyze to what extent the availability of such previous geo-annotations by the same user influences the performance of the system.

The paper is structured as follows. First, we detail the nature of the data sets that have been used, as well as the preprocessing methods that were applied. The subsequent two sections individually consider the two strategies that lie at the basis of our system: finding the most plausible area, using a standard language modeling approach, and finding the most likely location within that area, using similarity search. Next, in Section 4.5 we explain how these two methods can be combined, and show that this combination performs better than the two components on which it is based. Finally, we provide an overview of related work and conclude.

4.2 Data acquisition and preprocessing

As training data, we used a collection of 8 685 711 photos, containing the 3 185 258 georeferenced Flickr photos that were provided to participants of the Placing

Task, together with an additional crawl of 5 500 368 georeferenced Flickr photos. In addition to the coordinates themselves, Flickr provides information about the accuracy of coordinates as a number between 1 (world-level) and 16 (street level). When crawling the additional data, we only crawled Flickr photos having an accuracy of at least 12, to ensure that all coordinates were meaningful w.r.t. within-city location. Once retrieved, photos that did not contain tags or valid coordinates were removed from the collection. Next, we ensured that at most one photo was retained in the collection with a given tag set and user name, in order to reduce the impact of bulk uploads [3]. Once filtered, the remaining dataset contained 3 271 022 photos. A visual representation of this dataset is shown in Figure 4.1.

The test videos provided for the Placing Task contain videos that are part of bulk uploads, in the sense that some videos were uploaded around the same time as some photos in the training set by the same user, often resulting in a very high similarity between the tag set of the corresponding videos and photos. To avoid any undesirable effects of bulk uploads in our evaluation, we crawled a collection of 10 000 Flickr videos that have been uploaded later than the most recent photo from the training set. We furthermore restricted ourselves to videos with an accuracy level of 16, to ensure that our gold standard was as accurate as possible. This data set was then split into 7 400 videos that are owned by a user who also has at least one photo in our training set, and 2 600 videos by users who do not appear in the training set.

Next, the locations of the photos in the training set were clustered in a set of disjoint areas $\mathcal{A}$ using the k-medoids algorithm with geodesic distance, considering a varying number of clusters k. We consider ten different resolutions and thus ten different sets of areas $\mathcal{A}_k$. The datasets were clustered into 50, 500, 2 500, 5 000, 7 500, 10 000, 12 500, 15 000, 17 500 and 20 000 disjoint areas respectively.

Subsequently, a vocabulary V consisting of 'interesting' tags is compiled, which are tags that are likely to be indicative of geographic location. We used χ^2 feature selection to determine for each area in $\mathcal{A}$ the m most important tags. Let $\mathcal{A}$ be the set of areas that is obtained after clustering. Then for each area a in $\mathcal{A}$ and each tag t occurring in photos from a, the χ^2 statistic is given by:

$$\chi^2(a,t) = \frac{(O_{ta} - E_{ta})^2}{E_{ta}} + \frac{(O_{t\bar{a}} - E_{t\bar{a}})^2}{E_{t\bar{a}}} + \frac{(O_{\bar{t}a} - E_{\bar{t}a})^2}{E_{\bar{t}a}} + \frac{(O_{\bar{t}\bar{a}} - E_{\bar{t}\bar{a}})^2}{E_{\bar{t}\bar{a}}}$$

where O_{ta} is the number of photos in area a in which tag t occurs, $O_{t\bar{a}}$ is the number of photos outside area a in which tag t occurs, $O_{\bar{t}a}$ is the number of photos in area a in which tag t does not occur, and $O_{\bar{t}\bar{a}}$ is the number of photos outside area a in which tag t does not occur. Furthermore, E_{ta} is the number of occurrences of tag t in photos of area a that could be expected if occurrence of t were independent

of the location in area a, i.e. $E_{ta} = N \cdot P(t) \cdot P(a)$ with N the total number of photos, $P(t)$ the percentage of photos containing tag t and $P(a)$ the percentage of photos that are located in area a, i.e.:

$$P(a) = \frac{|X_a|}{\sum_{b \in \mathcal{A}} |X_b|} \tag{4.1}$$

where, for each area $a \in \mathcal{A}$, we write X_a to denote the set of images from our training set that were taken in area a. Similarly, $E_{t\overline{a}} = N \cdot P(t) \cdot (1 - P(a))$, $E_{\overline{t}a} = N \cdot (1 - P(t)) \cdot P(a)$, $E_{\overline{t}\overline{a}} = N \cdot (1 - P(t)) \cdot (1 - P(a))$. The vocabulary V was then obtained by taking for each area a, the m tags with highest χ^2 value. In the default configuration of our system, the m values are 640 000 for the coarsest clustering, 6 400, 256, 64, 28, 16, 10, 7, 5 for the intermediate resolutions and 4 for the finest clustering level. This choice of features ensures that the language models, introduced next, require approximately the same amount[1] of space for each clustering level. In Section 4.6, we will analyze the impact of the choice of the m values.

4.3 Language models

4.3.1 Outline

Given a previously unseen resource x, we try to determine in which area x was most likely taken by comparing its tags with those of the images in the training set. Specifically, using standard generative unigram language modeling, the probability of area a, given the tags that are available for resource x is given by

$$P(a|x) \propto P(a) \cdot \prod_{t \in x} P(t|a) \tag{4.2}$$

where we identify the resource x with its set of tags. The prior probability $P(a)$ of area a can be estimated using maximum likelihood, as in (4.1), which means that in absence of other information, resources are assigned to the area containing the largest number of photos from the training set. To obtain a reliable estimate of $P(t|a)$, some form of smoothing is needed, to avoid a zero probability when x is associated with a tag that does not occur with any of the photos in area a from the training set. We have experimented with Laplace smoothing, Jelinek-Mercer smoothing, and Bayesian smoothing with Dirichlet priors, the latter yielding the best results in general (with Jelink-Mercer producing similar results). These findings conform to experimental results in other areas of information retrieval [9], and to earlier work on georeferencing Flickr photos [3]. Specifically, using Bayesian

[1] Space requirements increase quadratically with the number of clusters.

smoothing with Dirichlet priors, we take:

$$P(t|a) = \frac{O_{ta} + \mu \left(\frac{\sum_{a' \in \mathcal{A}} O_{ta'}}{\sum_{a' \in \mathcal{A}} \sum_{t' \in V} O_{t'a'}} \right)}{(\sum_{t' \in V} O_{t'a}) + \mu}$$

where, as before, we write O_{ta} for the number of occurrences of term t in area a, and V is the vocabulary (after feature selection). The parameter μ takes a value in $]0, +\infty[$ and was set to 1750, although good results were found for a large range of values. The area a_x assigned to resource x is then the area maximizing the right-hand side of (4.2):

$$a_x = \arg\max_{a \in \mathcal{A}} P(a) \cdot \prod_{t \in x} P(t|a) \tag{4.3}$$

Thus an area is found which is assumed to contain the true location of x. It may be useful to convert this area to a precise location, e.g. for comparison with other methods. To this end, an area a can be represented as its medoid $med(a)$:

$$med(a) = \arg\min_{x \in a} \sum_{y \in a} d(x, y) \tag{4.4}$$

where $d(x, y)$ represents the geodesic distance. Note that the medoid is the most central element of an area, rather than its center-of-gravity. The latter is avoided here because it is too sensitive to outliers.

4.3.2 Experimental results

Whether or not (4.4) provides a good estimation depends on the number of clusters that are considered. If this number is too small, the clusters will be too coarse, and the medoid will not be a good approximation of the true location in general. If this number is too large, however, the chances of classification error increase. Thus there is a trade-off to be found, as can clearly be seen in Figure 4.2. This figure depicts the median error that was obtained for a variety of cluster sizes, i.e. the median of the geodesic distance between the medoid of the cluster that was found by (4.3) and the true location. The figure reports the results of three experimental set-ups: one experiment considers the 7 400 videos whose owner appears in our training set (*Overlap*), one experiment considers the results for these same videos when the photos from these video owners have been excluded from the training set (*Filtered*), and one experiment considers the 2 600 videos whose owners are distinct from the owners of the photos in the (complete) training set (*Distinct*).

Regarding the influence of previously geo-annotated photos by the same user, the bad performance of the *Distinct* experiment is particularly noticeable. Closer inspection of the results has revealed that the bad results are to a large extent due

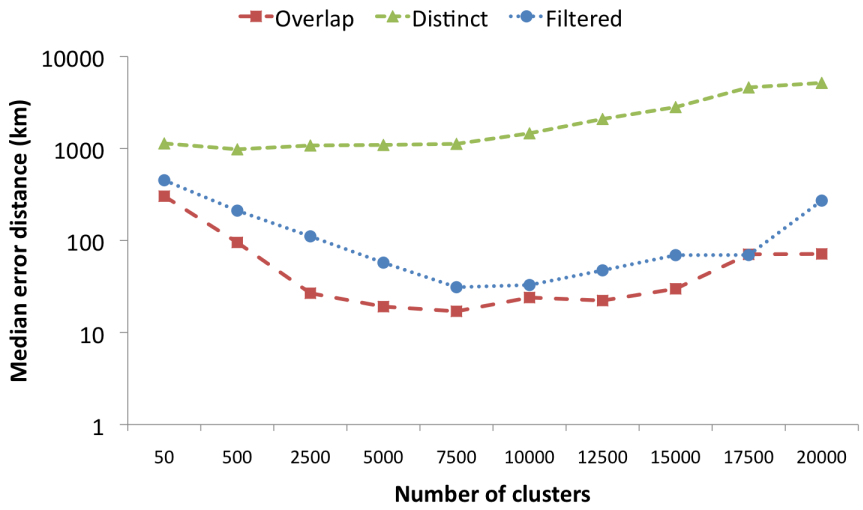


Figure 4.2: Median error between the medoid of the found cluster and the true location of the videos in the test set.

to the fact that the videos in the corresponding test set have less (and less informative) tags. For instance, while the average number of tags per video is 9.39 for the *Overlap* experiment, we only have 5.92 tags on average for the videos of the *Distinct* experiment. We may speculate that users owning a larger number of resources tend to put more effort in accurately tagging these resources. As the users of the videos in the *Overlap* experiment own photos as well as videos, they are more likely to belong to this latter category. The *Filtered* experiment confirms this intuition, showing that the mere lack of geo-annotated objects by the same user has a much milder impact, although the optimal median error is still worse by almost a factor two. This suggests that the number of (good) tags has a much stronger influence than the presence or absence of geo-annotated objects by the same user. To test this hypothesis, we have separately evaluated those videos that contain a given number of tags, starting from a combined test set containing all 10 000 videos. The results, which are shown in Figure 4.3, clearly show that videos with more tags also tend to contain more descriptive tags and can therefore be more accurately localized. However, the results for videos with more than 10 tags are, somewhat surprisingly, worse than those for videos with 6 to 10 tags. This appears to be due to the fact that among the videos with more than 10 tags, many contain tags that have not been manually added, e.g. *taxonomy:phylum=chordata*. In particular, we found that 9.25% of all tag occurrences contain a ':' in the [11,75] category, as opposed to 0.45% in the [6,10] category. Clearly, the assumption that the number of tags provides an indication of how much effort the user has spent to

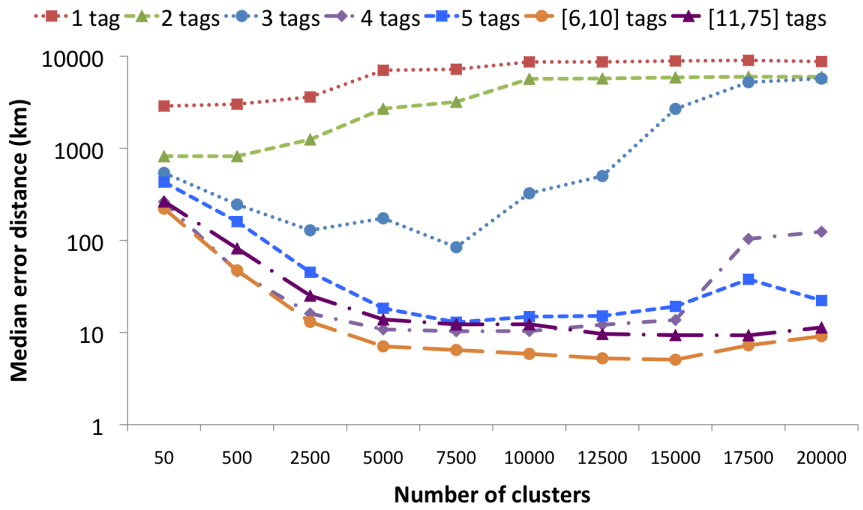


Figure 4.3: Median error between the medoid of the found cluster and the true location, each time using all test videos containing a given number of tags.

describe the video no longer applies when tags are added automatically. Figure 4.4 shows that the same conclusions can be drawn, when restricted to the videos from the *Distinct* set-up, providing evidence that it is indeed the lack of appropriate tags that cause the overall results of the *Distinct* and *Overlap* configurations to be so different.

For the *Overlap* experiment, the optimal median error of 17.02 km is obtained when using 7 500 clusters, for the *Distinct* experiment, the optimal median error of 979.86 km is obtained when using 500 clusters, and for the *Filtered* experiment, we again need 7 500 clusters to obtain the optimal median error of 31.10 km. The lower optimal number of clusters in the case of *Distinct* suggests that the less informative the tags of a given video are, the coarser the clustering should ideally be. This is also confirmed by the results in Figure 4.3 which show the optimal number of clusters to be 50 for photos with 1 tag (2876.46 km), 500 for photos with 2 tags (820.61 km), 7 500 for photos with 3 (84.33 km), 4 (10.32 km), or 5 (12.92 km) tags, 15 000 for photos with 6 to 10 tags (5.07 km), and 17 500 for photos with more than 10 tags (9.33 km).

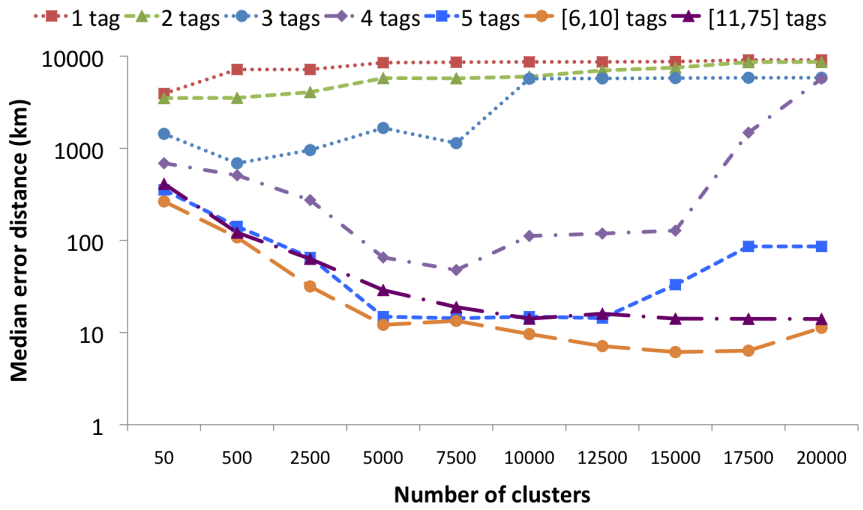


Figure 4.4: Median error between the medoid of the found cluster and the true location, using only the test videos from the Distinct *set-up containing a given number of tags.*

4.4 Similarity search

4.4.1 Outline

Rather than converting the problem at hand to a classification problem, a more direct strategy to find the location of a photo or video x consists of identifying the photos from the training set that are most similar to x, and estimate the location of x by averaging these locations. Specifically, let $y_1, ..., y_k$ be the k most similar photos from our training set. We then propose to estimate the location of x as a weighted center-of-gravity of the locations of $y_1, ..., y_k$:

$$loc(x) = \frac{1}{k} \sum_{i=1}^{k} sim(x, y_i)^{\alpha} \cdot loc(y_i) \tag{4.5}$$

where the parameter $\alpha \in]0, +\infty[$ determines how strongly the result is influenced by the most similar photos only. The similarity $sim(x, y_i)$ between resources x and y_i was quantified using the Jaccard measure:

$$s_{jacc}(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

where we identify a resource with its set of tags *without feature selection*. In principle, Jaccard similarity may be combined with other types of similarity, e.g. based on visual features.

In (4.5), locations are assumed to be represented as Cartesian (x, y, z) coordinates rather than as (lat, lon) pairs[2]. In practice, we thus need to convert the (lat_i, lon_i) coordinates of each photo y_i to its Cartesian coordinates:

$$x_i = cos(lat_i) \cdot cos(lon_i)$$
$$y_i = cos(lat_i) \cdot sin(lon_i)$$
$$z_i = sin(lat_i)$$

Subsequently, the right-hand side of (4.5) is evaluated, yielding a point (x^*, y^*, z^*), which is usually not on the surface of the earth. Unless this point is exactly the center of the earth, its latitude lat^* and longitude lon^* can be determined:

$$lat^* = atan2(z^*, \sqrt{x^{*2} + y^{*2}})$$
$$lon^* = atan2(y^*, x^*)$$

In addition to the choice of the parameter α, the performance of (4.5) depends on the set of resources R_x that is considered when determining the k most similar photos $y_1, ..., y_k$. In principle, we could take R_x to be the entire training set. However, we also experiment with putting a threshold on the similarity with x, considering in R_x only those resources that are sufficiently similar. This restriction is motivated by the fact that center-of-gravity methods are sensitive to outliers. Note that using medoids to alleviate the influence of outliers is not appropriate when the number of points is small. Also note that as a result of this restriction, sometimes less than k similar photos may be used. In each case R_x will contain the most similar photo, even if its similarity is below the threshold. Other photos are added only if they are sufficiently similar.

4.4.2 Experimental results

Three parameters influence the performance of the similarity search: the threshold on the similarity with the object to be classified, the number k of similar photos to consider, and the exponent α in (4.5). Table 4.1 displays the result for different choices of the threshold on similarity, and different values of k, in case of the *Overlap* configuration. Regarding the similarity threshold, we find that a small threshold of 0.05 slightly improves the results for the smaller values of k. Indeed, the smaller the value of k, the more the result may be influenced by outliers, and the more important it thus becomes to avoid them. Surprisingly, small values of k appear to be better than larger values, although the optimal choice $k = 2$ is substantially better than $k = 1$.

[2] See http://www.geomidpoint.com/calculation.html for an explanation of this coordinate transformation, and a comparison with alternative methods to calculate "average locations".

Table 4.1: Influence of the similarity threshold on the median error distance for the Overlap *configuration (using an exponent α of 1).*

threshold	0	0.05	0.10	0.15	0.20
1	2528	2528	2528	2528	2528
2	477	424	477	773	1150
3	685	604	662	880	1150
4	748	741	773	899	1181
5	790	821	835	952	1242
6	824	799	837	954	1238
7	808	823	850	961	1247
8	843	829	856	980	1246
9	855	856	871	971	1242
10	863	868	872	968	1243

k

Table 4.2: Influence of the exponent α on the median error distance for the Overlap *configuration (using a similarity threshold of 0.05).*

α	1	25	50	75	100
1	2528	2528	2528	2528	2528
2	424	343	341	341	341
3	604	435	413	411	410
4	741	417	383	370	370
5	821	410	368	350	349
6	799	419	399	395	393
7	823	422	400	395	395
8	829	440	427	420	419
9	856	459	450	441	440
10	868	475	459	451	449

k

Tables 4.2 illustrates the influence of varying the exponent α in (4.5), where we take the similarity threshold fixed at 0.05. Choosing a higher value of α essentially serves the same purpose as choosing a higher similarity threshold, i.e. reducing the impact of potential outliers on the result. We can observe that higher values of α tend to produce better results. Again the choice of $k = 2$ turns out to be optimal.

In general, it seems that similarity search performs a lot worse than the language models, yielding an optimal error of 340.69 km, as opposed to 17.02 km in the case of language models. Similar effects are witnessed for the *Distinct* and *Filtered* configurations (not shown), where we respectively find an optimal error of 1302.95 km (instead of 979.86 km) and 578.22 km (instead of 31.10 km). However, as we will see in the next section, when combined with the language models, similarity search may be of great value.

4.5 A hybrid approach

4.5.1 Outline

The two methods that have been presented in the previous sections can be combined in a natural way: first an area is determined using the language modeling approach from Section 4.3 and then the similarity based method from Section 4.4 is applied, but restricted to the photos in the found area. When no photo in the clustering satisfies the chosen similarity threshold, the medoid of the found cluster can be used instead. Thus, we may take advantage of the language modeling's ability to implicitly discriminate between occurrences of more and less important tags, while keeping the advantage of the similarity search that a precise coordinate is obtained.

A second extension is related to choosing the right number of clusters. In particular, when we discover that a given resource has no tag in common with the vocabulary of the chosen clustering, we fall-back to the next (coarser) clustering.[3] In this way, if a resource contains no tags that are indicative of a precise location (e.g. *eiffeltower*) but does contain some tags that define a larger-scale area (e.g. *france*), it may not have any tags in common with the vocabulary of the finest clusterings, but after falling-back to a coarser clustering, a suitable area can still be determined.

4.5.2 Experimental results

Figure 4.5 shows the median distance that is obtained when language models are combined with similarity search. Interestingly, we find that choosing $k = 1$ with

[3]Recall that in absence of any tags, without fall-back, the prior probabilities determine to which cluster a resource is assigned, hence the cluster containing the largest number of resources will be chosen.

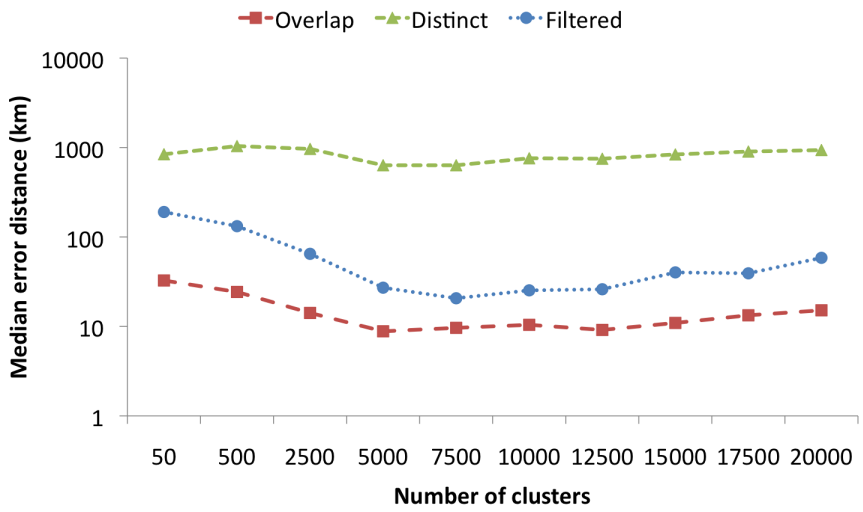


Figure 4.5: Median error obtained using the hybrid method with $k = 1$ and without a similarity threshold.

Table 4.3: Number of the test videos for which the location that was found is within a given distance of the true location.

	1km	5km	10km	50km	100km
Overlap (7 400)	2135	3362	3773	4500	4694
Distinct (2 600)	465	803	903	1066	1128
Filtered (7 400)	1428	2770	3248	4012	4265

similarity threshold 0 (shown in Figure 4.5) performs slightly better than choosing $k = 2$ with similarity threshold 0.05 (not shown), despite that the latter configuration is clearly better when similarity search is applied alone. This can be explained by the fact that within a cluster, all photos are relatively close to each other anyway, hence the problem of outliers is alleviated. As a result, the positive effect of filtering photos that are not sufficiently similar becomes counter-productive. A more detailed analysis of the results is presented in Table 4.3. For all three set-ups, a marked improvement is witnessed over the results of the language models from Section 4.3, the optimal results now being attained for 5 000 clusters in the case of the *Overlap* (8.82 km) and *Distinct* (633.36 km) set-ups, and for 7 500 clusters in the case of the *Filtered* (20.64 km) set-up.

Figures 4.6 and 4.7 provide a more detailed picture of the performance of our method, considering all test videos and those from the *Distinct* set-up respectively. As in Section 4.3, we find that the bad performance in the *Distinct* set-up can

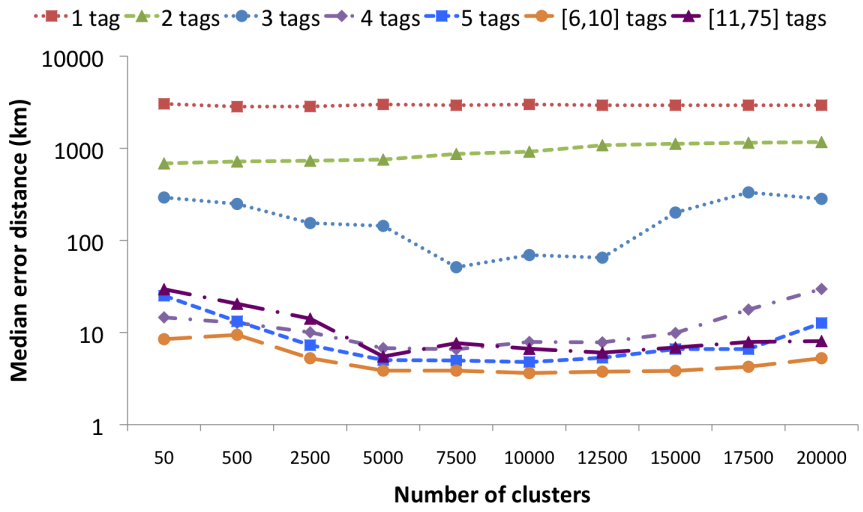


Figure 4.6: Median error between the medoid of the found cluster and the true location, each time using all test videos containing a given number of tags.

be attributed to the fewer number of videos with sufficient tags. In particular, if we only consider those videos with 6 to 10 tags (21.77% of the test videos), a median distance of 3.90 km is attained when using either 15 000 or 17 500 clusters. In case of the *Overlap* experiment (not shown), the median distance in the [6,10] range (23.19% of the test videos) is only slightly better, with 3.54 km being attained when using 10 000 clusters. These results indicate that rather precise coordinates can be found for most videos, provided that a sufficient number of (manually chosen) tags are available.

Finally we analyze the impact of feature selection. The purpose of feature selection is to eliminate all tags that are not spatially relevant, before the language models are built. This may be useful not only for speeding up calculations, but also to improve classification accuracy. Figures 4.8, 4.9 and 4.10 display how choosing a different number of features impacts the median error distance. The results for all features refers to the set of features that have been used in the experiments throughout the paper, applying χ^2 feature selection as explained in Section 4.2. The other results show what happens when only the best 25%, 50% and 75% of these features (according to the χ^2 statistic) are retained. The main observation is that the optimal value is quite robust w.r.t. the number of selected features. Only when a suboptimal number of clusters is chosen we find some differences, favoring fewer features for the coarser clusterings.

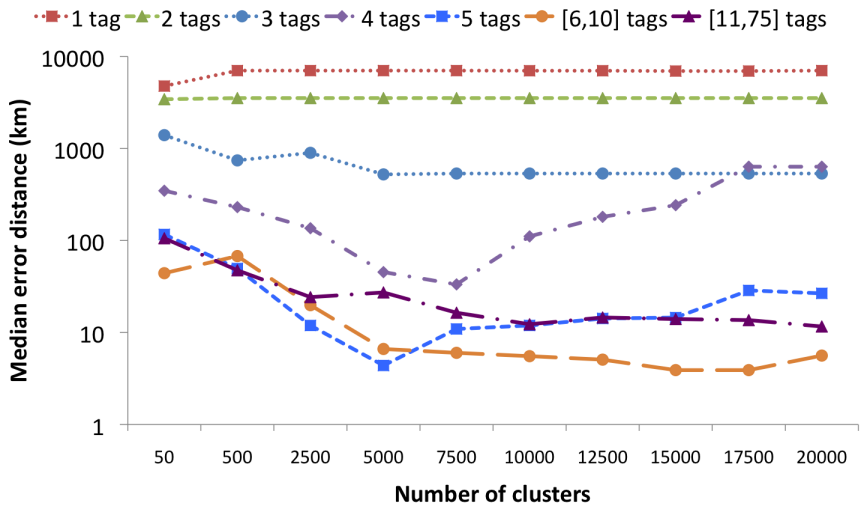


Figure 4.7: Median error between the medoid of the found cluster and the true location, using only the test videos from the Distinct *set-up containing a given number of tags.*

4.6 Related work

The related work falls in two categories: finding the geographic scope of resources, and using it when it is available.

Finding locations of tagged photos The task of deriving geographic coordinates for multimedia objects has recently gained in popularity. A recent benchmark evaluation of this task was carried out at MediaEval 2010 [6], where an earlier version of our system was shown to substantially outperform all other approaches. This result confirms and strengthens earlier support for using language models in this task [3].

Most existing approaches are based on clustering, in one way or another, to convert the task into a classification problem. For instance, in [10] target locations are determined using mean shift clustering, a non-parametric clustering technique from the field of image segmentation. To assign locations to new images, both visual (keypoints) and textual (tags) features were used. Experiments were carried out on a sample of over 30 million images, using both Bayesian classifiers and linear support vector machines, with slightly better results for the latter. Two different resolutions were considered corresponding to approximately 100 km (finding the correct metropolitan area) and 100 m (finding the correct landmark). It was found that visual features, when combined with textual features, substantially improve accuracy in the case of landmarks. In [11], an approach is presented which is

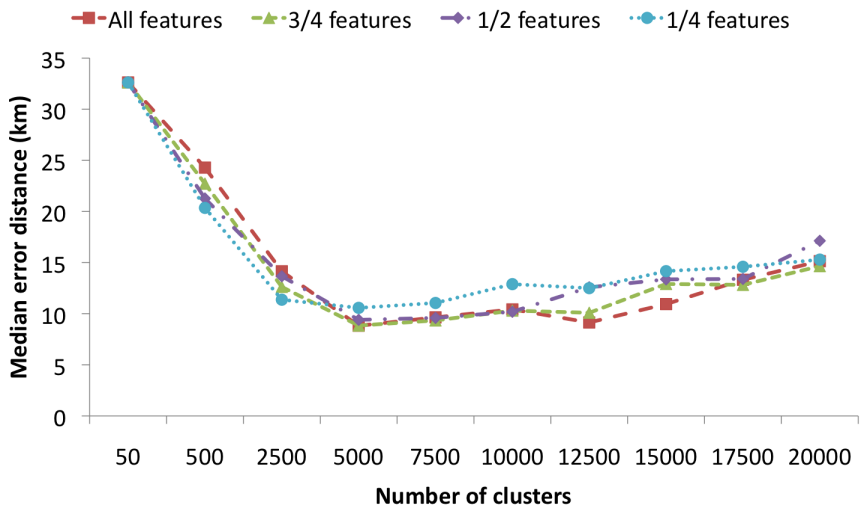


Figure 4.8: Impact of the amount of feature selection, in case of the Overlap *set-up*

based purely on visual features. For each new photo, the 120 most similar photos with known coordinates are determined. This weighted set of 120 locations is then interpreted as an estimate of a probability distribution, whose mode is determined using mean-shift clustering. The resulting value is used as prediction of the image's location.

Next, [3] investigates the idea that when georeferencing images, the spatial distribution of the classes (areas) could be utilized to improve accuracy. Their starting point is that typically, not only the correct area will receive a high probability, but also the areas surrounding the correct area. An appropriate adaptation of the standard language modeling approach is shown to yield a small, but statistically significant improvement.

Using locations of tagged photos When available, the coordinates of a photo may be useful for a variety of purposes. In [12], for instance, coordinates of tagged photos are used to find representative textual descriptions of different areas of the world. These descriptions are then put on a map to assist users in finding images that were taken in a given location of interest. The approach is based on spatially clustering a set of geotagged Flickr images, using k-means, and then relying on (an adaptation of) tf-idf weighting to find the most prominent tags of a given area. Similarly, [13] looks at the problem of suggesting useful tags, based on available coordinates. Some authors have looked at using geographic information to help diversify image retrieval results [14, 15].

Geotagged photos are also useful from a geographic perspective, to better un-

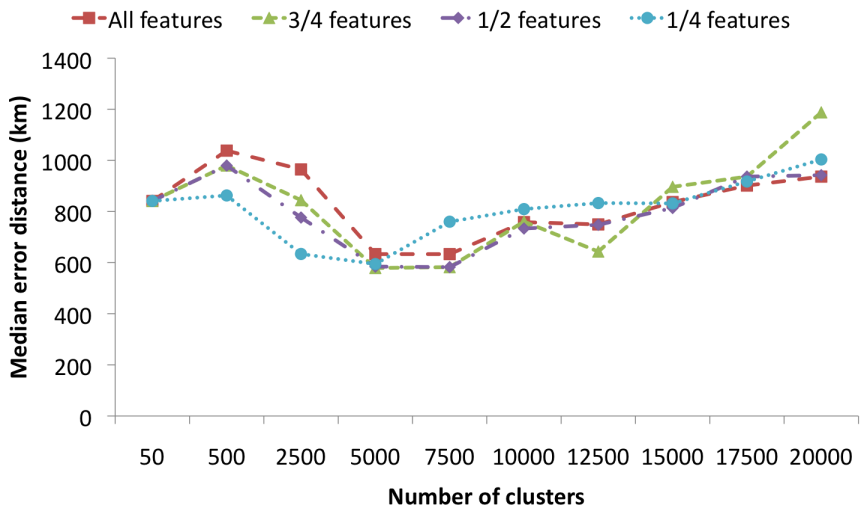


Figure 4.9: Impact of the amount of feature selection, in case of the Distinct *set-up*

derstand how people refer to places, and overcome the limitations and/or costs of existing mapping techniques [4]. For instance, by analyzing the tags of georeferenced photos, Hollenstein [5] found that the city toponym was by far the most essential reference type for specific locations. Moreover, [5] provides evidence suggesting that the average user has a rather distinct idea of specific places, their location and extent. Despite this tagging behaviour, Hollenstein concluded that the data available in the Flickr database meets the requirements to generate spatial footprints at a sub-city level. Finding such footprints for non-administrative regions (i.e. regions without officially defined boundaries) using georeferenced resources has also been addressed in [2] and [16]. Another problem of interest is the automated discovery of which names (or tags) correspond to places. Especially for vernacular place names, which typically do not appear in gazetteers, collaborative tagging-based systems may be a rich source of information. In [17], methods based on burst-analysis are proposed for extracting place names from Flickr.

4.7 Concluding remarks

We have advocated a two-step approach for georeferencing tagged multimedia objects. In the first step, the task of finding suitable geographic coordinates is treated as a classification problem, where the classes are areas that have been obtained by clustering the locations of the objects in the training set. Once the most likely area has been identified, we determine a precise location by interpolating the locations of the most similar objects, in that area. Experimental results confirm the useful-

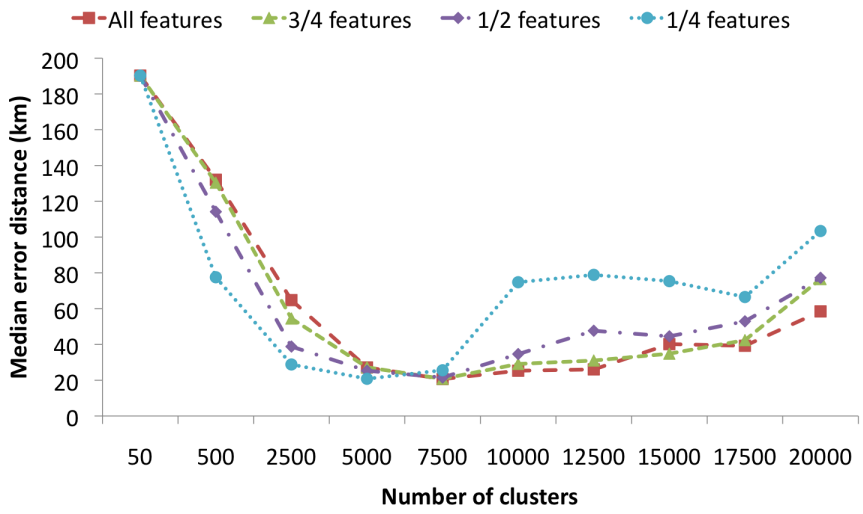


Figure 4.10: Impact of the amount of feature selection, in case of the Filtered *set-up*

ness of this hybrid methodology. We have also analysed the influence of previously geo-annotated resources by the same user, and found that, while the availability of such resources in the training set positively influences the performance, the difference in performance all but disappears if a sufficient number of tags is available for that resource.

We have experimented with several gazetteers (Geonames, DBpedia, and the US and world sets of USGS/NGA), but have not been able to improve our results. On the other hand, preliminary analyses that use an oracle for disambiguating toponyms show that using gazetteers together with our current method has the potential of reducing the median distance considerably. It thus remains unclear whether (or how) such resources could be useful for this task. In addition to gazetteers, other types of information could be taken into account, which we have not examined, including visual features and information about the profile and social network of the corresponding user.

Acknowledgments We thank Johannes Deleu for interesting discussions on the use of gazetteers.

References

[1] M. Ames and M. Naaman. *Why we tag: motivations for annotation in mobile and online media*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 971–980, 2007.

[2] S. Schockaert and M. De Cock. *Neighborhood restrictions in geographic IR*. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 167–174, 2007.

[3] P. Serdyukov, V. Murdock, and R. van Zwol. *Placing flickr photos on a map*. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 484–491, 2009.

[4] M. Goodchild. *Citizens as sensors: the world of volunteered geography*. GeoJournal, 69:211–221, 2007.

[5] L. Hollenstein and R. Purves. *Exploring place through user-generated content: Using Flickr to describe city cores*. Journal of Spatial Information Science, 1(1):21–48, 2010.

[6] O. Van Laere, S. Schockaert, and B. Dhoedt. *Finding locations of Flickr resources using language models and similarity search*. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval, pages 48:1–48:8, 2011.

[7] F. Wu and D. Weld. *Automatically refining the Wikipedia infobox ontology*. In Proceeding of the 17th International Conference on World Wide Web, pages 635–644, 2008.

[8] P. Schmitz. *Inducing ontology from Flickr tags*. In Proceedings of the Collaborative Web Tagging Workshop, pages 210–214, 2006.

[9] C. Zhai and J. Lafferty. *A study of smoothing methods for language models applied to Ad Hoc information retrieval*. In Proceedings of the 24th Annual International ACM SIGIR Conference, pages 334–342, 2001.

[10] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. *Mapping the world's photos*. In Proceedings of the 18th International Conference on World Wide Web, pages 761–770, 2009.

[11] J. H. Hays and A. A. Efros. *IM2GPS: Estimating geographic information from a single image*. In Proceedings of the 21st IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1–8, 2008.

[12] S. Ahern, M. Naaman, R. Nair, and J. H.-I. Yang. *World explorer: visualizing aggregate data from unstructured text in geo-referenced collections*. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 1–10, 2007.

[13] E. Moxley, J. Kleban, and B. Manjunath. *Spirittagger: a geo-aware tag suggestion tool mined from Flickr*. In Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, pages 24–30, 2008.

[14] L. Kennedy and M. Naaman. *Generating diverse and representative image search results for landmarks*. In Proceedings of the 17th International Conference on World Wide Web, pages 297–306, 2008.

[15] A. Popescu and I. Kanellos. *Creating visual summaries for geographic regions*. In IR+SN Workshop (at ECIR), 2009.

[16] F. Wilske. *Approximation of neighborhood boundaries using collaborative tagging systems*. In Proceedings of the GI-Days, pages 179–187, 2008.

[17] T. Rattenbury and M. Naaman. *Methods for extracting place semantics from Flickr tags*. ACM Transactions on the Web, 3(1):1–30, 2009.

5

Spatially-aware Term Selection for Geotagging

Current approaches to georeferencing rely on term selection techniques, such as TF-IDF, χ^2 or Information Gain (IG), which ignore the spatial nature of the domain. In this chapter, we implement the idea of spatial smoothing of term occurrences by using Kernel Density Estimation (KDE) to model each term as a two-dimensional probability distribution over the surface of the Earth. As an alternative, feature selection methods are presented that use Ripley's K function from geographical epidemiology. Experimental results are provided which demonstrate an improvement over the standard term selection methods.

⋆ ⋆ ⋆

Olivier Van Laere, Jonathan Quinn, Steven Schockaert and Bart Dhoedt.

Accepted for publication in IEEE Transactions on Knowledge and Data Engineering, IEEE, February 2013. Originally submitted, May 2012. Major revision submitted, December 2012. Minor revision submitted, February 2013.

Abstract The task of assigning geographic coordinates to textual resources plays an increasingly central role in geographic information retrieval. The ability to select those terms from a given collection that are most indicative of geographic

location is of key importance in successfully addressing this task. However, this process of selecting spatially relevant terms is at present not well understood, and the majority of current systems are based on standard term selection techniques, such as χ^2 or Information Gain, and thus fail to exploit the spatial nature of the domain. In this paper, we propose two classes of term selection techniques based on standard geostatistical methods. First, to implement the idea of spatial smoothing of term occurrences, we investigate the use of Kernel Density Estimation (KDE) to model each term as a two-dimensional probability distribution over the surface of the Earth. The second class of term selection methods we consider is based on Ripley's K statistic, which measures the deviation of a point set from spatial homogeneity. We provide experimental results which compare these classes of methods against existing baseline techniques on the tasks of assigning coordinates to Flickr photos and to Wikipedia articles, revealing marked improvements in cases where only a relatively small number of terms can be selected.

5.1 Introduction

The advent of mobile devices has gone hand-in-hand with an increased interest in geographic information retrieval. Indeed, as more information about the location of users becomes available, there is a growing need to identify the geographic scope of web resources: a promotion for UK railway tickets may be of little interest to a user in Australia, while photos of Portland Timbers games may mainly be of interest to residents of Portland. Gazetteers have traditionally been the main tool to assess the geographic scope of textual resources. However, gazetteers are limited to manually compiled lists of toponyms, which are necessarily restricted in scope. Many local landmarks or geographic features may not be contained in these lists, and vernacular places names and events are often not accounted for. Moreover, apart from place names and events, there may be a variety of other textual cues that can be used to estimate the geographic scope of a resources, such as slang words, regional product names, etc.

Large collections of georeferenced text provide an opportunity to complement traditional gazetteers, by identifying correlations between occurrences of terms and particular places. In this respect, the large number of photos on Flickr (currently about 175 million[1]) that have been provided with geographic coordinates is of particular interest. For example, by training language models from already georeferenced photos, [1] shows how coordinates can be estimated for previously unseen photos, based only on the associated textual tags. Such language models have even been shown to be capable of finding the coordinates of other resources, such as Wikipedia pages [2]. In [3], an approach is presented which discovers tags

[1] http://www.flickr.com/map/, accessed 12 February 2012.

on Flickr that refer to events and to places, whereas [4] uses Flickr to characterize the vague boundaries of neighborhoods in cities (e.g. places such as *downtown*) and [5] derives information about tourist attractions from georeferenced Flickr photos. Other authors use georeferenced Twitter messages to analyze correlations between terms and location; e.g. [6] studies the lexical variation across different geographic regions.

Most of the aforementioned approaches can essentially be seen as spatial forms of text categorization. As in standard text categorization, an important form of preprocessing consists of reducing the vocabulary by selecting only the most relevant terms. Indeed, it is well known that effective term selection can improve classifier effectiveness, and it can substantially decrease the computational cost, allowing existing methods to scale to collections of larger sizes [7, 8]. In our context, relevant terms are those that bear a certain spatial connotation, i.e. terms that occur disproportionally often in particular regions. So far, most authors have relied on standard term selection techniques, by discretizing the set of possible locations into a finite set of areas, and interpreting these areas as categories. However, there are a number of reasons why such an approach might be sub-optimal. First, the scale and boundaries of the chosen areas will inevitably be to some extent arbitrary. Furthermore, while in standard text categorization relevant terms are often tied to one particular category, this is to a much lesser extent the case when looking for spatially-relevant terms; e.g. toponyms are often highly ambiguous (e.g. geonames[2] contains 3771 records for *San Antonio*) and they may cover many of the considered areas (e.g. names of countries). Rather than identifying terms that are tied to one particular category, it might therefore be more appropriate to look for terms that are tied to a select number of geographic regions, something which traditional methods may not be entirely appropriate for. Other authors rely on more heuristic approaches, and directly look for clusters of occurrences of terms [9] or at burst-analysis techniques [3]. However, such techniques are strongly influenced by a priori assumptions on the distribution of location-relevant terms (e.g. regarding the scale at which to look for bursts), and it is not always clear which criterion they really optimize.

In this paper, we propose a number of approaches for the identification of location-relevant terms based on geostatistics. The first class of methods is based on the use of kernel density estimation (KDE [10]) in unison with established statistical and information theoretic measures. In this way, we aim to combine the best of both worlds, relying on KDE for endowing the methods with a form of "spatial awareness" while keeping the proposed scores easily interpretable due to their relationship with well-known measures. The second class of methods is based on Ripley's K statistic [11] which measures the extent to which a set of points diverges from a homogeneous distribution. We compare the proposed techniques

[2] http://www.geonames.org/search.html?q=san+antonio&country=, accessed 12 February 2012.

with standard term selection methods by analyzing the effect on the performance on the task of geotagging textual resources. We look at the performance when georeferencing Flickr photos, i.e. estimating their geographical coordinates based on the tags that have been assigned to them, and when georeferencing Wikipedia articles.

Our results demonstrate marked improvements over standard term selection techniques, especially in cases where relatively few terms are retained. Moreover, our results elucidate desireable characteristics of a term selection method in the area of geotagging, suggesting that we need to favour *common* terms whose occurrences are (i) correlated with spatial location and (ii) clustered around a small number of locations. While several of our methods identify spatial correlation, those that do not identify spatial clustering tend to perform relatively poorly. For example, occurrences of terms such as 'beach' or 'mountain' may be strongly correlated with spatial location and although these terms are likely to be useful for disambiguation, they are unlikely to be as useful for geotagging purposes as names of places. This observation also explains why the heuristic method from [9], called geographical spread, performs particularly well, as it is directly aimed at identifying common terms whose occurrences are clustered around a single location.

The paper is structured as follows. After reviewing related work in the next section, Section 5.3 discusses the proposed term selection methods, as well as a number of baselines. Subsequently, in Section 5.4 we detail the geotagging task that is used to quantitatively compare the different methods, and explain how our training and test data was obtained. Finally, we present the experimental results in Section 5.5.

5.2 Related work

Standard term selection techniques such as χ^2 and information gain have been widely studied. We refer to [7] and [8] for an overview and experimental comparison of such techniques.

Several authors have looked at techniques for identifying events from unstructured text, by looking for co-occurrences of dates and places [12], or of dates and named entities in general [13]. In addition to the fact that explicit references to dates occur in texts, many documents are also dated (e.g. news stories, blogs, emails), which means that document collections can often be seen as streams of text. Events are then extracted from such a stream by trying to identify bursts in one way or another [3, 14–16]. Such works essentially try to identify terms or phrases that are "temporally relevant", which is similar to our goal of finding terms that are "spatially relevant".

Large collections of georeferenced texts, however, have only relatively recently become available. Examples are georeferenced Flickr photos (where the terms

take the form of tags that have been assigned by users), georeferenced Twitter messages, and georeferenced Wikipedia pages. As a result, techniques which are similar to those for event detection can now be used to find location-specific terms. For example, the approach proposed in [3] to automatically detect places from such collections consists of applying the idea of burst detection to the spatial domain. In this paper, we focus on the extraction of terms with a clear geographical scope; essentially, extraction of a particular form of place semantics. In [27], an overview is presented of a number of different motivations for extracting spatially relevant terms. The motivation for selecting location-specific terms is often to find the most appropriate description for a given place at a given scale [3, 17]. The aim of [18], on the other hand, is to suggest location-relevant tags when users are uploading georeferenced photos, while [19] identifies locations mentioned in arbitrary text gathered from Flickr photos and demonstrates the use of neighbourhood and hierarchical smoothing techniques. The aims of the aforementioned works are different from our goal of limiting the set of terms to improve the effectiveness and efficiency of text classification in a spatial context. In the context of geotagging Flickr photos, sometimes χ^2 term selection is used [20] and sometimes no term selection at all [1]. Furthermore, the TagMaps TF-IDF method proposed in [27] reflects an idea for spatially aware term ranking. For geotagging Twitter messages, [21] uses a combination of stop word removal and some form of stemming, and moreover only retains those terms that occur at least 50 times. In [22], a generative probabilistic model is used to determine words with a geographic scope within a tweet, and a form of neighborhood smoothing is employed to refine the estimations. Geotagging general web pages is mostly gazetteer based (e.g. [23]), in which case only toponyms are considered, although [2] and [24] use language modeling approaches to assign coordinates to Wikipedia pages.

Kernel density estimation [10] is a popular technique for analyzing geographic point data (see e.g. [25]). In the context of geographic information retrieval, it has, among others, been used to model the vague boundaries of vernacular regions, using point data that is mined from the web [4, 26].

5.3 Identifying location-relevant terms

5.3.1 Baseline techniques

Standard approaches to term selection aim to find terms whose occurrence increases or decreases the probability that the document in which it occurs belongs to a given category. In other words, those terms are selected that are most discriminative w.r.t. a given set of classes and a given collection of classified documents. We briefly review four standard term selection techniques, as well as one recent method that has been specifically proposed in the context of geotagging Flickr

resources [9].

χ^2 term selection

One popular method to implement this idea uses the χ^2 statistic to assess to what extent there is a statistically significant difference between the actual number of occurrences of a term t in documents of class c and the number of occurrences that we would expect to see if the probability of seeing t did not depend on whether the corresponding document is in class c. Let us write O_{tc} for the number of times term t occurs in a document of class c. Similarly, let $O_{t\overline{c}}$ be the number of times t occurs in documents outside class c, $O_{\overline{t}c}$ the number of occurrences of terms other than t in documents of class c, and $O_{\overline{t}\overline{c}}$ the number of occurrences of terms other than t in documents outside class c. Moreover, we write E_{tc} for the expected frequency of term t in class c, and similar for $E_{\overline{t}c}$, $E_{t\overline{c}}$ and $E_{\overline{t}\overline{c}}$:

$$
\begin{aligned}
E_{tc} &= \frac{\sum_c O_{tc}}{\sum_{c'} \sum_{t'} O_{t'c'}} \cdot \sum_{t'} O_{t'c} \\
E_{\overline{t}c} &= \frac{\sum_c \sum_{t' \neq t} O_{t'c}}{\sum_{c'} \sum_{t'} O_{t'c'}} \cdot \sum_{t'} O_{t'c} \\
E_{t\overline{c}} &= \frac{\sum_c O_{tc}}{\sum_{c'} \sum_{t'} O_{t'c'}} \cdot \sum_{c' \neq c} \sum_{t'} O_{t'c'} \\
E_{\overline{t}\overline{c}} &= \frac{\sum_c \sum_{t' \neq t} O_{t'c}}{\sum_{c'} \sum_{t'} O_{t'c'}} \cdot \sum_{c' \neq c} \sum_{t'} O_{t'c'}
\end{aligned}
$$

The terms that are most discriminative w.r.t. class c are then chosen as those that maximize the χ^2 statistic:

$$
\chi^2(t,c) = \frac{(O_{tc} - E_{tc})^2}{E_{tc}} + \frac{(O_{\overline{t}c} - E_{\overline{t}c})^2}{E_{\overline{t}c}} + \frac{(O_{t\overline{c}} - E_{t\overline{c}})^2}{E_{t\overline{c}}} + \frac{(O_{\overline{t}\overline{c}} - E_{\overline{t}\overline{c}})^2}{E_{\overline{t}\overline{c}}}
$$

To select the best terms overall, we select those with the maximum χ^2 score over all classes. This approach was found to yield the best overall results for standard text categorization [7].

In the context of geotagging, we face the problem that there are no natural categories w.r.t. which we can evaluate the χ^2 statistic. This can be solved by first discretizing the locations on Earth into a finite number of areas, and let these play the role of categories. Throughout this paper, this will be accomplished by placing a 512×512 grid over the surface of the Earth (cf. [1, 17, 27]).

Log-likelihood

As an alternative to the χ^2 term selection, we consider Dunning's *log-likelihood* statistic [28]. For each term t and class c, the log-likelihood is given by:

$$\begin{aligned} G^2(c,t) = \quad & 2(O_{tc}\log O_{tc} + O_{t\bar{c}}\log O_{t\bar{c}} + O_{\bar{t}c}\log O_{\bar{t}c} + O_{\bar{t}\bar{c}}\log O_{\bar{t}\bar{c}} \\ & + N\log N \\ & - (O_{tc}+O_{t\bar{c}})log(O_{tc}+O_{t\bar{c}}) - (O_{tc}+O_{\bar{t}c})log(O_{tc}+O_{\bar{t}c}) \\ & - (O_{t\bar{c}}+O_{\bar{t}\bar{c}})log(O_{t\bar{c}}+O_{\bar{t}\bar{c}}) - (O_{\bar{t}c}+O_{\bar{t}\bar{c}})log(O_{\bar{t}c}+O_{\bar{t}\bar{c}})) \end{aligned}$$

where $O_{tc}, O_{t\bar{c}}, O_{\bar{t}c}$ and $O_{\bar{t}\bar{c}}$ are defined as before and N is the total number of photos in the training data. Similarly, the most relevant features for a given class can then be selected by choosing the features with the highest value for the G^2 statistic. To select the best overall terms, we need to aggregate the rankings obtained for every class c into a single ranking. This is accomplished by first selecting the best term from each of the rankings, then the terms at position 2, etc.

Information gain

Rather than looking for statistically significant anomalies, we may also use information theoretic measures to find the terms that are most informative w.r.t. a given classification. In particular, information gain measures the expected change in entropy about the class membership of a document d after learning that term t occurs in it [7]:

$$\begin{aligned} IG(t) =& H(C) - (p(t)\cdot H(C|t) + p(\bar{t})\cdot H(C|\bar{t})) \\ =& -\left(\sum_c p(c)\cdot\log(p(c))\right) \\ & + p(t)\cdot\left(\sum_c p(c|t)\cdot\log p(c|t)\right) \\ & + p(\bar{t})\cdot\left(\sum_c p(c|\bar{t})\cdot\log p(c|\bar{t})\right) \end{aligned} \tag{5.1}$$

where maximum likelihood estimates are used for all probabilities, e.g. $p(c|t) = \frac{O_{tc}}{\sum_{c'} O_{tc'}}$. In contrast to χ^2, information gain immediately provides us with an overall ranking of the terms. Again, the classes correspond to the cells of a 512×512 grid.

Most used

A particularly simple term selection technique that is sometimes used consists of selecting the terms that occur in the largest number of documents. Despite the simplicity of the method, it often performs remarkably well in practice [7].

Geographical spread

In [9], a term selection technique was proposed which aims at finding location-relevant terms. It explicitly looks at the extent to which occurrences of a tag are clustered around a small number of locations. Specifically, for each term, [9] proposes to calculate the geographical spread as follows. First, a grid is placed on the surface of the Earth, as before, and for each grid cell c, the number of occurrences O_{tc} of a given term t is determined. Then a graph G is constructed, in which the nodes correspond to the cells $c_{i,j}$ for which $O_{tc_{i,j}} > 0$. There is an edge between the nodes corresponding to cells $c_{i,j}$ and $c_{i',j'}$ if $c_{i',j'} \in \{c_{i-1,j}, c_{i+1,j}, c_{i,j-1}, c_{i,j+1}\}$, i.e. if $c_{i,j}$ and $c_{i',j'}$ are adjacent cells (where adjacency along the diagonals is not considered). The geographic spread (or geospread) $S(t)$ of a term t is then defined as

$$S(t) = \frac{|\mathcal{C}|}{\max_{C \in \mathcal{C}} \left(\sum_{c \in C} O_{tc} \right)}$$

where $\mathcal{C}$ is the set of all connected components of G, and each connected component C is identified with the corresponding set of grid cells c. In [9] the grid is chosen such that each cell is 1 degree longitude by 1 degree latitude. For conformity with the other methods, however, we will use a 512×512 grid in this paper.

The core idea of this measure is that the connected components correspond to a cluster of occurrences of the term t. Some terms may refer to a very precise area (e.g. landmarks such as *eiffeltower*) while others may refer to a broader region (e.g. countries such as *france*). The geospread measure treats such terms more or less as equal, by only looking at the number of clusters (each of which may correspond to a precise or a broad region) and the absolute size of the largest cluster.

Due to the procedural nature of its definition, it is not clear what exactly is optimized by the geospread measure, which stands in stark contrast to measures such as χ^2 or information gain. Nonetheless, experimental results reveal that, at least for the task of geotagging, the geospread measure outperforms both of the aforementioned methods.

5.3.2 KDE based methods

Our aim in this section is to introduce a number of measures to identify Flickr tags that relate to location. Such tags can refer to toponyms (e.g. *paris*, *france*, *mediterranean*), but also to geographic features (e.g. *beach*, *forest*, *lake*), names of landmarks (e.g. *empirestatebuilding*, *eiffeltower*), events (e.g. *911*, *ironman*), slang words, etc. For each of these types of tags, the distribution of tag occurrences should deviate substantially from that of general tags such as *birthdayparty* or *iphone*. Nonetheless, in many cases, there may be several grid cells that contain

a large number of tag occurrences. Classical term selection techniques are poorly equipped to differentiate between situations where these cells define a small number of regions and situations where such cells occur at many different places. The geospread measure, on the other hand, does explicitly look for clusters of grid cells, but is difficult to interpret.

The alternatives that we propose in this paper associate with each tag t a probability distribution $p(\mathcal{A}|t)$ of locations, where locations are again taken to be the cells of a 512×512 grid $\mathcal{A}$. This probability distribution is obtained using kernel density estimation (KDE [10]).

Kernel Density Estimation is a statistical analysis tool, used to generate a non-parametric probability density function estimation. KDE is somewhat similar in principle to histogram generation, but suffers less from the effects of quantization. A KDE is a linear combination of local kernel density estimates, where the smoothness of a kernel is controlled by a bandwidth parameter θ in degrees latitude/longitude. This process results in a high statistical efficiency. A standard KDE is computed using:

$$q(\mathcal{A}|t) = \frac{1}{|A|} \sum_{a \in \mathcal{A}} K \frac{(t - t_a)}{\theta}, \qquad t \in \mathbb{R}.$$

Where K is the kernel. Choosing an optimal value of θ is highly important, as the solution for f depends significantly upon it. The bandwidth can be selected using methods such as the Mean Integrated Squared Error [29]. However, the *optimal* value, as chosen by such methods, is often inappropriate for the task, in a similar way that the optimal selection for the number of clusters in a clustering algorithm may not be suitable. For example, a method may benefit more from a higher level of smoothing than is inherent in the data. In this paper, we use the robust diffusion KDE method described in [29], but define our own value for θ. $q(\mathcal{A}|t)$ must be normalised for use as a probability distribution $p(\mathcal{A}|t)$.

In the same way, from the set of all locations of the photos in the training set, a background distribution $p(\mathcal{A})$ is estimated using KDE. For clarity, we will write $p_{KDE}(\mathcal{A}|t)$ and $p_{KDE}(\mathcal{A})$ for the distributions that result from the KDE process. The background distribution $p_{KDE}(\mathcal{A})$, visualized in Figure 5.1, shows large peaks in North America, Europe and Japan. Figure 5.2 shows the KDE for the tag *oregon*. Whilst the main peaks exist in the state of Oregon, ambiguities are still present, such as for example, the City of Oregon, and locations in the states of Ohio, Illinois, and Missouri. Figure 5.3 shows the KDE for the tag *beach*, specifically for the European region. Peaks consistently occur at coastal regions. In each example, the log of the KDE is shown to aid in the visualisation. Note how our use of KDE in this way corresponds to a form of spatial smoothing. The larger the value of the bandwidth parameter θ, the more the occurrence of a tag t in a given cell influences the distribution for its neighbouring cells.

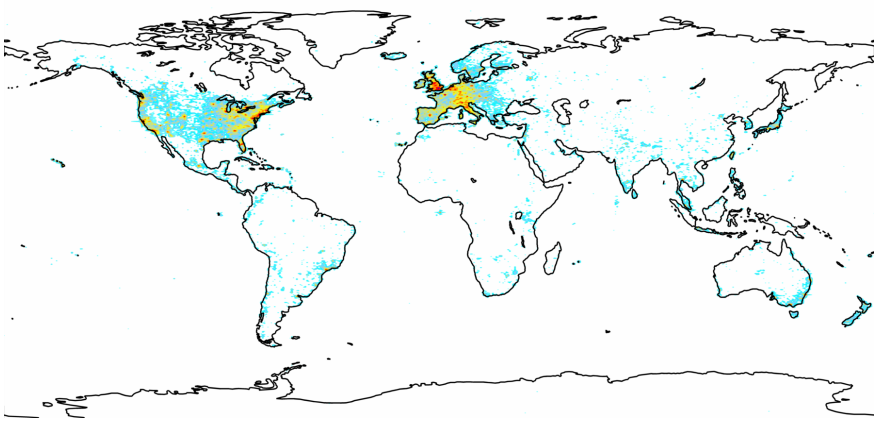


Figure 5.1: Log of the background distribution KDE. Bandwidth was chosen as 10^{-7} *degrees latitude/longitude.*

Geographical spread as entropy

The first idea we explore to identify such tags follows the same intuitions as the geospread measure: the more the occurrences of a tag are clustered around a few locations and the more often it occurs, the more likely that the tag bears a location-specific meaning. In contrast to the geospread measure, however, we propose an easily interpretable score, and we introduce a parameter that allows us to control the trade-off between favouring location-specific tags and avoiding rare tags.

Intuitively, the idea that tags that are clustered around a few locations should be favoured is closely aligned with the information theoretic notion of entropy. The more the probability distribution $p(\mathcal{A}|t)$ is centered around a few peaks, the lower the entropy of that distribution will be, and the more desirable tag t is. However, when estimating $p(\mathcal{A}|t)$ based on KDE, the total number of occurrences of tag t is not taken into account. For instance, a tag which occurs only once will trivially correspond to a distribution with a minimal entropy of 0. To cope with this, we propose to further smooth $p(\mathcal{A}|t)$ with the amount of smoothing depending on the total number of occurrences of tag t. Using Bayesian smoothing with Dirichlet priors, we obtain

$$p_{Dir}(a|t) = \frac{p_{KDE}(a|t) \cdot N_t + \mu \cdot p_{KDE}(a)}{N_t + \mu}$$

where N_t is the total number of occurrences of tag t and $\mu \in [0, +\infty[$ is a parameter which controls how many samples we should see to abandon the idea that occurrences of t follow the general distribution. Taking a lower value of μ will result in more rarely occurring tags being selected. In particular if $\mu = 0$ we recover $p_{KDE}(\mathcal{A}|t)$ while for very large values of μ, $p_{Dir}(\mathcal{A}|t)$ will tend to $p_{KDE}(\mathcal{A})$.

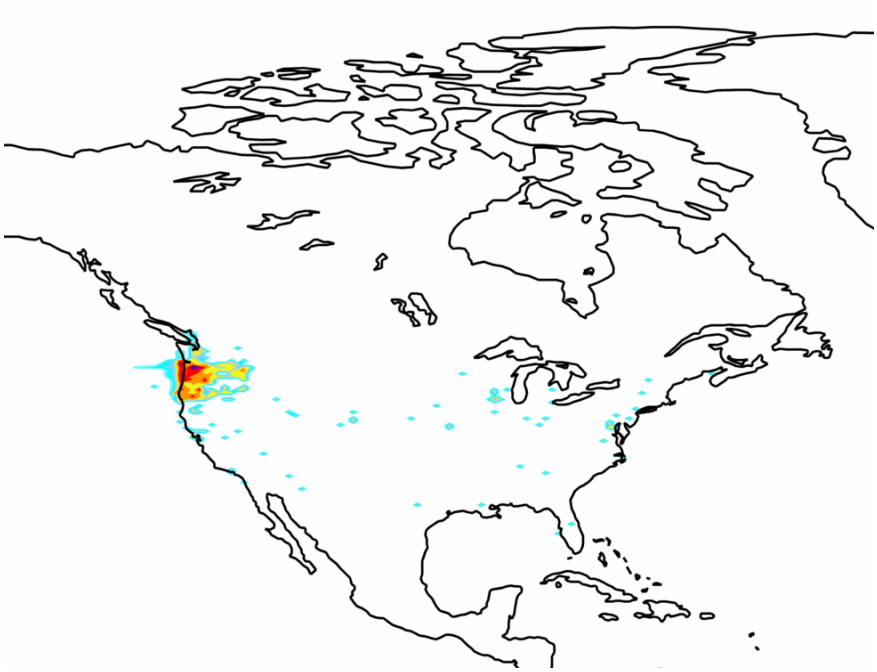


Figure 5.2: Log of the KDE for the tag oregon*, shown for the North American region. Bandwidth was chosen as* 10^{-7} *degrees latitude/longitude.*

Alternatively, using a uniform prior, we have

$$p_{uni}(a|t) = \frac{p_{KDE}(a|t) \cdot N_t + \frac{\mu}{|\mathcal{A}|}}{N_t + \mu}$$

After this smoothing step, entropy can be used to rank the tags:

$$s^{ent}_{Dir}(t) = H_{Dir}(\mathcal{A}|t) = -\sum_{a \in \mathcal{A}} p_{Dir}(a|t) \cdot \log(p_{Dir}(a|t)) \tag{5.2}$$

$$s^{ent}_{uni}(t) = H_{uni}(\mathcal{A}|t) = -\sum_{a \in \mathcal{A}} p_{uni}(a|t) \cdot \log(p_{uni}(a|t)) \tag{5.3}$$

Note the difference between these scores and the information gain method from (5.1). Rather than quantifying the uncertainty that is removed by the occurrence or non-occurrence of a tag t, here we are interested in the entropy itself as a measure of how much the probability density $p(\mathcal{A}|t)$ is spread out over the surface of the Earth. Furthermore, note how the use of KDE implies that having occurrences of the tag in neighbouring cells is less penalized than having occurrences in disjoint cells, with the bandwidth parameter θ controlling how far apart occurrences need to be considered disjoint.

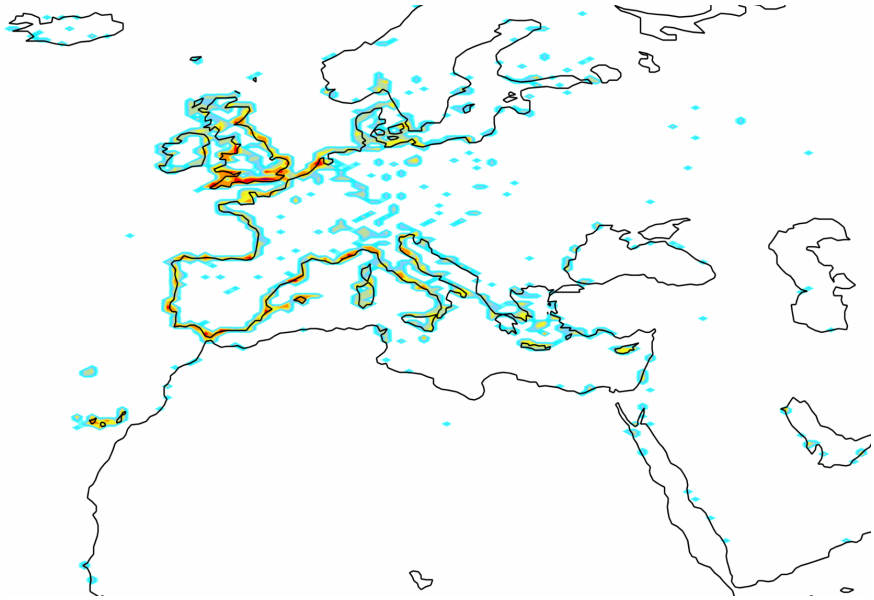


Figure 5.3: Log of the KDE for the tag beach*, shown for the European region. Bandwidth was chosen as* 10^{-7} *degrees latitude/longitude.*

Divergence from the background distribution

The idea underlying (5.2)–(5.3) is that useful terms are those that occur mainly at a few selected locations. Here, we take a slightly different view, whereby terms are assumed to be location-relevant to the extent that the distribution of their occurrences diverges from the background distribution $p_{KDE}(\mathcal{A})$. This can be quantified using the Kullback-Leibler divergence between $p_{KDE}(\mathcal{A})$ and $p_{Dir}(\mathcal{A}|t)$, i.e.

$$s^{kl}(t) = D_{KL}(p_{Dir}(\mathcal{A}|t) \,\|\, p_{KDE}(\mathcal{A})) = \sum_{a \in \mathcal{A}} p_{Dir}(a|t) \cdot \log \left(\frac{p_{Dir}(a|t)}{p_{KDE}(a)} \right)$$

Note that the role of KDE is slightly different here. Essentially, the spatial smoothing that is obtained from using KDE ensures that the Kullback-Leibler divergence will be low as long as most occurrences of t are *near* the cells with the highest probability in the background distribution. In other words, the idea is that any artefacts from the training data are smoothed out.

A related idea is to use a goodness-of-fit test to assess with which degree of confidence we can reject the null hypothesis that the occurrences of tag t have been sampled from the background distribution $p_{KDE}(\mathcal{A})$. Using the χ^2 test this leads

to the following score:

$$s^{\chi^2}(t) = \sum_{a \in \mathcal{A}} \frac{(O_{ta} - p_{KDE}(a) \cdot N_t)^2}{p_{KDE}(a) \cdot N_t} \tag{5.4}$$

with O_{ta} the number of occurrences of tag t in grid cell a and N_t the total number of occurrences of t, as before.

5.3.3 Ripley's K based methods

Ripley's K function [11] is a statistic which is used to analyse whether a given point set is likely to have been generated from a homogeneous Poisson distribution. Among others, it is used in epidemiology to analyse the distribution of disease cases [30], and in ecology to analyse the spatial distribution of plants [31]. Given a set of N points Q spread over an area of size A, it can be estimated as

$$K(\lambda) = A \cdot \frac{|\{(p,q) \,|\, p,q \in Q, p \neq q, d(p,q) \leq \lambda\}|}{N^2}$$

In other words, $K(\lambda)$ is proportional to the pairs of points from Q that are within distance λ from each other. Note that A is constant for all terms, and as we are only interested in ranking terms, we can safely ignore it, i.e. we evaluate for each term t with N_t occurrences the right-hand side of the following expression:

$$K(\lambda) \propto \frac{|\{(p,q) \,|\, p,q \in Q_t, p \neq q, d(p,q) \leq \lambda\}|}{N_t^2}$$

where the set Q_t contains the locations of the photos to which term t has been assigned. For reasons of efficiency, we use the Manhatten distance for d, as this allows us to efficiently retrieve all points within distance λ of a given point by indexing the points using a k-d tree.

A simple idea would be to rank terms according to their $K(\lambda)$ value for a suitable choice of λ. However, the values for $K(\lambda)$ are not comparable between terms with a different number of occurrences: the larger the number of occurrences N_t of a given term t, the easier it becomes to obtain a larger value of $K(\lambda)$ by chance. Therefore, we will rank terms based on the probability that such a value could have been obtained by chance. As analytical solutions for the critical values of $K(\lambda)$ are not available, we have used a Monte Carlo simulation to this end. To find the critical values for a term with N occurrences, we randomly selected 10000 sets of N points, by choosing locations of photos in our training data as possible points. We repeated this process for N going from 3 to 1000, and for λ equal to 1km, 10km and 100km. Figure 5.4 shows the critical values for $K(10)$ that allow us to conclude that the spatial clustering of a term is not due to chance with 95% confidence.

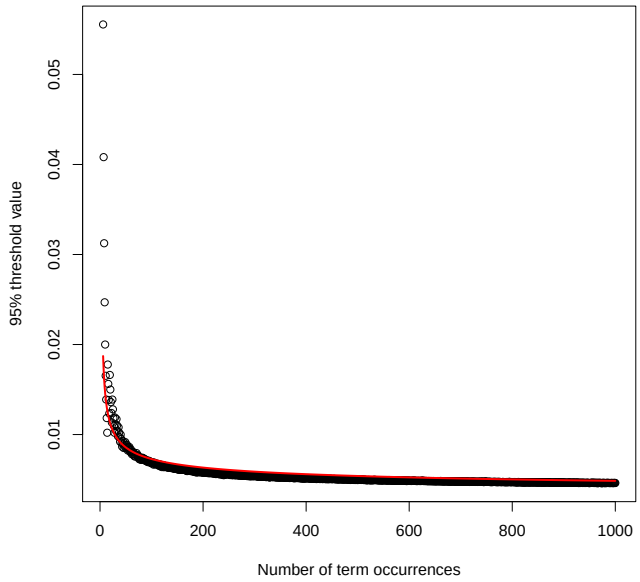


Figure 5.4: Critical $K(10)$ values for the 95% confidence level, in function of the number of term occurrences N.

Unfortunately, in the training data (described in Section 5.4), out of 293 673 terms with at most 1000 occurrences, 287 901 terms have a value for $K(10)$ that is above the critical value. However, as Figure 5.4 shows, the critical values for $K(10)$ are approximately proportional to $\frac{1}{\log N}$. Similar results were found for other values of λ and other confidence thresholds. Therefore, we propose to rank terms according to the following score:

$$s^K_{log}(t) = \log N_t \cdot \frac{|\{(p,q) \,|\, p,q \in Q_t, p \neq q, d(p,q) \leq \lambda\}|}{N_t^2}$$

In our experiments, we also consider the following variant:

$$\begin{aligned} s^K_{lin}(t) &= N_t \cdot \frac{|\{(p,q) \,|\, p,q \in Q_t, p \neq q, d(p,q) \leq \lambda\}|}{N_t^2} \\ &= \frac{|\{(p,q) \,|\, p,q \in Q_t, p \neq q, d(p,q) \leq \lambda\}|}{N_t} \end{aligned}$$

The latter variant will favour terms that occur more often, based on the view that when we can only select a limited number of terms, we should choose those that are both spatially clustered and common. A third variant is based on the observation that the geographical spread measure from [9] mainly favours terms whose occurrences are centered around a small number of points. For example, while occurrences of a term such as 'beach' will be strongly location-dependent, its geographical spread score will be rather low. To favour terms whose occurrences are centered around only a few points, we use the following score

$$s^K_{lin\text{-}\omega}(t) = N_t \cdot \frac{\sum_{p \in Q_t} \left(|\{q \,|\, q \in Q_t, p \neq q, d(p,q) \leq \lambda\}|\right)^\omega}{N_t^2}$$

Note that for $\omega = 1$, we have $s^K_{lin\text{-}\omega}(t) = s^K_{lin}(t)$. For $\omega > 1$, terms are favoured that occur centered around a small number of locations. For example, the term 'landscape' is ranked at position 2162 for $\omega = 1$ and at position 259113 for $\omega = 5$. The reason is that while occurrences of this term are strongly correlated to areas on the globe that are suitable for landscape photography, its occurrences are not centered around a few particular places. Conversely, the term 'westmidlandfireservice' only has 3 occurrences in the training set, all of which are near the same location. Due to the small number of occurrences, for $\omega = 1$ it is ranked towards the end of the list, at position 214262. For $\omega = 5$, however, the clustering of the term occurrences is considered more important and the term moves up to position 128650.

5.4 Assigning coordinates to textual resources

Since 2010, the MediaEval workshop has featured a benchmarking initiative called the Placing Task, whose goal is to allow for the comparison of current approaches in automated geotagging of Flickr resources. While the use of gazetteers and visual and audio features is tempting for this task, the best performing systems in the past two years have been purely based on statistically analysing tags [9, 32], thus confirming and strengthening the support for language models, initially proposed in [1] for this task. To evaluate and compare the proposed term selection techniques, we follow a similar language modeling based approach to geotagging Flickr photos.

First, we crawled a representative set of geotagged Flickr photos using the public API in April 2011. This resulted in a collection of around 105M photos, equivalent to about 70% of the total number of geotagged photos available on Flickr at that time. From this set, we removed all photos that did not contain any tags or were tagged with invalid coordinates and we subsequently filtered the data set for bulk uploads by the users (following [1]). In this way, a reduced set of 43.7M photos was obtained. Among these photos, 25M photos were randomly chosen as training data and 100k photos as test data, ensuring that the set of photo owners was disjoint for training and test data. Note that the original 2011 Placing Task test set was not used for testing purposes as it only contains 5347 test items, most of which are moreover owned by users appearing in the training set.

From the initial training data of 25M photos, we selected those tags that were used by at least 3 users. To dampen the potential impact of the tagging behavior of any single user, for term selection we furthermore limited the number of considered occurrences of terms to one per user. In addition, both for term selection and for georeferencing, we ignored photos with a reported accuracy below the maximal level of 16 (which corresponds to a street-level accuracy). Thus we arrived at a final set of 9 472 388 photos, making up our actual training data.

Both for term selection and for georeferencing, a form of discretization needs to be applied. For term selection, it is important to have areas that are sufficiently fine grained and of a comparable size, so a rectangular 512×512 grid was used to this end. Note that by doing this, we have not used a particular map projection, simplifying the implementation. Importantly, it also ensures that our results are easily comparable to other methods in the literature that use a similar grid. For georeferencing, on the other hand, we need to ensure that sufficient training data is available for each of the areas we end up with. This means that the total number of areas should not be too high, and that larger area sizes may be needed for parts of the world where fewer training photos are available. Following [20], we found k-medoids clustering with $k = 5000$ clusters to yield good results.

To implement the actual georeferencing step, a multinomial Naive Bayes clas-

sifier can be trained over the set of areas $\mathcal{A}$. When representing an unseen test photo x as its set of tags, the probability $P(a|x)$ is then estimated as being proportional to

$$P(a|x) \propto P(a) \cdot \prod_{t \in x} P(t|a) \tag{5.5}$$

The prior probability $P(a)$ is estimated using maximum likelihood estimation, while for $P(t|a)$, we applied Bayesian smoothing with Dirichlet priors ($\mu = 1000$), as this was found to give the best results in [1] and [20]. The area a maximizing the right-hand side of (5.5) is then converted to an actual location estimate by choosing the coordinates of the medoid of the corresponding cluster, i.e. the photo x minimizing $\sum_{x' \in a} d(x, x')$ with d the geodesic distance, over all photos in area a.

To assess the generality and robustness of the proposed methods, after using the Flickr test set to find suitable parameter values, we also test our methods on a test set consisting of Wikipedia articles. This test set consists of 21 839 geotagged Wikipedia documents that can be considered a spot (i.e. can be located to a certain, narrow, geographical scope). To construct this test set, we downloaded the DBPedia 3.7 "Geographic Coordinates" English (nt) Wikipedia dump[3], containing the geographical coordinates and Wikipedia ID's (e.g. "Abbotsford_House") of 442 775 entities. From these, we retained the 47 493 documents whose coordinates are located within the bounding box of the United Kingdom[4]. Wikipedia contains numerous documents that are hard to pinpoint to a precise location, discussing for example architectural styles, schools of thought, people or concepts. As we consider techniques for estimating precise coordinates, it is useful to restrict the evaluation to articles that have a limited spatial extent, such as landmarks, buildings, schools, or railway stations. To this end, we have further filtered the dataset, keeping only the documents whose coordinates either refer to a location of type "railwaystation, landmark or edu", or have a reported scale of 1:10000 or finer.

To apply the georeferencing method described above to the Wikipedia test data, we map the test documents to sets of Flickr tags. This can easily be achieved by converting the Wikipedia test documents to lowercase, and scanning for terms or concatenations of up to 3 consecutive terms that correspond to Flickr tags. The Wikipedia test set and the set of 301968 features used in our evaluations are available online[5].

[3] http://downloads.dbpedia.org/3.7/en/geo_coordinates_en.nt.bz2

[4] Note that, as with any test collection, our choice of restricting test documents to those that are located in the UK introduces a particular bias. It remains to be verified to what extent the conclusions we present in this paper generalize to other test collections, and in particular, to areas of the world for which available training data is more sparse.

[5] https://github.com/ovlaere/spatial_feature_selection

5.5 Experimental results

In this section, we will analyze the effect of the term selection methods on the performance of the geotagging system that was discussed in Section 5.4. To this end, we will use the following two metrics:

1 **Median distance** is the median of the distance between the estimated location and the true location, over all 100k photos in the test set.

2 **Accuracy at n km** is the percentage of photos from the test set for which the estimated location is at most n kilometre from the true location.

First, in Section 5.5.1, we analyse the behaviour of the proposed KDE based methods in detail, looking in particular at the relative performance of the proposed scores and the influence of the underlying parameters. Then, in Section 5.5.2, we analyse the proposed scores based on Ripley's K statistic in a similar fashion. Finally, Section 5.5.3 compares the KDE and K based methods against a number of baseline techniques, and provides a discussion on the practical impact of our result.

5.5.1 KDE based methods

There are two parameters that play a key role in the performance of the methods. First, the bandwidth parameter θ (in degrees latitude/longitude) of the KDE algorithm determines the degree of spatial smoothing that is applied. Choosing a higher value for θ means that the influence of noise will be reduced, but at the same time, useful local effects may be cancelled. Second, the parameter μ controls the extent to which rare tags are penalized: the larger the value of μ the more frequent tags need to be considered desirable. Note, however, that s^{χ^2} does not depend on μ. Both in the case of θ and μ, each of the methods are robust against small changes; only the order-of-magnitude of these parameters is important.

Figure 5.5 compares the different KDE based methods for a basic configuration with $\theta = 10^{-7}$ and $\mu = 1000$. Due to the difference in smoothing method, s^{ent}_{Dir} will favour terms in areas which are popular overall, whereas s^{ent}_{uni} will select terms that are region-specific, regardless of whether they are in a popular area. Indeed, because of the use of the Dirichlet prior a spike in an unpopular area (i.e. an area to which the background distribution $p(\mathcal{A})$ assigns a low probability) will cancel out spikes in the background distribution, thus reducing entropy, unless the tag occurs sufficiently often. The performance of s^{kl} and in particular s^{χ^2} is clearly worse. As already hinted at in the introduction, identifying spatial correlation by itself is not enough, as terms such as 'beach' or 'forest' may diverge substantially from the background distribution, while perhaps not being as interesting as toponyms and other types of spatially relevant terms, in the context of geotagging. The entropy

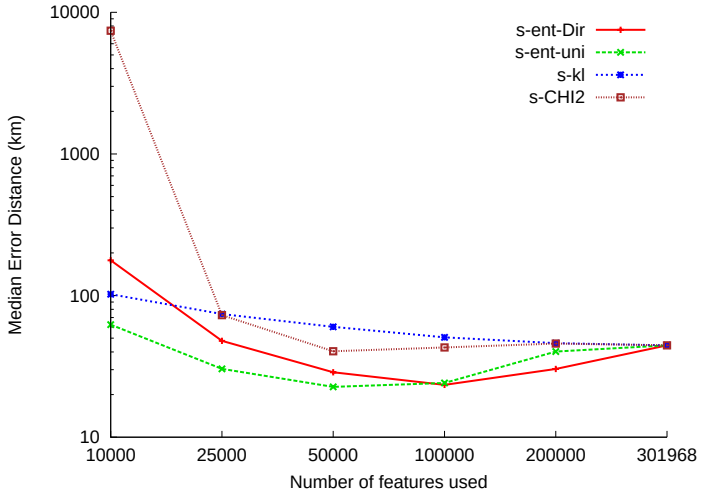


Figure 5.5: Comparison of the different KDE based term selection methods, using $\mu = 1000$ and $\theta = 10^{-7}$ where applicable. In each case, the median is reported of the distance between the estimated location and the true location of the 100k photos of the test set.

based methods, on the other hand, identify terms whose occurrences are highly clustered.

To better understand the influence of the bandwidth parameter, Figures 5.6–5.8 show how the performance of s^{ent}_{Dir}, s^{kl} and s^{χ^2} changes with varying values of the bandwidth parameter. From Figure 5.6 we can see that s^{ent}_{Dir} performs comparably for $\theta = 10^{-5}$, but gets substantially worse for larger values of θ. The figure also illustrates the actual contribution of using KDE: the results for *no KDE* were obtained by estimating $p(\mathcal{A})$ and $p(\mathcal{A}|t)$ directly from the data using maximum likelihood, instead of using KDE. While the result is slightly better for 10k tags, it is much worse overall. On the other hand, in the case of s^{kl}, the value of the bandwidth parameter, and the use of KDE in general, seems to have a minor effect only. In the case of Figure 5.8, finally, we find that increasing the bandwidth parameter to 10^{-1} improves the results considerably. While the entropy based measures need a small bandwidth parameter to ensure that any spikes in the data are not lost, s^{χ^2} needs larger values of θ to eliminate noise.

We now turn to the parameter μ, which is analyzed in Figure 5.9 for s^{ent}_{Dir} and s^{ent}_{uni} and in Figure 5.10 for s^{kl}; recall that s^{χ^2} is independent of μ. In general, smaller values of μ lead to less popular tags being selected sooner. As can be seen, when taking $\mu = 100$, this initially leads to worse results. This is to be expected, since, all things being equal, popular tags are more useful for classification than rare tags. However, when at least 100k terms are selected, the results for $\mu = 100$

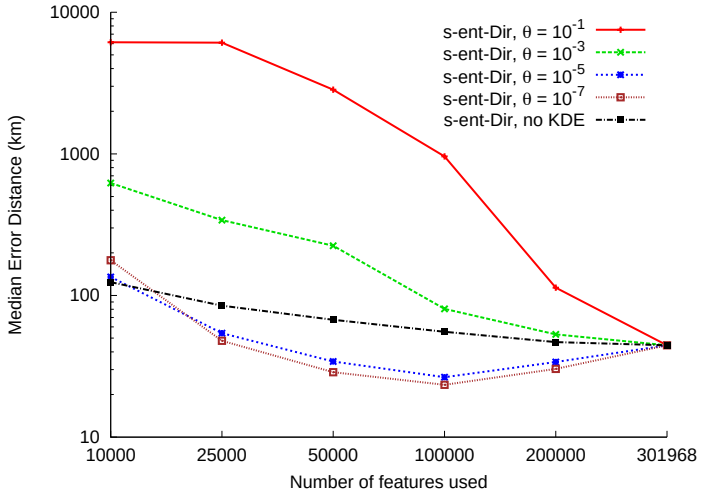


*Figure 5.6: Influence of the bandwidth parameter on method s^{ent}_{Dir}. For comparison, we also report the results of a variant of s^{ent}_{Dir} in which the distributions $p_{KDE}(\mathcal{A})$ and $p_{KDE}(\mathcal{A}|t)$ are replaced by maximum likelihood estimations (*no KDE*).*

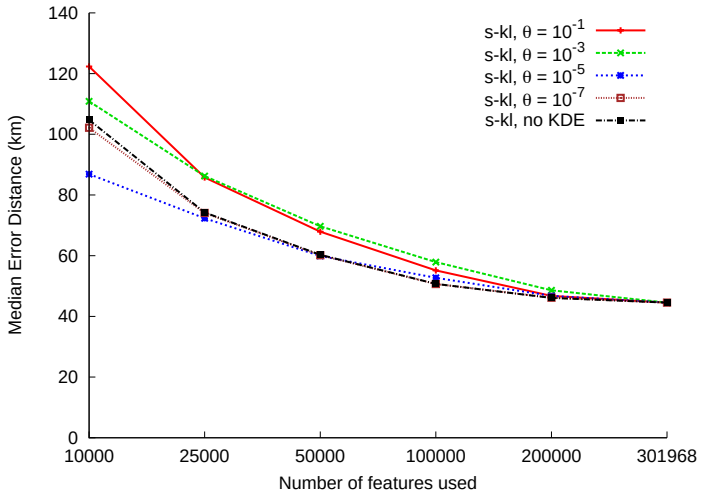


Figure 5.7: Influence of the bandwidth parameter on method s^{kl}. Again, for comparison results are shown of a variant in which the KDE estimations are replaced by maximum likelihood estimations.

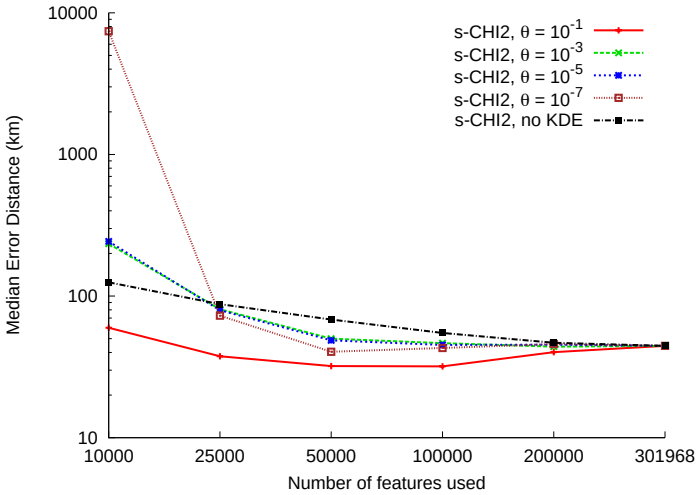


Figure 5.8: Influence of the bandwidth parameter on method s^{χ^2}. Again, for comparison results are shown of a variant in which the KDE estimations are replaced by maximum likelihood estimations.

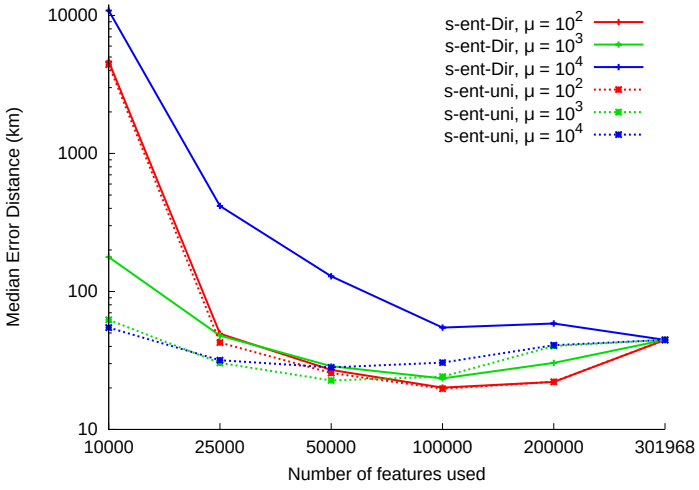


Figure 5.9: Influence of the smoothing parameter μ on the methods s^{ent}_{Dir} and s^{ent}_{uni}.

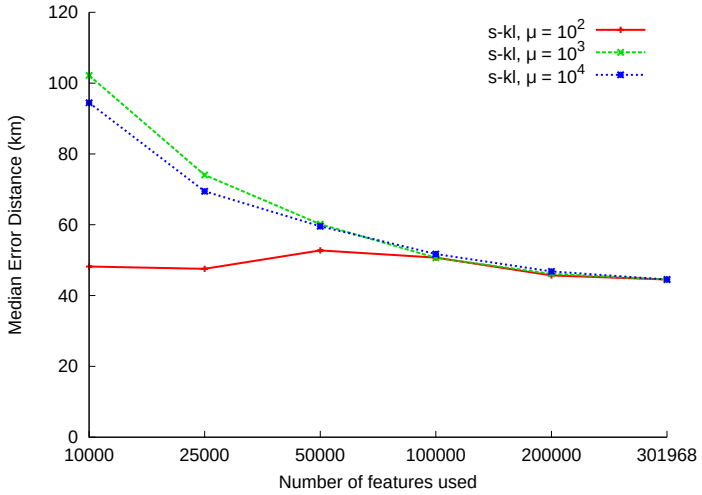


Figure 5.10: Influence of the smoothing parameter μ on the method s^{kl}.

outperform those of $\mu = 1000$. Hence the parameter μ appears to be effective in controlling the trade-off between using a safe strategy (focusing on popular tags) which is initially effective but may miss out some interesting rarer tags if a sufficiently high number of tags is selected, and using a more adventurous approach which may initially choose some sub-optimal tags, but does eventually find all or most of the relevant ones. In the case of s^{ent}_{uni}, the results for $\mu = 10000$ are quite similar to those for $\mu = 1000$, showing the robustness of this score against particular choices of μ: only for sufficiently small values of μ will the actual choice have a real impact. In the case of s^{ent}_{Dir}, however, choosing μ too large will put too much emphasis on regions which have spikes in the background distribution, missing out too much of the relevant tags in other parts of the world. The situation for s^{kl} in Figure 5.10 is slightly different, where $\mu = 100$ leads to a better performance overall. Together with the findings from Figure 5.7, this suggests that s^{kl} quickly loses its ability to discriminate location-relevant tags when too much smoothing is applied. Again the results for $\mu = 10000$ and $\mu = 1000$ are comparable.

5.5.2 Ripley's K based methods

When using the scores based on Ripley's K statistic, the main parameter is the distance λ within which two points are considered sufficiently close. Figure 5.11 clearly shows that choosing $\lambda = 100km$ leads to better results than 1km or 10km. As for the KDE based methods, we are not aiming to find the optimal value for the parameters, but only an indication of the order-of-magnitude of reasonable values.

Table 5.1 compares the performance of the methods $s^K_{lin\text{-}\omega}$. While choosing

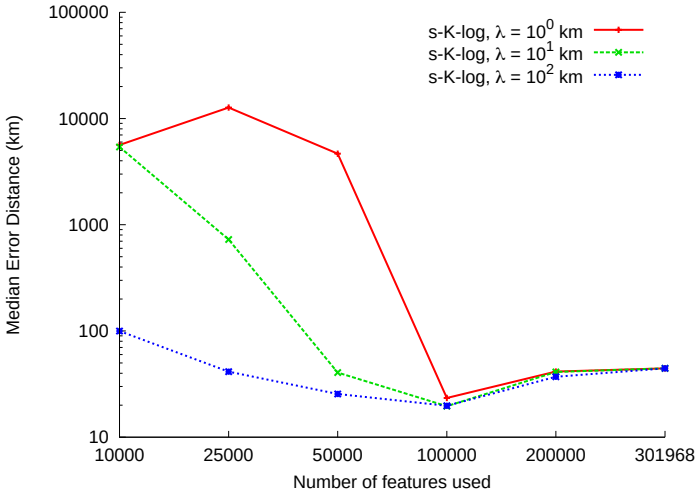


Figure 5.11: Influence of the scale parameter λ on the method s^K_{log}.

$\lambda = 100km$ remains a good choice, for these methods, $\lambda = 10km$ can lead to slightly better performance. We find that $\omega = 2$ leads to the best result initially, while choosing $\omega = 5$ leads to the best result overall. This observation is consistent with our results for the entropy-based KDE methods. Indeed, like these entropy-based methods, $s^K_{lin\text{-}5}$ will favour those terms whose occurrences are clustered around one or a few locations, favouring place names over spatially correlated terms which occur across the globe (e.g. the names of geographic features). As this strategy puts less emphasis on the statistical significance with which we can conclude that a given tag's occurrences have not been sampled from a homogeneous Poisson distribution, it can be seen as more adventerous, which may explain the worse initial performance. Similar to what we found for the KDE based methods, configurations that are optimal for aggressive feature selection (viz. only selecting 10k terms) tend to be not globally optimal.

5.5.3 Comparison with existing methods

The aims of this section are (i) to compare the KDE and K based methods to the existing methods that have been described in Section 5.3.1, (ii) to evaluate our methods on a different test set to assess their generality and robustness, and (iii) to further clarify why certain methods perform better than others.

Figure 5.12 compares the baseline methods against each other. We can observe that both χ^2 and information gain are clearly outperformed by geospread, which supports the view that classical term selection methods are less suitable for this task. While the most-used technique is outperformed by geospread when at

Table 5.1: Comparison of the methods $s^K_{lin\text{-}\omega}$

Function	Threshold	10000	25000	50000	100000	200000	301968
$s^K_{lin\text{-}1}$	1km	180.19	53.1	32.25	29.93	41.49	44.51
	10km	63.41	36.62	31.47	31.24	41.16	44.51
	100km	60.21	43.02	37.86	36.93	40.33	44.51
$s^K_{lin\text{-}2}$	1km	5313.26	380.45	50.49	21.22	41.49	44.51
	10km	145.73	38.67	24.92	22.55	41.16	44.51
	100km	**48.39**	**29.29**	24.39	25.61	38.65	44.51
$s^K_{lin\text{-}3}$	1km	5795.09	4895.66	130.36	21.96	41.49	44.51
	10km	305.29	65.31	29.79	19.66	41.16	44.51
	100km	54.69	32.09	**23.91**	20.09	36.75	44.51
$s^K_{lin\text{-}4}$	1km	13020.06	5664.03	418.18	23.01	41.49	44.51
	10km	4279.31	98.53	33.73	**19.53**	41.16	44.51
	100km	61.80	34.65	24.74	19.62	35.19	44.51
$s^K_{lin\text{-}5}$	1km	13203.96	5768.14	3931.84	23.91	41.49	44.51
	10km	4859.58	210.64	38.61	19.69	41.16	44.51
	100km	72.65	37.88	25.65	19.73	**33.40**	44.51

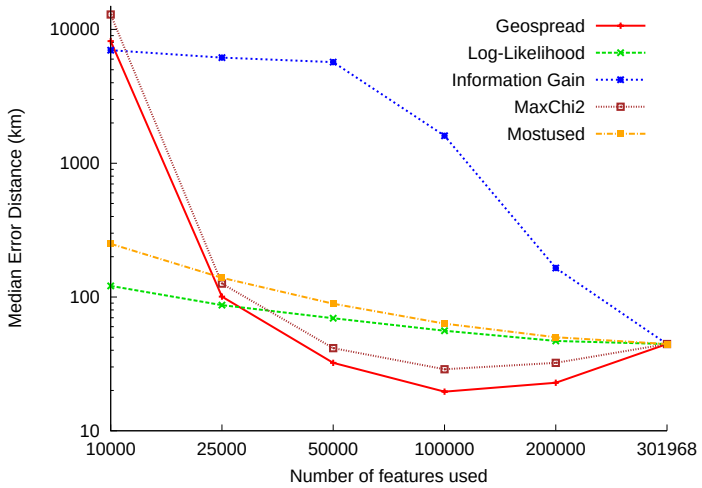


Figure 5.12: Comparison of the baseline methods on the Flickr test set.

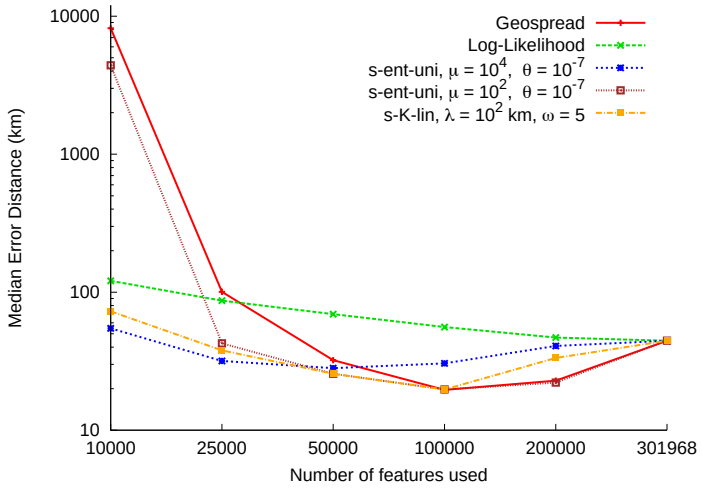


Figure 5.13: Comparison of the best performing methods based on KDE and Ripley's K statistic with the best performing baseline methods on the Flickr test set.

least 25k terms are selected, it performs surprisingly well in the case of aggressive term selection: when only 10k terms are selected the most-used method leads to a median distance of 251.0 km, as opposed to 5314.9 km for geospread. The initial performance of the log-likelihood method is remarkable, but the optimal result for log-likelihood is substantially worse than the optimal performance of geospread and χ^2 (at 50k features).

In Figure 5.13 we compare some of the best performing methods based on KDE and Ripley's K against the log-likelihood and geospread measures. Figure 5.14 compares the same methods on the Wikipedia test set. What can be observed from these figures is that the optimal result for the geospread measure is approximately the same as the optimal result for the KDE and K based methods. However, when fewer than 50k terms are selected, the geospread measure performs considerably worse, especially on the Flickr test set. From an application perspective, this means that the geospread measure is more sensitive to an appropriate choice of the number of features. More fundamentally, it means that the geospread measure is not suitable when aggressive term selection is needed. For instance, in a service which filters a continuous stream of Twitter posts based on their estimated geographic location, the efficiency of the georeferencing process plays a crucial role. This efficiency depends to a large extent on the number of terms that are selected. The results from Figures 5.13 and 5.14 show that the methods we propose in this paper can be used to obtain a reasonable median error distance when using only 10k terms. The log-likelihood measure performs remarkably well on the Flickr test set, but fails to confirm this performance on the Wikipedia test set.

Table 5.2: Percentage of photos for which the coordinates that are found are within 5km and 100km of their true location, when selecting 10k, 25k, 50k, 100k and 200k tags.

	Accuracy at 5km					Accuracy at 100km				
	10k	25k	50k	100k	200k	10k	25k	50k	100k	200k
χ^2	17.23	25.64	28.25	31.76	32.45	31.67	48.58	56.31	59.57	58.34
log-likelihood	25.22	27.45	28.75	29.91	30.73	48.32	50.96	52.67	54.17	55.32
information gain	2.95	5.45	9.54	15.95	25.81	6.48	11.29	18.91	29.89	46.68
most used	24.13	27.16	28.89	30.06	30.72	42.38	47.57	50.72	53.04	54.78
geospread	24.94	29.64	33.21	35.4	34.53	38.75	49.11	56.79	62.46	61.6
$s^{ent}_{Dir}, \mu = 1000, \theta = 10^{-7}$	29.02^{α}	32.31^{α}	34.11^{α}	34.74^{β}	33.03^{β}	47.08^{α}	53.75^{α}	57.77^{α}	60.14^{β}	58.76^{β}
$s^{ent}_{Dir}, \mu = 100, \theta = 10^{-7}$	28.89^{α}	31.95^{α}	34.04^{α}	35.25^{β}	34.61^{β}	44.88^{α}	52.53^{α}	57.84^{α}	62.06^{β}	$\mathbf{61.78}^{\beta}$
$s^{ent}_{uni}, \mu = 1000, \theta = 10^{-7}$	31.15^{α}_{β}	$\mathbf{33.67}^{\alpha}_{\beta}$	$\mathbf{35.03}^{\alpha}_{\beta}$	34.32^{α}_{β}	31.47^{β}	52.08^{α}_{β}	$\mathbf{57.57}^{\alpha}_{\beta}$	$\mathbf{60.81}^{\alpha}_{\beta}$	60.91^{α}_{β}	56.51^{β}
$s^{ent}_{uni}, \mu = 100, \theta = 10^{-7}$	29.06^{α}	32.45^{α}_{β}	34.38^{α}_{β}	35.35^{α}_{β}	$\mathbf{34.7}^{\beta}$	45.7^{α}	53.83^{α}_{β}	58.88^{α}_{β}	62.5^{α}_{β}	$\mathbf{61.78}^{\beta}$
$s^{kl}, \mu = 1000, \theta = 10^{-7}$	27.85^{α}_{β}	28.9^{α}_{β}	29.65^{α}_{β}	30.25^{β}	30.7^{β}	49.83^{α}_{β}	51.98^{α}_{β}	53.49^{α}_{β}	54.8^{β}	55.47^{β}
$s^{kl}, \mu = 100, \theta = 10^{-7}$	30.53^{α}_{β}	31.08^{α}_{β}	30.48^{α}_{β}	30.46^{β}	30.83^{β}	$\mathbf{55.02}^{\alpha}_{\beta}$	55.28^{α}_{β}	54.37^{α}_{β}	54.74^{β}	55.53^{β}
$s^{\chi^2}, \theta = 10^{-1}$	$\mathbf{31.44}^{\alpha}_{\beta}$	32.93^{α}_{β}	33.27^{α}_{β}	32.88^{α}_{β}	31.56^{β}	52.98^{α}_{β}	56.36^{α}_{β}	57.93^{α}_{β}	58.29^{α}_{β}	56.53^{β}
s^{K}_{log-1}	29.78^{α}	32.38^{α}	34.41^{α}_{β}	35.52^{α}_{β}	31.91^{β}	50.01^{α}	55.42^{α}	59.87^{α}_{β}	62.68^{α}_{β}	57.25^{β}
s^{K}_{lin-5}	30.49^{α}	32.71^{α}	34.44^{α}_{β}	$\mathbf{35.53}^{\alpha}_{\beta}$	32.47^{β}	51.71^{α}	56.35^{α}	59.9^{α}_{β}	$\mathbf{62.69}^{\alpha}_{\beta}$	58.18^{β}

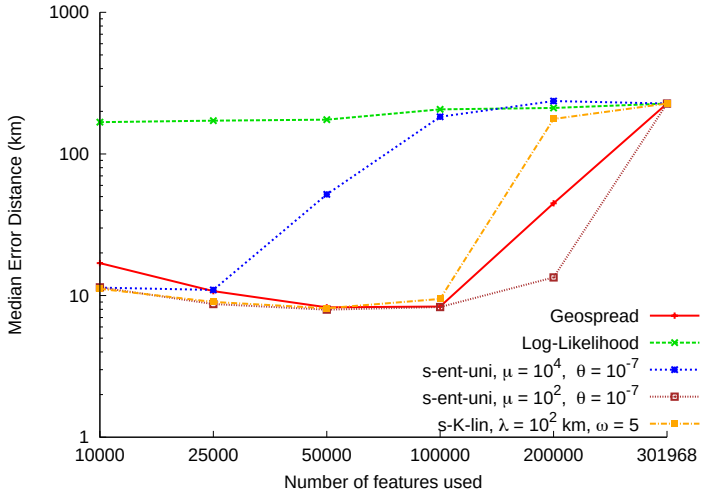


Figure 5.14: Comparison of the best performing methods based on KDE and Ripley's K statistic with the best performing baseline methods on the Wikipedia test set.

Table 5.2 compares the different methods in terms of their ability to find coordinates for photos within 5km and 100km of their true location. The overall conclusions remain the same. For more aggressive forms of term selection, up to 50k tags, s^{ent}_{Dir}, s^{ent}_{uni} and s^{χ^2} outperform each of the baselines, provided that the bandwidth is chosen sufficiently small for s^{ent}_{Dir} and s^{ent}_{uni}, and sufficiently large for s^{χ^2}. Furthermore, if μ is chosen sufficiently small, s^{ent}_{Dir} and s^{ent}_{uni} can compete with geospread at 100k and 200k tags. For results marked with $^{\alpha}$ and $_{\beta}$, the corresponding method improves, respectively, geospread and log-likelihood in a statistically significant way[6] with a p-value $< 10^{-12}$.

Figure 5.15 presents a more detailed view on the errors made by different methods on the Flickr test set when using 10k terms. The value reported for test item n (of the X-axis) is the error made for the photo whose estimated location is the n^{th} best among all 100k test photos. In particular, the value reported for 50k is the median error, which we have mainly focused on so far, while the values for 25k and 75k correspond to the first (Q1) and third (Q3) quartile. Figures 5.16 and 5.17 show the result of using 50k and 100k terms respectively. As could be expected, in light of the aforementioned results on the median error distance, geospread performs considerably worse than several of our methods when using 10k terms. More surprisingly, when using 50k terms, our methods perform an order of magnitude better than geospread between the 65^{th} and 80^{th} percentile, and perform comparably elsewhere. When using 100k terms, the methods based on

[6]To evaluate the statistical significance, we have used the sign test as its sensitivity to outliers makes the Wilcoxon signed-rank test less reliable in this context.

Ripley's K perform almost identical to geospread overall. Log-likelihood is outperformed by the proposed methods overall, although it performs remarkably well around the third quartile.

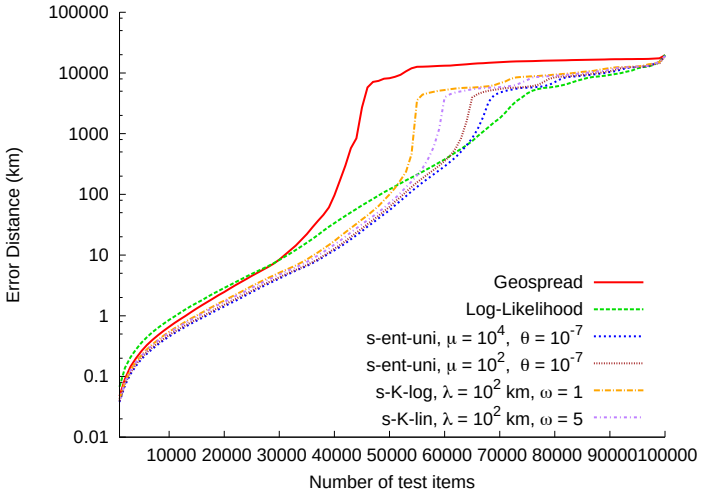


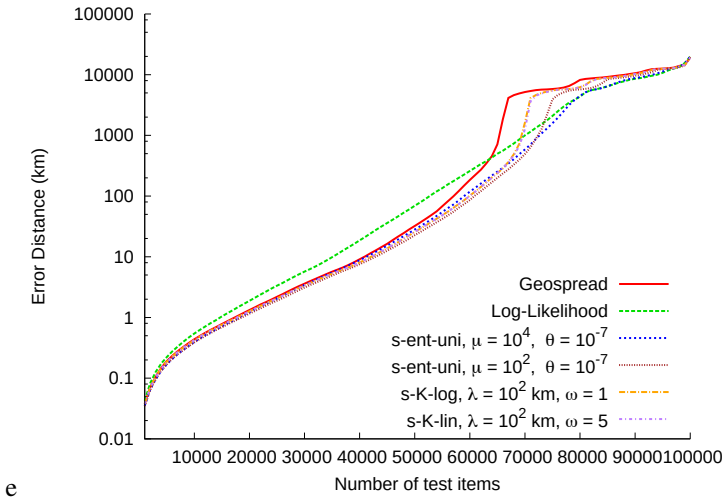
Figure 5.15: Detailed comparison of the errors made on the Flickr test set by the different methods when using 10k terms.

The full rankings of a number of selected methods from this paper are available online[7].

5.6 Conclusions

In this paper, we have studied the use of kernel density based methods and methods based on Ripley's K statistic for selecting location-relevant tags from a collection of georeferenced Flickr photos. Similar in motivation to standard term selection methods, our aim was to improve the effectiveness and efficiency of classifiers that are used to estimate the geographic location of textual resources in general, and tagged Flickr photos in particular. Experimental results clearly reveal that standard term selection methods perform poorly, being outperformed both by the geospread method from [9] and by the methods introduced in this paper. Furthermore, our methods were found to outperform the geospread method by up to two orders of magnitude when aggressive term selection is needed. When a larger percentage of the terms can be retained, our methods perform comparably. In this paper, whilst we have considered a fairly broad spectrum for the values of μ and θ, we have avoided tailoring these values to the problem. However, we highlight that there is

[7] https://github.com/ovlaere/spatial_feature_selection/tree/master/rankings



e

Figure 5.16: Detailed comparison of the errors made on the Flickr test set by the different methods when using 50k terms.

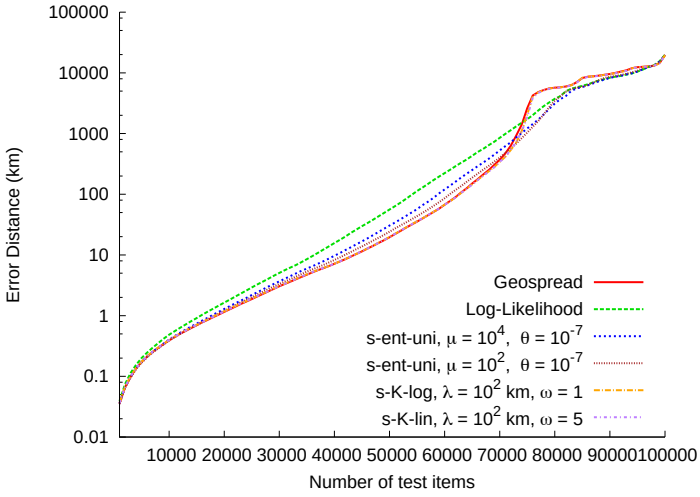


Figure 5.17: Detailed comparison of the errors made on the Flickr test set by the different methods when using 100k terms.

potential for significant improvement in the results through the optimal selection, for the KDE based methods, of μ and θ for the particular data.

The importance of our results is two-fold. On the one hand, our methods allow considerably better results when only few features can be selected. This is paramount in applications where a large number of resources needs to be georeferenced in limited time, e.g. to implement a geographic filter for a stream of Twitter posts, news articles or photos. Second, the analysis of various configurations for our methods clearly shows that identifying tags whose occurrence is correlated to spatial location is not sufficient, and that we should rather select those terms that appear in one or a few clear clusters. This observation is also consistent with the strong performance of the geospread measure from [9].

References

[1] P. Serdyukov, V. Murdock, and R. van Zwol. *Placing flickr photos on a map*. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 484–491, 2009.

[2] C. De Rouck, O. Van Laere, S. Schockaert, and B. Dhoedt. *Georeferencing Wikipedia pages using language models from Flickr*. In Proceedings of the Terra Cognita 2011 Workshop, pages 3–10, 2011.

[3] T. Rattenbury, N. Good, and M. Naaman. *Towards automatic extraction of event and place semantics from flickr tags*. In Proceedings of the 30th Annual International ACM SIGIR Conference, pages 103–110, 2007.

[4] L. Hollenstein and R. Purves. *Exploring place through user-generated content: Using Flickr to describe city cores*. Journal of Spatial Information Science, 1(1):21–48, 2010.

[5] A. Popescu and G. Grefenstette. *Deducing trip related information from Flickr*. In Proceedings of the 18th International Conference on World Wide Web, pages 1183–1184, 2009.

[6] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. *A latent variable model for geographic lexical variation*. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1277–1287, 2010.

[7] Y. Yang and J. O. Pedersen. *A Comparative Study on Feature Selection in Text Categorization*. In Proceedings of the 14th International Conference on Machine Learning, pages 412–420, 1997.

[8] M. Rogati and Y. Yang. *High-performing feature selection for text classification*. In Proceedings of the 11th International Conference on Information and Knowledge Management, pages 659–661, 2002.

[9] C. Hauff and G.-J. Houben. *WISTUD at MediaEval 2011: Placing task*. In Working Notes of the MediaEval Workshop, 2011.

[10] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[11] B. Ripley. *Spatial statistics*. Wiley, 1981.

[12] D. A. Smith. *Detecting and Browsing Events in Unstructured text*. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 73–80, 2002.

[13] R. Swan and J. Allan. *Extracting significant time varying features from text*. In Proceedings of the Eighth International Conference on Information and Knowledge Management, pages 38–45, 1999.

[14] H. L. Chieu and Y. K. Lee. *Query based event extraction along a timeline*. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 425–432, 2004.

[15] Q. Zhao, P. Mitra, and B. Chen. *Temporal and information flow based event detection from social text streams*. In Proceedings of the 22nd Aational Conference on Artificial Intelligence, pages 1501–1506, 2007.

[16] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. *Parameter free bursty events detection in text streams*. In Proceedings of the 31st International Conference on Very Large Data Bases, pages 181–192, 2005.

[17] O. Z. Chaudhry and W. A. Mackaness. *Automated extraction and geographical structuring of Flickr tags*. In Proceedings of the 4th International Conference on Advanced Geographic Information Systems, Applications, and Services, pages 134–139, 2012.

[18] E. Moxley, J. Kleban, and B. S. Manjunath. *Spirittagger: a geo-aware tag suggestion tool mined from Flickr*. In Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, pages 24–30, 2008.

[19] N. O'Hare and V. Murdock. *Modeling locations with social media*. Information Retrieval, pages 1–33, 2012.

[20] O. Van Laere, S. Schockaert, and B. Dhoedt. *Finding locations of flickr resources using language models and similarity search*. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval, pages 48:1–48:8, 2011.

[21] Z. Cheng, J. Caverlee, and K. Lee. *You are where you tweet: a content-based approach to geo-locating Twitter users*. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pages 759–768, 2010.

[22] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. *Spatial variation in search engine queries*. In Proceedings of the 17th International Conference on World Wide Web, pages 357–366, 2008.

[23] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. *Web-a-where: geotagging web content*. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 273–280, 2004.

[24] B. Wing and J. Baldridge. *Simple supervised document geolocation with geodesic grids*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 955–964, 2011.

[25] C. Brunsdon. *Estimating probability surfaces for geographical point data: An adaptive kernel algorithm*. Computers & Geosciences, 21(7):877 – 894, 1995.

[26] C. B. Jones, R. S. Purves, P. D. Clough, and H. Joho. *Modelling vague places with knowledge from the Web*. Int. J. Geogr. Inf. Sci., 22:1045–1065, January 2008.

[27] T. Rattenbury and M. Naaman. *Methods for extracting place semantics from Flickr tags*. ACM Transactions on the Web, 3(1):1–30, 2009.

[28] T. Dunning. *Accurate methods for the statistics of surprise and coincidence*. Comput. Linguist., 19(1):61–74, March 1993.

[29] Z. I. Botev, J. F. Grotowski, and D. P. Kroese. *Kernel density estimation via diffusion*. Annals of Statistics, 38(5):2916–2957, 2010.

[30] D. Pfeiffer, T. Robinson, M. Stevenson, K. Stevens, D. Rogers, and A. Clements. *Spatial analysis in epidemiology*. Oxford University Press New York, 2008.

[31] P. Haase. *Spatial pattern analysis in ecology based on Ripley's K-function: Introduction and methods of edge correction*. Journal of Vegetation Science, 6(4):575–582, 1995.

[32] O. Van Laere, S. Schockaert, and B. Dhoedt. *Ghent university at the 2010 Placing Task*. In Working Notes of the MediaEval Workshop, 2010.

6

Georeferencing Flickr photos using language models at different levels of granularity: an evidence based approach

"If you torture the data long enough, it will confess."

– Ronald Coase (1910 -), Nobel Prize in Economics (1991).

⋆ ⋆ ⋆

This chapter provides an overview of a multi-scale approach to georeferencing Flickr videos. We present an adaptive technique that assigns locations to photos at the right level of granularity, or, in some cases, even refrains from making any estimations regarding location at all. To this end, we consider the idea of training language models at different levels of granularity, and combining the evidence provided by these language models using Dempster and Shafer's theory of evidence. We provide experimental results which clearly confirm that the increased spatial awareness that is thus gained allows us to make better informed decisions, and moreover increases the overall accuracy of the individual language models.

⋆ ⋆ ⋆

Olivier Van Laere, Steven Schockaert and Bart Dhoedt.

Journal of Web Semantics, Volume 16 (1): 17-31, 2012. Originally submitted, April 2011. Major revision submitted December 2011. Minor revision submitted April 2012. Accepted, May 2012.

Abstract The topic of automatically assigning geographic coordinates to Web 2.0 resources based on their tags has recently gained considerable attention. However, the coordinates that are produced by automated techniques are necessarily variable, since not all resources are described by tags that are sufficiently descriptive. Thus there is a need for adaptive techniques that assign locations to photos at the right level of granularity, or, in some cases, even refrain from making any estimations regarding location at all. To this end, we consider the idea of training language models at different levels of granularity, and combining the evidence provided by these language models using Dempster and Shafer's theory of evidence. We provide experimental results which clearly confirm that the increased spatial awareness that is thus gained allows us to make better informed decisions, and moreover increases the overall accuracy of the individual language models.

6.1 Introduction

In addition to topical relevance, the geographic scope of a web resource is often paramount for assessing its relevance. Inspired by this observation, geographic information retrieval (GIR) systems attempt to identify spatial constraints in queries, and to determine which web resources satisfy them [1, 2]. This requires appropriate, structured geographic background information, which is available in the form of gazetteers. However, as gazetteers are often restricted to administrative places or are otherwise incomplete, many of the names people use to refer to places (i.e. vernacular place names) are not recognized. Moreover, in determining the geographic scope of a web resource, other terms than toponyms may play a key role (e.g. the names of local events). As a result, there has been a recent interest in the automated acquisition of geographic knowledge from online resources which are already georeferenced, e.g. utilizing information provided by users in tagging-based systems such as Flickr [3–6], other types of social websites [7, 8], or even local business directories such as Yahoo! local [9]. What is common to these approaches is that they rely on resources containing both geographic coordinates and textual descriptions (typically in the form of tags) to find correlations between locations and linguistic descriptions. These correlations are then used to obtain *geographic information* in the sense of [11–13], i.e. tuples of the form $< x, y, z, t, U >$ where U represents a 'thing' which was present at location (x, y, z) at time t. Note that in the aforementioned works U is referred to by some web object; e.g. a Flickr photo

or Twitter post refers to the presence of a user at a particular location.

Given this importance of large-scale repositories of georeferenced resources, it is of interest to increase the number of resources for which appropriate geo-annotations exist. In the case of Flickr, for instance, coordinates are only available for a small fraction[1]. A number of recent research efforts have been directed towards automatically finding (approximate) coordinates of Flickr photos [2, 14, 16]. The importance of this task is twofold. On one hand, it shows how we may directly georeference online resources, without the intermediate construction of a gazetteer or other forms of explicit spatial semantics of toponyms. On the other hand, it allows to make a larger number of georeferenced Flickr photos available, which is interesting per se (e.g. to allow spatial browsing by displaying them on a map). Note that the idea of using Flickr tags to derive geo-annotations, as a form of semantic information about a photo, fits within a broader trend to use Web 2.0 data sources to bootstrap the semantic web. For example [17] suggests building *collective knowledge systems* by integrating user-contributed content from the Social Web and machine-gathered (semantic) data. Taking this idea one step further, the *DBPedia Mobile* client proposed in [18] allows a user to browse location related information and semantically interlinked data sources, but at the same time also to contribute to the overall geospatial semantic web by publishing content that is linked with nearby DBPedia resources.

Existing work indicates that language models are particularly suitable for the task of assigning coordinates to Flickr photos [14, 19]. The geographic space is then discretized into a set of disjoint areas. After training a language model for each of these areas, we may determine which one is most likely to contain the true location of a given photo. A drawback of this approach is that it must be decided a priori what is the most suitable granularity at which the location of each photo should be determined. Clearly, such a view is at odds with the observation that the tags of some photos are more indicative of a specific place (e.g. *Central Park, New York*) than others (e.g. *picnic*).

The solution we propose in this paper is to train language models at different levels of granularity, and subsequently decide the most appropriate granularity level for each individual photo. Although we then still need to choose a specific number of clusters for each granularity level, this avoids having to fix the overall scale at which each photo should be georeferenced. In this decision, there is a trade-off between accuracy and informativeness. Essentially, we choose the finest granularity at which the most likely area is sufficiently probable. In contrast to standard language modeling approaches, the actual probabilities that come out of the language models thus become important, rather than only the ranking that is imposed by them. Since such probabilities are known to be poorly calibrated, in

[1] http://www.flickr.com/map/ shows that around 168M photos are geotagged of over 6.46 billion photos (http://www.flickr.com/explore) on Flickr. Accessed on December 6th, 2011.

this paper, we study the effect of two forms of post-processing that are applied to these probabilities. First, we consider a standard approach for calibrating classifier probabilities, based on the well-known PAV (pair-adjacent violators) algorithm. The second form of post-processing relies on the spatial dimension of the problem setting. In particular, we propose an approach based on Dempster and Shafer's theory of evidence [20, 21], which allows us to deal with probabilistic information at different levels of granularity in a natural way. Moreover, the theory dictates how evidence coming from different sources — in this case the language models of areas at different granularity levels — can be combined.

The paper is structured as follows. First, in Section 6.2 we explain how our training and test data was selected, what relevant meta-data is available for Flickr photos, and which preprocessing we have performed. Next, Section 6.3 recalls the basic approach to georeferencing Flickr photos based on language models, and it explains how the resulting probabilities can be calibrated. The core of our approach is presented in Section 6.4, where we show how the probabilities produced by language models may be encoded as belief functions in the sense of Shafer, and how these belief functions may be combined with each other to arrive at a single belief function capturing all available evidence. Section 6.5 then explains how we may use belief functions in practice. Subsequently, Section 6.6 presents our experimental findings. Finally, we provide an overview of related work and conclude.

This paper is a substantially revised and extended version of [22]; the main extensions are as follows. First, the belief functions are now built from calibrated language model probabilities, whereas we used the raw probabilities in [22]. Second, we now consider more combination operators, and a different decision rule based on pignistic probability. Furthermore, to have a better mapping among different granularity levels, we now use one hierarchical clustering, rather an independent flat clustering for each level. Finally, the experimental results have been significantly extended, using a more representative data set.

6.2 Data acquisition and preprocessing

For each photo that is uploaded to its website, Flickr maintains several types of meta-data, which can be obtained via its publicly available API. In this paper, two types of meta-data will be relevant: descriptive tags that have been provided by the photo owners, and for some photos, information about where they were taken. The location information includes a geographical coordinate (latitude and longitude), and information about the accuracy of the location, encoded as a number between 1 (world-level) and 16 (street-level).

The data set we have used consists of two parts. The first part contains the 3 185 343 photos that were provided to the participants of the 2010 MediaEval

Table 6.1: Size of the considered data sets

Training set	2 176 719 photos
Calibration set	1 038 612 photos
Test set	50 000 photos
Total	3 265 331 photos

Table 6.2: Mean and standard deviation for the number of tags per photo in each data set.

Data set	Mean	Standard deviation
training set	9.34	8.24
calibration set	9.27	8.07
test set	9.20	7.95

Placing Task[2], a recent benchmarking initiative on the topic of automatically georeferencing Flick videos. In July 2010, we crawled Flickr in order to expand this initial data set. The query used for this additional crawl constrained the resulting photos to those with an accuracy of at least 12, to ensure that all coordinates were meaningful w.r.t. within-city location. Once retrieved, photos that did not contain any tags or whose coordinates were not valid were removed from the collection. As a result, we obtained an additional data set containing the 5 500 368 most recently georeferenced images (at that time). Combining these two sets resulted in a data set consisting of 8 685 711 georeferenced photos covering more or less the entire world.

In a preprocessing phase, we removed duplicates, i.e. photos of the same user that have an identical tag set (to reduce the impact of bulk uploads [14]). Once filtered, the remaining data set of 3 265 331 photos was divided into a training set of 2 176 719 photos (2/3rd), a separate training set of 1 038 612 photos (1/3rd - 50K) that will be used for calibration of the probabilities, and a test set of 50 000 photos. When separating training data from calibration and test data, we ensured that all photos from the same user were either in the training set, or in the calibration and test sets (to avoid an unfair exploitation of user-specific tags [23]). Tables 6.1 and 6.2 provide some characteristics of the different data sets. A plot of the coordinates of the photos from the training set is shown in Figure 6.1.

The task of estimating the location where a photo was taken can be seen as a classification problem: for each unseen photo t from the test set, we then determine which area a from a given set of areas $\mathcal{A}$ is most likely to contain this location. To create this set of areas $\mathcal{A}$, a k-medoids clustering algorithm (PAM - Partitioning Around Medoids) with geodesic distance was used to cluster the locations of the photos in the training set into 2000 disjoint areas. Note that the k-medoids algorithm was preferred over k-means as it handles the occurrence of outliers better.

[2]http://www.multimediaeval.org/mediaeval2010/placing/index.html

Table 6.3: Mean and standard deviation of the size of the clusters in terms of kilometers.

Granularity	Mean (km)	Standard deviation (km)
50	529.91	457.74
250	177.84	180.62
500	113.44	117.97
1000	68.58	70.76
2000	39.68	41.82

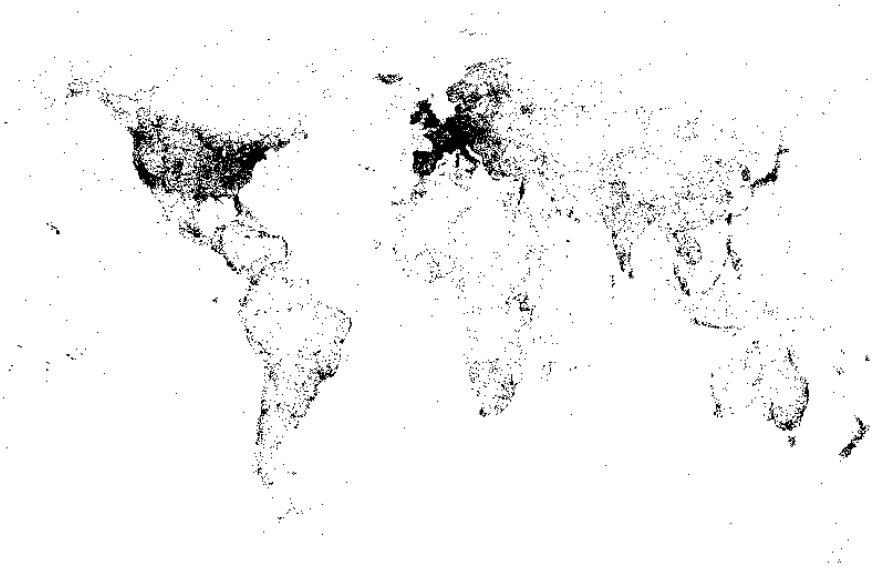


Figure 6.1: Plot of the training set

Among all coordinates, the initial k medoids are randomly chosen. In a subsequent step, initial clusters are obtained by associating the remaining coordinates to the closest medoid (in terms of geodesic distance). Next, for each cluster C, the new medoid is chosen as the element $c \in C$ minimizing

$$\sum_{c' \in C} d(c, c')$$

where the cluster C is identified with its set of coordinates, and d refers to geodesic distance. New clusters can then be obtained from these medoids by again assigning each coordinate to the closest medoid. This process is repeated until the cluster configuration does not change anymore. In this paper, we will consider different levels of granularity, with 2000 areas being the finest level. To obtain coarser granularity levels, we subsequently used agglomerative hierarchical clustering on this

initial clustering, leading to clusterings into 1000, 500, 250 and 50 areas. This step of agglomerative clustering was accomplished by repeatedly merging those two clusters whose medoids were closest to each other w.r.t. geodesic distance. Note that each cluster at one of the coarser granularity levels then exactly corresponds to the union of one or more of the areas of the finest clustering. Note that alternative clustering algorithms, such as a grid based approach [46], mean shift clustering [14] or even a classification based on administrative boundaries can be used for this task; all we require is that each cluster from a coarser granularity level can be seen as the union of one or more clusters from the finest clustering. Examples of our clusterings are shown in Figure 6.2 and Figure 6.3, showing only the clusters located in Europe for clarity. To illustrate the characteristics of the different granularity levels, Table 6.3 provides the mean and standard deviation of the size of the clusters, where the size of a cluster C is taken to be the maximal distance between the medoid and any other member of the cluster.

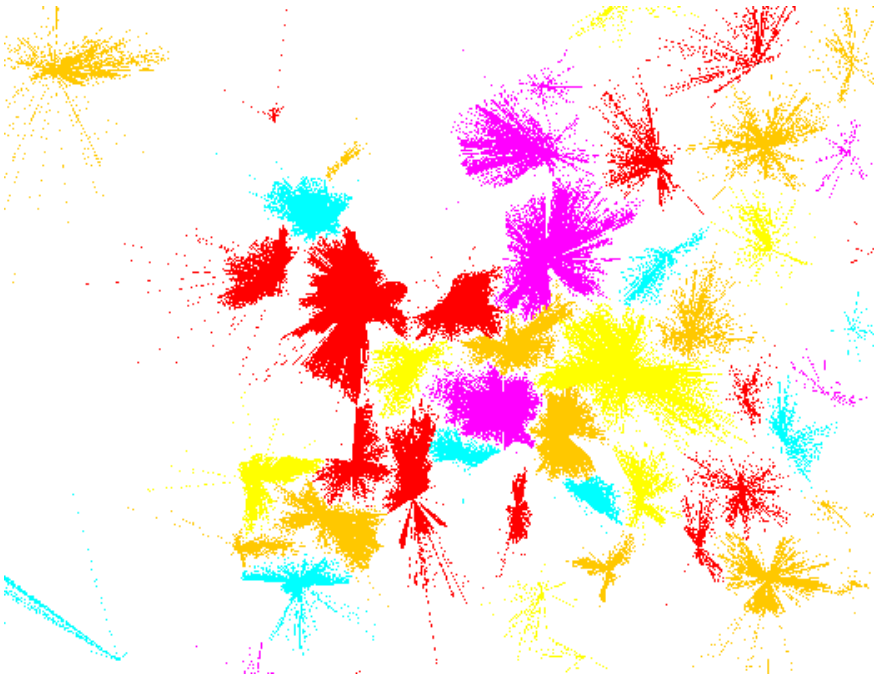


Figure 6.2: Coarse clustering of Europe ($|\mathcal{A}| = 250$)

Next, a vocabulary V consisting of 'interesting' tags is compiled, which are tags that are likely to be indicative of geographic location. We used χ^2 feature selection to determine for each area in $\mathcal{A}$ the m most important tags[3]. The vocab-

[3]Initial experiments have shown χ^2 feature selection to perform slightly better than mutual infor-

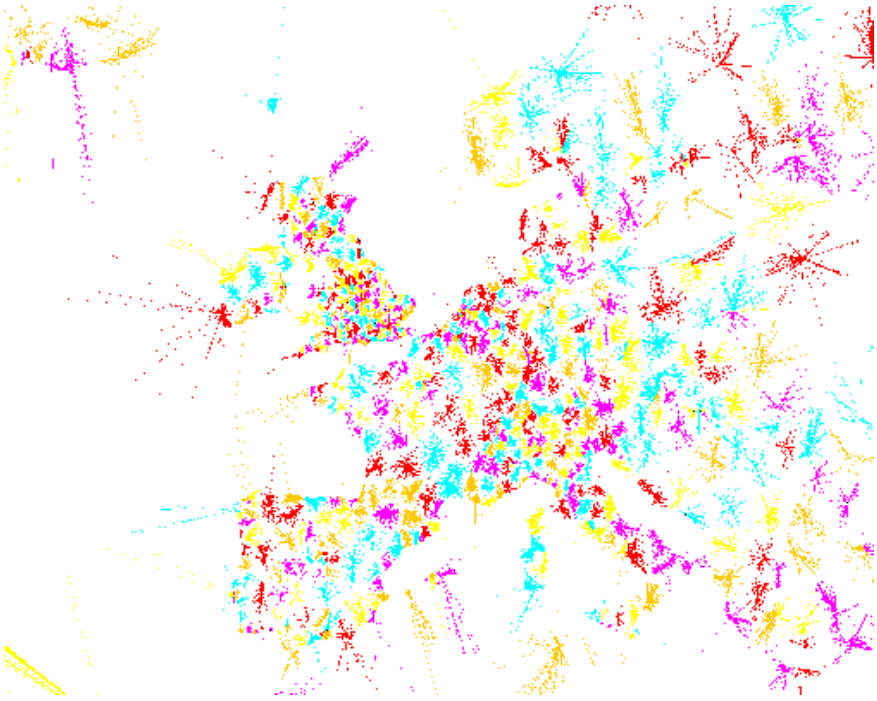


Figure 6.3: Fine clustering of Europe ($|\mathcal{A}|$ = 2000)

ulary V was then obtained by taking for each area a, the m tags with highest χ^2 value. The m values which we have used are 62 500 for the coarsest clustering, 12 500, 2 500, 500 for the intermediate resolutions and 100 for the finest clustering level. This choice of features ensures that the language models, introduced next, require approximately the same amount of memory space for each clustering level[4].

6.3 Calibrated language models for estimating location

6.3.1 Language models

Let $\mathcal{A}$ be a set of (disjoint) areas, obtained by clustering the locations of the photos in our training set. For the ease of presentation, we identify an area $a \in \mathcal{A}$ with the corresponding set of photos that were taken in it. Given a previously unseen

mation on this task.

[4]Space requirements increase quadratically with the number of clusters.

photo x, we try to determine in which area x was most likely taken by comparing its tags with those of the images in the training set. Previous work [2, 14, 19] has revealed that probabilistic (unigram) language models [24] are particularly useful to this end. The probability $p(a|x)$ that image x was taken in area a is then taken to be proportional to

$$p(a|x) \propto p(a) \cdot \prod_{t \in x} p(t|a) \tag{6.1}$$

which corresponds to using a multinomial Naive Bayes classifier to assign areas to photos. The prior probability $p(a)$ of area a can be estimated using the maximum likelihood method:

$$p(a) = \frac{|X_a|}{N}$$

To avoid a zero probability when x contains a tag that does not occur in area a, some form of smoothing is needed when estimating $p(t|a)$. Let $D_a(t)$ be the occurrence count of tag t in area a. The total tag occurrence count D_a of area a is then defined as follows:

$$|D_a| = \sum_{t \in V} D_a(t)$$

where V is the vocabulary that was obtained after feature selection, as explained in Section 6.2. One possible smoothing method is Bayesian smoothing with Dirichlet priors, in which case we have ($\mu > 0$):

$$p(t|a) = \frac{D_a(t) + \mu\, p(t|C)}{|D_a| + \mu}$$

in which the probabilistic model of the collection $p(t|C)$ is defined using maximum likelihood:

$$p(t|C) = \frac{\sum_{a' \in \mathcal{A}} D_{a'}(t)}{\sum_{a' \in \mathcal{A}} \sum_{t' \in V} D_{a'}(t')}$$

Another possibility is to use Jelinek-Mercer smoothing, in which case we have ($\lambda \in [0, 1]$):

$$p(t|a) = \lambda \frac{D_a(t)}{|D_a|} + (1 - \lambda)\, p(t|C)$$

We have experimentally found these two smoothing techniques to yield comparable results (for optimal values of the parameters $\mu = 1750$ and $\lambda = 0.80$), although Bayesian smoothing was found to be more robust w.r.t. the choice of the parameter. These findings conform to experimental results in other areas of information retrieval [25, 26], and to earlier work on georeferencing Flickr photos [14].

As we focus on the effect of different granularity levels in this paper, we restrict ourselves to a rather standard language modelling approach. Note, however, that the model presented in this section can be refined in different ways, using additional information about the owner, information from visual features, etc. For example, [15] and [44] use the home location of the user, while [54] uses information about her social network. As another form of refinement, in [14] a location-aware from of smoothing is used.

6.3.2 Calibration

In principle, an estimation of the actual value of $p(a|x)$, for all $a \in \mathcal{A}$, is found from (6.1) after normalization. However, it is well-known that Naive Bayes does not produce well-calibrated probability estimates [27]. As our approach will strongly depend on the actual values of the probability estimates, we need to apply some form of calibration. In [28], an approach called *binning* is shown to produce such well-calibrated probabilities. In [29], an extension of this method based on the PAV (pair-adjacent violators [30]) algorithm is proposed, which we have adopted in our experiments. In particular, let us write $n(a|x)$ for the normalised Naive Bayes output, i.e.:

$$n(a|x) = \frac{score(a|x)}{\sum_{a' \in \mathcal{A}} score(a'|x)} \tag{6.2}$$

where $score(a|x)$ denotes the estimation of the right-hand side of (6.1).

Some care needs to be taken to avoid underflow or a significant loss of precision, as the values $score(a|x)$ tend to be very small. As usual, these values can be calculated in log-space, i.e.

$$\log score(a|x) = \log p(a) + \sum_{t \in x} \log p(t|a)$$

The normalization cannot be carried out in log-space, so we rewrite the denominator in equation (6.2) in the following way:

$$\log \sum_{a' \in \mathcal{A}} score(a'|x) \tag{6.3}$$

$$= (\log \sum_{a' \in \mathcal{A}} \gamma \cdot score(a'|x)) - \log \gamma \tag{6.4}$$

$$= (\log \sum_{a' \in \mathcal{A}} \exp \log(\gamma \cdot score(a'|x)) - \log \gamma \tag{6.5}$$

$$= (\log \sum_{a' \in \mathcal{A}} \exp(\log(\gamma) + \log(score(a'|x))) - \log \gamma \tag{6.6}$$

By choosing γ sufficiently high, problems of reduced precision can be avoided; we have used

$$\log \gamma = \max_{a' \in \mathcal{A}} \textit{abs}(\log(score(a'|x)))$$

In this way, $\exp(\log(\gamma)+\log(score(a'|x)))$ in equation (6.6) becomes $exp(0) = 1$ for the most plausible areas a', which avoids both underflow and overflow for the probability of those areas. Note that, if underflow occurs for the probability of less plausible areas, this is then because their probability is extremely small compared to the most plausible area, in which case we can safely ignore them.

The PAV algorithm is now used to map the scores $n(a|x_i)$ to accurate probability estimates, as follows [31, 32]:

- Assume that the photos $x_1, ..., x_m$ from the training set are ranked such that $n(a|x_i) \geq n(a|x_{i+1})$ for all i.
- At each stage of the algorithm, a list of bins is maintained. Let us write $B(i, j)$ for the bin that contains the images $x_i, x_{i+1}, ..., x_j$. Initially the list L contains one bin for each photo, i.e. $L = \{B(i, i)|1 \leq i \leq m\}$. For a given bin $B^1 = B(i, j)$, we write $avg(B^1)$ for the percentage of photos in bin B^1 that actually belong to the area a.
- Let $L = (B^1, ..., B^p)$. Until it holds that $avg(B^i) \geq avg(B^{i+1})$ for all i, repeat the following
  1 Find all maximal subsequences of bins $B^i, ..., B^j$ in the list such that $avg(B^r) \leq avg(B^{r+1})$ for all $r \in \{i, i + 1, ..., j - 1\}$.
  2 Replace these subsequences in the list L by the single bin $B = B^i \cup B^{i+1} \cup ... \cup B^j$.

To ensure that meaningful probability estimates are obtained, as an additional step, we also merge each bin containing fewer than 100 items with the bin succeeding it. This is especially important for the first bin, which we otherwise found to provide an unrealistically optimistic estimation. For instance, if the highest ranked photo was correctly georeferenced, the highest bin would always be associated with a probability of 1.

Let $L = (B^1, ..., B^p)$ be the final list of bins that is obtained from this procedure. Each bin B naturally corresponds to an interval $bounds(B) = [\underline{n}, \overline{n}]$ where $\underline{n} = \min_{x \in B} n(a|x)$ and $\overline{n} = \max_{x \in B} n(a|x)$. For a given photo x from the test set, we then determine the bin B for which $bounds(B)$ contains $n(a|X)$, or whose bounds are closest to $n(a|x)$. A probability estimate $p(a|x)$ is then given by $avg(B)$. Note that $\sum_a p(a|x)$ may be different from 1. However, we refrain from normalizing these estimates at this stage, as initial experiments have shown that this may largely nullify the effect of the calibration process.

6.4 Combining language models of different granularity levels

The language modeling approach that was outlined in Section 6.3 is not spatially aware in the sense that e.g. neighboring areas are treated in the same way as areas that are located in different parts of the world. To see why this difference might be important, assume that the probability $p(.|x)$ takes a high value for two different areas a and b. If a and b are adjacent or close to each other, it makes sense to estimate the location of x at a coarser level of granularity, using an area c as result which encompasses both a and b. Indeed, the fact that all areas that are considered plausible are spatially close suggests that our estimation will be near the actual location of x, while the available information is not sufficient to distinguish reliably between a and b. In contrast, when a and b are not close, the choice between a and b is likely to be a problem of disambiguation. In such as case, it makes more sense to first determine the most likely area c at a coarser granularity level, and take a to be the result if c contains a (but not b), and b if c contains b (but not a).

Our solution uses Dempster-Shafer evidence theory [20, 21] to combine the probability distributions obtained from language models that operate at different resolutions. Based on the agreement between fine-grained models and coarse-grained models, we may then try to find the most plausible region in which a photo was taken, at the most appropriate resolution given the available information. Essentially, our approach then finds the smallest region for which all models agree (to a sufficient degree) to contain the true location with high probability.

6.4.1 Belief functions

Let $\{\mathcal{A}_1, ..., \mathcal{A}_k\}$ be different clusterings of the locations in the training set such that $|\mathcal{A}_1| > |\mathcal{A}_2| > ... > |\mathcal{A}_k|$, i.e. $\mathcal{A}_1$ corresponds to the finest clustering and $\mathcal{A}_k$ corresponds to the coarsest clustering. For each clustering, a language model is obtained which (after calibration) results in a probability distribution $p_i(.|x)$ in the universe $\mathcal{A}_i$ for each image x. A key observation is that the spatial extension of each area a in $\mathcal{A}_i$ corresponds to the union of the spatial extensions of a set of areas from the finest level $\mathcal{A}_1$, as the different clusterings have been obtained in a hierarchical fashion. Let us write $areas(a)$ for this set of areas from $\mathcal{A}_1$ that are included in a. Then, if a is the area maximizing $p(.|x)$, we can take this as evidence that the correct area, at the finest level, is among those of the set $areas(a)$. In other words, the probability distributions $p_2, ..., p_k$ naturally correspond to probability distributions on the power set of $\mathcal{A}_1$, i.e. to belief functions on $\mathcal{A}_1$.

Recall that a belief function [21] on a finite universe U is any $2^U \rightarrow [0, 1]$ mapping m satisfying $\sum_{X \subseteq U} m(X) = 1$ and $m(\emptyset) = 0$; belief functions are also called mass assignments. Intuitively, $m(X)$ represents the amount of evidence

that the correct value of some variable is among those in X. Subsets X such that $m(X) > 0$ are called focal elements. Starting from a belief function m, two measures of uncertainty are usually considered:

$$Bel(X) = \sum_{Y \subseteq X} m(Y) \qquad Pl(X) = \sum_{Y \cap X \neq \emptyset} m(Y)$$

for any $X \subseteq U$. The degree of belief $Bel(X)$ can be interpreted as a lower bound on the probability that X contains the correct value, while the degree of plausibility $Pl(X)$ is an upper bound for this probability.

Probability distributions essentially model variability, i.e. the phenomenon that the outcome of a given experiment may not always be the same, but they lack the capability of genuinely modelling epistemic uncertainty, i.e. the uncertainty resulting from a lack of information. For example, suppose that we know with perfect certainty that the outcome of rolling a die was among the values $\{1, 2, 3\}$. In probability theory, we are left with assigning an equal probability to each of these values, i.e. $p(1) = p(2) = p(3) = \frac{1}{3}$. However, this probability distribution is not a faithful representation of the beliefs that we hold: why should we be able to infer that it is twice as likely that the outcome was odd than that the outcome was even, if all we started off with was the knowledge that the outcome was in $\{1, 2, 3\}$. Using belief functions, on the other hand, we can distinguish between the mass assignment m_1 defined by $m_1(\{1, 2, 3\}) = 1$, and the mass assignment m_2 defined by $m_2(\{1\}) = m_2(\{2\}) = m_2(\{3\}) = \frac{1}{3}$. In other words, belief functions are capable of modelling both variability and epistemic uncertainty.

Note that in the special case where all focal elements are singletons, belief functions simply correspond to probability distributions. Specifically, if we define $p(x) = m(\{x\})$, it holds that $P(X) = Bel(X) = Pl(X)$ for every $X \subseteq U$, where P is the probability measure associated with p, and Bel and Pl are the belief and plausibility measures associated with m.

Nonetheless, when it comes to making decisions based on our available beliefs, the choice between m_1 and m_2 may actually not matter. When deciding whether or not to accept a bet, for instance, all we can do is assume an equal probability for each outcome, i.e. apply the maximum entropy principle. The point of using belief functions, however, is to apply this maximum entropy principle *after* all the available evidence is combined. In other words, a difference is made between the *credal* level, which is concerned with modelling the beliefs of an agent, and the decision or *pignistic* level (from the Latin word *pignus* for bet). Specifically, when it comes to decision making based on belief functions, a mass assignement m is often converted into the associated *pignistic probability distribution* p defined by [33]

$$p(x) = \sum_{\emptyset \subset X \subseteq U, x \in X} \frac{m(X)}{|X|}$$

after which decisions may be made using standard approaches (e.g. based on maximizing expected utility).

6.4.2 Obtaining mass assignments

In the context of this paper, the universe U will always be the set of areas (clusters) of the most fine-grained clustering $\mathcal{A}_1$. For a given photo x, the different granularity levels lead to mass assignments $m_1, ..., m_k$ defined as follows. First, at each granularity level i, a set $\mathcal{S}_i$ containing the most likely areas from $\mathcal{A}_i$ is determined. In principle, we could take $\mathcal{A}_i = \mathcal{S}_i$, but in practice, a smaller set $\mathcal{S}_i$ is desirable to keep the approach time- and space-efficient. In our experiments, the set $\mathcal{S}_i$ was obtained by adding areas in decreasing order of likelihood until $\sum_{a \in \mathcal{S}_i} p_i(a|x) \geq \theta$ for some fixed parameter θ_i (e.g. $\theta_i = 0.95$). Recall that the probability estimates $p_i(a|x)$ are not necessarily normalized, i.e. they do not necessarily sum to 1. However, in all but a few cases we have that $\sum_{a \in \mathcal{S}_i} p_i(a|x) < 1$. Then we define:

$$m_i^x(X) = \begin{cases} p_i(a|x) & \text{if } X = \textit{areas}(a) \text{ for } a \in \mathcal{S}_i \\ 1 - \sum_{a \in \mathcal{S}_i} p_i(a|x) & \text{if } X = \mathcal{A}_1 \\ 0 & \text{otherwise} \end{cases} \tag{6.7}$$

Note that the probability $p_i(a|x)$ is translated to the mass $m_i^x(a)$ for areas a in $\mathcal{S}_i$. The remaining mass corresponding to the areas outside $\mathcal{S}_i$, i.e. $\sum_{a \in (\mathcal{A}_i \setminus \mathcal{S}_i)} p_i(a)$ is assigned to the entire universe $\mathcal{A}_1$. This mass will be approximately equal to $1 - \theta_i$ and reflects the probability that we are ignorant about the location of x. Choosing a lower value of θ_i will thus lead to a more cautious and less informative mass assignment.

Finally, in the rare cases where $s^* = \sum_{a \in \mathcal{S}_i} p_i(a|x) \geq 1$, the probability estimates are first normalized as

$$p_i^*(a|x) = \frac{p_i(a|x)}{s^*}$$

and the mass assignment is defined as in (6.7), but based on the normalized estimates $p_i^*(a|x)$ instead of $p_i(a|x)$.

6.4.3 Combining evidence

Different belief functions may encode the evidence provided by different sources, in which case a *combination operator* may be used to obtain a single, combined belief function. In particular, given two belief functions m and m' in a universe U, Dempster [20] proposes to model the combined evidence using the mass assign-

ment $m \oplus m'$ defined as

$$(m \oplus m')(\emptyset) = 0 \tag{6.8}$$

$$(m \oplus m')(X) = \frac{\sum_{Y \cap Z = X} m(Y) \cdot m'(Z)}{1 - \sum_{Y \cap Z = \emptyset} m(Y) \cdot m'(Z)} \tag{6.9}$$

for any subset $\emptyset \subset X \subseteq U$, and provided that

$$\sum_{Y \cap Z = \emptyset} m(Y) \cdot m'(Z) < 1$$

The denominator in (6.9) is a normalization factor, which corresponds to the mass that would normally be assigned to the empty set, i.e. it is a measure of the amount of conflict between m and m'. It can be shown that this combination rule is associative.

By treating the different granularity levels as independent sources, the overall evidence about the location of a photo x may thus be described by the belief function m^x:

$$m^x = m_1^x \oplus m_2^x \oplus ... \oplus m_k^x \tag{6.10}$$

Example 1. *Let us go back to the scenario outlined in the beginning of Section 6.4. In particular, assume that there are only two granularity levels, and $\mathcal{S}_1 = \{a, b\}$ and $\mathcal{S}_2 = \{u, v\}$. At the finest level, we are thus faced with the choice of a or b as the location of a given photo x. First assume that $areas(u)$ contains both a and b. In this case, the focal elements of m^x are $\{a\}$, $\{b\}$, $areas(u)$, $areas(v)$, and $\mathcal{A}_1$; we obtain*

$$\begin{aligned}
m^x(\{a\}) &= K \cdot m_1^x(\{a\}) \cdot (m_2^x(areas(u)) + m_2^x(\mathcal{A}_1)) \\
m^x(\{b\}) &= K \cdot m_1^x(\{b\}) \cdot (m_2^x(areas(u)) + m_2^x(\mathcal{A}_1)) \\
m^x(areas(u)) &= K \cdot m_1^x(\mathcal{A}_1) \cdot m_2^x(areas(u)) \\
m^x(areas(v)) &= K \cdot m_1^x(\mathcal{A}_1) \cdot m_2^x(areas(v)) \\
m^x(\mathcal{A}_1) &= K \cdot m_1^x(\mathcal{A}_1) \cdot m_2^x(\mathcal{A}_1)
\end{aligned}$$

where K is the normalization constant. Assuming that $m_1^x(\mathcal{A}_1)$ and $m_2^x(\mathcal{A}_1)$ are sufficiently small, we have

$$m^x(areas(u)) \approx m^x(areas(v)) \approx m^x(\mathcal{A}_1) \approx 0$$

and thus $K \approx m^x(\{a\}) + m^x(\{b\})$*; we obtain*

$$\begin{aligned}
&Bel(\{a\}) \\
&\approx \frac{m_1^x(\{a\}) \cdot m_2^x(areas(u))}{m_1^x(\{a\}) \cdot m_2^x(areas(u)) + m_1^x(\{b\}) \cdot m_2^x(areas(u))} \\
&= \frac{p_1(a|x)}{p_1(a|x) + p_1(b|x)} \\
&Bel(\{b\}) \\
&\approx \frac{m_1^x(\{b\}) \cdot m_2^x(areas(u))}{m_1^x(\{a\}) \cdot m_2^x(areas(u)) + m_1^x(\{b\}) \cdot m_2^x(areas(u))} \\
&= \frac{p_1(b|x)}{p_1(a|x) + p_1(b|x)} \\
&Bel(areas(u)) \approx 1
\end{aligned}$$

Note that because v *does not overlap with any area of* $\mathcal{S}_1$*, most of the mass* $m_2^x(areas(v))$ *disappears in the normalization constant* K*. If* u *is a clear winner at the second level, i.e.* $p_2(u|x) \gg p_2(v|x)$*, without a clear winner at the first level, we thus obtain strong evidence that the correct location is in* u*, but much weaker evidence for* a *or* b *individually.*

Now consider a second scenario in which $a \in areas(u)$ *while* $b \in areas(v)$*. We then get*

$$m^x(\{b\}) = K \cdot m_1^x(\{b\}) \cdot (m_2^x(areas(v)) + m_2^x(\mathcal{A}_1))$$

and $m^x(\{a\})$*,* $m^x(areas(u))$*,* $m^x(areas(v))$ *and* $m^x(\mathcal{A}_1)$ *as before. Again assuming that* $m_1^x(\mathcal{A}_1)$ *and* $m_2^x(\mathcal{A}_1)$ *are sufficiently small, we have*

$$\begin{aligned}
&Bel(\{a\}) \\
&\approx \frac{m_1^x(\{a\}) \cdot m_2^x(areas(u))}{m_1^x(\{a\}) \cdot m_2^x(areas(u)) + m_1^x(\{b\}) \cdot m_2^x(areas(v))} \\
&Bel(\{b\}) \\
&\approx \frac{m_1^x(\{b\}) \cdot m_2^x(areas(u))}{m_1^x(\{a\}) \cdot m_2^x(areas(u)) + m_1^x(\{b\}) \cdot m_2^x(areas(v))}
\end{aligned}$$

If we moreover again make the assumption that $p_2(u|x) \gg p_2(v|x)$*, we get*

$$\begin{aligned}
Bel(\{a\}) &\approx Bel(areas(u)) \approx 1 \\
Bel(\{b\}) &\approx Bel(areas(v)) \approx 0
\end{aligned}$$

Hence in this case, we do obtain strong evidence for a*. Note that in the latter scenario the evidence from the second granularity level has allowed us to make a decision between* a *and* b*, while in the former scenario it has rather provided a more cautious alternative, avoiding a somewhat arbitrary choice between* a *and* b*.*

The combination rule (6.8)–(6.9) is the first and most widely known combination rule, already proposed by Dempster in the 1960s. It has been argued by several authors that it constitutes the only principled way to combine independent and reliable sources in a conjunctive way [34, 35]. On the other hand, from an application perspective, when the degree of conflict $\sum_{Y \cap Z = \emptyset} m(Y) \cdot m'(Z)$ is close to 1, it is reputed to provide counterintuitive results [36]. Moreover, when the degree of conflict is equal to 1, the result of the combination is not even defined. To cope with this, when using Dempster's rule, we first apply some form of discounting, i.e. each mass assignment m is replaced by the mass assignment m_δ, defined by

$$m_\delta(A) = \delta \cdot m(A)$$

if A is different from the universe U, and

$$m_\delta(U) = \delta \cdot m(U) + (1 - \delta)$$

In our experiments, we use $\delta = 0.99$. Note that this indeed guarantees that the degree of conflict is strictly smaller than 1.

Another solution, which is adopted in the transferable belief model (TBM) of Smets [37], is to simply allow a non-zero mass for the empty set. We thus obtain the following combination operator:

$$(m_1 \odot m_2)(A) = \sum_{B \cap C = A} m_1(B) \cdot m_2(C) \tag{6.11}$$

After the final mass assignment has been determined, the mass of the empty set is than added to the mass of the universe. The resulting combination operator is sometimes called Yager's rule [38] ($\emptyset \subset A \subset U$):

$$(m_1 \otimes' ... \otimes' m_k)(A) = (m_1 \odot ... \odot m_k)(A) \tag{6.12}$$

$$(m_1 \otimes' ... \otimes' m_k)(U) = (m_1 \odot ... \odot m_k)(U) + (m_1 \odot ... \odot m_k)(\emptyset) \tag{6.13}$$

$$(m_1 \otimes' ... \otimes' m_k)(\emptyset) = 0 \tag{6.14}$$

Note that unlike $\otimes$ and $\odot$, the operator $\otimes'$ underlying Yager's combination rule is not associative. Dubois and Prade have proposed the following alternative way of distributing the mass of the empty set [39]:

$$(m_1 \otimes'' m_2)(A) = (m_1 \odot m_2)(A) + \sum_{B \cup C = A, B \cap C = \emptyset} m_1(B) \cdot m_2(C) \tag{6.15}$$

The underlying intuition here is that in the presence of conflicts, we should take the point of view that one of the sources is correct, which leads to a disjunctive combination of conflicting evidence and the requirement that $B \cup C = A$ in the right-hand side of (6.15).

6.5 Using belief functions in geographic information retrieval

By combining $m_1^x, ..., m_k^x$ using either of the combination operators, we obtain a single mass assignment m^x summarizing the available evidence about the location of x. In many cases, some post-processing of this mass assignment will be needed to obtain usable approximations of the location of x, e.g. in the form of a precise point, a precise region (i.e a polygon), or a fuzzy region (i.e. a nested set of polygons). Indeed, unlike simple representations such as points and polygons, mass assignments cannot readily be spatially indexed, which is a prerequisite if we are to use georeferencing of photos to support online location-based querying [1]. Moreover, mass assignments, unlike probability distributions and fuzzy regions, cannot be visualized in a way which is sufficiently intuitive for end users. How exactly x's location should be represented in the result depends on the precise requirements of the application context:

Supporting location-based queries Consider a user indicating an interest in photos that were taken in Manhattan. In such a case, we could simply use the mass assignment m^x of each photo x to calculate the belief or plausibility that x was taken in Manhattan, the latter being represented as a union of elements from $\mathcal{A}_1$. Similarly, if a user is interested in photos that were taken in the vicinity of a particular point-of-interest, we could determine the belief or plausibility that each photo in the collection was taken within a given radius of that point. When the mass assignments have been converted to points (the most likely location of x) or polygons (a confidence region for x) that have been spatially indexed a priori, location-based querying becomes computationally feasible.

Helping users georeference their photos When users upload a photo to Flickr, they have the option to indicate on a map where it was taken. When the user has already provided a number of tags for the photo, it makes sense to analyze these tags, and already zoom in on this map at where the photo was likely taken. In this way, less effort is required by the user, which may lead to more users georeferencing their photos, with a higher accuracy level. This application not only requires the system to determine where to center the map, but also to determine at which zoom level it should be shown. This boils down, conceptually, to finding the smallest area containing the true location of x with a given confidence level, i.e. a confidence region for x.

Visualizing plausible locations In some applications, we may simply provide the user with a visual summary of where a photo was likely taken. One of the most obvious ways to do this is by presenting a heat map, which may

conceptually be seen as a mapping from locations to the unit interval [0,1], i.e. a possibility distribution [40] of locations. This requires to determine an appropriate approximation of m^x.

It seems that from an application point of view, mass assignments are mainly useful (i) to find the most likely area, at a given granularity level, in which the photo was taken, (ii) to find the most fine-grained area that contains the true location of the photo at a given confidence level, and (iii) to obtain a visual summary of the plausible locations. These three uses are discussed below.

6.5.1 Finding the most plausible area

The probability distribution $p_i(.|x)$ obtained by calibrating the language models of the areas in $\mathcal{A}_i$ naturally allows us to determine the most plausible area from $\mathcal{A}_i$, viz. the area a maximizing $p_i(a|x)$. The mass assignment m^x has been obtained by combining p_i with other pieces of evidence (i.e. the probability distributions over the other levels of granularity), and may thus allow us to determine the most plausible location of $\mathcal{A}_i$ in a better-informed way. In general, one could also think of combining p_i with belief functions encoding information from other sources of evidence such as gazetteers or visual feature information to obtain m^x. Obvious decision rules are choosing the area a maximizing the belief measure and choosing the area maximizing the plausibility measure:

$$choose_{Bel}(\mathcal{A}_i, m^x) = \arg\max_{a \in \mathcal{A}_i} Bel(areas(a)) \tag{6.16}$$

$$choose_{Pl}(\mathcal{A}_i, m^x) = \arg\max_{a \in \mathcal{A}_i} Pl(areas(a)) \tag{6.17}$$

A third decision rule uses the pignistic probability measure P^x induced by m:

$$choose_{P}(\mathcal{A}_i, m^x) = \arg\max_{a \in \mathcal{A}_i} P^x(areas(a)) \tag{6.18}$$

6.5.2 Determining confidence regions

Rather than first fixing the granularity level and then determining the most plausible area, it often makes sense to look for the smallest area that contains a given photo x with some predefined confidence level, where *confidence* may be measured in terms of belief, plausibility or pignistic probability. An important question is which areas are to be considered for the result. Either we may restrict ourselves to the areas in $\bigcup_i \mathcal{A}_i$, or we may allow arbitrary subsets of $\mathcal{A}_1$, possibly with the restriction that the chosen subset constitutes a connected (or even convex) region. The solution which we have adopted in the experiments is based on the former choice, which is considerably easier from a computational point of view. Moreover, as the areas in $\bigcup_i \mathcal{A}_i$ have all been obtained from clustering the training data,

they likely correspond to meaningful geographic entities. For instance, if all of the most plausible areas from $\mathcal{A}_1$ are in Manhattan, it often makes more sense to use the entire region of Manhattan as result, rather than the disjoint union of these specific areas within Manhattan. The situation where available information is ambiguous forms an exception to this view: if all we know is that a photo was taken in Washington, it makes sense to represent the result e.g. as the union of Washington D.C. and Washington state.

The procedure to determine a confidence region then becomes the following. First, we check whether our confidence in the most likely area a from $\mathcal{A}_1$ — determined e.g. using $choose_P$, $choose_{Bel}$ or $choose_{Pl}$ — is sufficiently high. This confidence could again be measured in terms of pignistic probability, belief or plausibility. If this is the case, region a is taken as the result. Otherwise, we check whether our confidence in the most likely area from $\mathcal{A}_2$ is sufficiently high, etc. If even our confidence in the most likely area from $\mathcal{A}_k$ is too low, it seems reasonable to acknowledge that no reliable location could be determined for the corresponding photo.

6.5.3 Approximation of mass assignments

Mass assignments have the disadvantage that they are difficult to visualize, and they may require considerable amounts of storage space (which may become problematic at the scale of billions of Flickr images). Therefore, there is an interest in approximating the mass assignments m^x in a way that alleviates these issues, without losing too much relevant information. Ideally, we want an approximation of the mass assignment as a mapping from $\mathcal{A}_1$ to $[0, 1]$ (or some other scale), as such mappings are easy to visualize, e.g. as a heat map. An obvious candidate would be to use the pignistic probability. However, this still has the disadvantage that a value must be stored for each element from $\mathcal{A}_1$. Here we present an alternative solution, which uses possibility theory [40].

The main idea is to determine a nested family of areas $B_1 \subseteq B_2 \subseteq ... \subseteq B_l \subseteq \mathcal{A}_1$, such that $B_1, ..., B_l$ correspond to increasingly more cautious approximations of the location of the photo x. They can be obtained by applying the procedure from Section 6.5.2, using a (fixed) set of different thresholds on the required confidence. In this way, all we have to store are the l regions and the corresponding confidence values. To visualize the mass assignment, we can then simply plot these areas, using gray-scale values that depend on the confidence levels. Moreover, the use of a small number of confidence regions also means that standard spatial indexing methods can be used, e.g. when implementing a system that needs to be able to retrieve all photos that are located in a given area with a predefined confidence.

Note that the nested family $B_1 \subseteq ... \subseteq B_l$ can be seen as a mapping π from

$\mathcal{A}_1$ to $[0, 1]$:

$$\pi(a) = \max_{i=1}^{l} \min \left(B_{l+1-i}(a), \frac{i}{l} \right) \tag{6.19}$$

where we identify the sets B_i with their characteristic mapping for the ease of presentation, i.e. we have $B_i(a) = 1$ iff $a \in B_i$ and $B_i(a) = 0$ otherwise. The mapping π is called a possibility distribution [40], and $\pi(a)$ the degree of possibility that the correct area is a. Where probability distributions can model variability but not epistemic uncertainty, possibility distributions can model epistemic uncertainty but not variability. A situation of complete ignorance can be modeled as $\pi(a) = 1$ for all $a \in \mathcal{A}_1$, whereas in a completely informed situation we have $\pi(a) = 1$ for exactly one $a \in \mathcal{A}_1$ and $\pi(a) = 0$ for all other areas. In general, the degree $\pi(a)$ is interpreted as the degree to which one would be surprised to learn that a is the real value of the considered variable, an interpretation which at least goes back to the work of Shackle [41].

Like probability distributions, possibility distributions also correspond to a special case of belief functions. To clarify this link, first note that with each possibility distribution π, two uncertainty measures Π and N can be associated, defined (in a universe U) by

$$\Pi(X) = \sup_{u \in U} \pi(u)$$
$$N(X) = 1 - \Pi(U \setminus X)$$

Intuitively, $\Pi(X)$ corresponds to the degree to which it is consistent with our beliefs to assume that the correct value is among those in X, whereas $N(X)$ corresponds to the degree to which this is implied by our beliefs. Now, let m be a mass assignment whose focal elements constitute a nested family of sets: $\emptyset \subset X_1 \subset X_2 \subset ... \subset X_l \subseteq U$. With the mass assignment m we can associate the possibility distribution π defined by $\pi(x) = \sum_{x \in X_i} m(X_i) = Pl(\{x\})$ [42]. Then we have that for any $X \subseteq U$, it holds that $Bel(X) = N(X)$ and $Pl(X) = \Pi(X)$. In general, a mass assignment m can be approximated by a possibility distribution in different ways. One approach is to still define $\pi(x) = Pl(\{x\})$, in which case π is called the contour function of m. A second approach is to use a predefined family of nested sets, as we did in (6.19).

Possibility distributions are not only useful for visualization. Their graded nature makes them suitable representations for modeling the boundaries of vague vernacular geographic regions [7, 9, 10]. Such flexible boundaries could be obtained by georeferencing a "virtual photo" whose tags are the name of the region, and the city and country in which it occurs. In fact, similar ideas have already been proposed, but without making the links with possibility theory explicit. For instance, [43] proposes a method in which spatial terms occurring on a web page

Table 6.4: Accuracy of the predictions at each of the five considered granularity levels.

	Accuracy				
	50	250	500	1000	2000
Probability – Raw	82.08	67.43	61.90	57.46	51.14
Probability – Calibrated	82.65	68.14	62.56	58.02	51.97
Belief – Dempster	84.30	72.38	67.95	63.29	**53.28**
Plausibility – Dempster	84.30	72.41	67.91	62.90	52.66
Pign. Prob. – Dempster	**84.33**	**72.44**	**68.20**	**63.41**	53.27
Belief – Yager	84.30	72.38	67.95	63.29	**53.28**
Plausibility – Yager	84.30	72.41	67.91	62.90	52.66
Pign. Prob. – Yager	82.62	71.50	67.54	63.25	53.27
Belief – Dubois-Prade	84.17	71.89	67.52	62.97	53.11
Plausibility – Dubois-Prade	84.15	72.03	67.44	62.38	52.05
Pign. Prob. – Dubois-Prade	83.67	71.17	66.87	62.29	52.90

are converted into polygons and the overall relevance of that web page w.r.t. a given location is calculated based on the number of polygons in which that location appears. However, seeing these polygons as the focal elements of a mass assignment, this corresponds exactly to determining the degree of plausibility of the considered location.

6.6 Evaluation

As the baseline of our experiments, we will consider the raw probabilities that are produced by the language models (i.e. the right-hand side of (6.1)). This baseline technique has been the basis of a system with which we participated in the 2010 and 2011 editions of the MediaEval Placing Task competition, where it was shown to compare favorably against other georeferencing techniques [16, 44]. This result confirms and strengthens earlier support for using language models in this task [14].

The techniques that we propose in this paper aim at improving the baseline in two different ways. First, by combining evidence from different granularity levels, we can hope that better informed decisions can be made about which is the most likely area at a given granularity level (as was illustrated in Example 1). This means that the Dempster-Shafer based techniques should allow us to obtain a higher overall accuracy. Second, by calibrating the probabilities and by combining evidence from different granularity levels, we can also hope that more reliable confidence estimates are obtained. Here, we are not interested in improving the overall accuracy, but in determining which of the photos we can georeference in an accurate way. This is important from an application point of view, as clearly

Table 6.5: Percentage of photos for which the found location was within 1km, 5km, 10km, 50km, 100km and 1000 km of the true location, and the median distance on the error (in kilometers), when using the raw probabilities (full test set).

Gran.	1	5	10	50	100	1000	10000	Median
50	00.15	00.54	00.89	03.13	05.49	62.50	97.37	732.80
250	01.10	06.48	09.53	20.69	31.03	78.02	96.76	188.97
500	02.39	11.34	16.54	33.58	47.66	76.97	96.49	110.46
1000	04.60	17.69	24.04	47.39	56.90	76.52	96.41	59.34
2000	09.91	25.35	32.98	52.47	59.56	76.28	96.30	40.61

Table 6.6: Percentage of photos for which the found location was within 1km, 5km, 10km, 50km, 100km and 1000 km of the true location, and the median distance on the error (in kilometers), when using pignistic probabilities obtained from Dempster's combination rule (full test set).

Gran.	1	5	10	50	100	1000	10000	Median
50	00.15	00.54	00.93	03.20	05.59	63.27	97.57	728.47
250	01.18	06.84	10.04	21.86	32.65	80.89	97.43	174.62
500	02.56	12.22	17.84	36.29	51.49	80.91	97.39	94.70
1000	04.91	18.90	25.77	51.68	61.66	80.66	97.20	45.81
2000	09.95	25.43	33.16	53.65	61.46	79.49	96.70	37.62

not all photos have sufficiently descriptive tags to allow meaningful coordinates to be found. What we need then, is a way of selecting a maximal set of photos such that at least, say 95% of these photos is correctly georeferenced. Both goals are more or less independent, in the sense that techniques which succeed in improving the overall accuracy may not necessarily be best suited to determine photos that are likely to be georeferenced correctly. In the following, we analyze both goals.

6.6.1 Overall accuracy

Considering the first goal, Table 6.4 summarizes the overall accuracies that are obtained at each of the 5 considered granularity levels, for each of the considered methods. The line *Probability - Raw* contains the results that are obtained when using the raw probabilities provided by the language models, and the line *Probability – Calibrated* contains the results of using the PAV algorithm to calibrate these probabilities as explained in Section 6.3.2. As can be seen from the table, calibration leads to a minor (but consistent) improvement in accuracy. This is somewhat surprising, as the aim of calibration was not to obtain better predictions but to obtain better confidence scores (in relation to the second goal). It should be emphasized here that we used a separate set for calibrating the probabilities, which did neither overlap with the test set nor with the training set that was used

for training the language models. As such, in applying the PAV algorithm, we may implicitly take the observation into account that the probabilities for some areas are systematically too large or too small, and thus influence which area is considered to be the most plausible one for a given photo.

Nonetheless, the improvement in accuracy that is witnessed by applying the PAV algorithm is rather small. One of the main reasons for applying this technique was that accurate probability estimates were needed by the Dempster-Shafer method, to compare the probabilities from language models at different granularity levels. Table 6.4 shows the results that were obtained using three different combination rules (Dempster (6.8)–(6.9), Yager (6.12)–(6.14), and Dubois-Prade (6.15)), each time considering three different decision rules (based on belief (6.16), plausibility (6.17) and pignistic probability (6.18)). For each of these 9 configurations, a clear improvement is found over the results of the (calibrated) language model probabilities. The difference is most pronounced at the intermediate granularity levels. It appears that the language models' results for the coarsest granularity level are difficult to improve, as (i) most of the incorrectly georeferenced photos are simply not tagged in a sufficiently descriptive way (i.e. the language model probabilities are nearly optimal), and (ii) there is little evidence to be found at the finer granularity levels to help make a decision at the coarsest level. Note that, at the coarsest level, there are only 50 clusters for the entire world, hence classification here basically amounts to finding the right country for a photo. Conversely, the results for the finest granularity level are also difficult to improve, which may be due to the same two reasons. While many photos contain tags that allow us to pinpoint the right city, finer predictions can often not be made. Moreover, evidence from the coarser granularity levels is usually not sufficiently specific to help make this decision. For the three intermediate granularity levels, larger improvements are obtained.

Comparing the three combination rules in Table 6.4, we notice that Dempster and Yager produce identical results when either belief or plausibility is used as the decision rule. This was to be expected, since Dempster's and Yager's rules only differ in how the mass of the empty set is redistributed. As a result, the ranking of areas according to their degree of belief or degree of plausibility is unaltered. When using pignistic probability, however, some changes may occur. Similarly, when using Dubois and Prade's combination rule, additional focal elements are introduced, which may affect which area is considered to be the most plausible one at a given granularity level. While Dubois and Prade's rule leads to similar results as Dempster's and Yager's, results of the latter combination rules are slightly better. Concerning the decision rule, pignistic probability was found to be slightly better when using Dempster's rule, while belief was slightly better in combination with Dubois and Prade's rule. In most cases, using belief was also the best choice in combination with Yager's rule.

Table 6.7: Percentage of photos for which the found location was within the correct city, administrative region and country, when using the raw probabilities. (restricted test set)

Granularity	City	Admin	Country
50	01.29	14.09	48.16
250	12.36	39.44	76.75
500	21.73	52.83	81.93
1000	27.38	59.48	82.45
2000	32.36	63.97	81.41

Table 6.8: Percentage of photos for which the found location was within the correct city, administrative region and country, when using pignistic probabilities obtained from Dempster's combination rule (restricted test set).

Granularity	City	Admin	Country
50	01.32	14.34	48.64
250	13.04	41.15	77.33
500	23.48	56.49	84.88
1000	29.26	63.77	86.15
2000	32.61	65.85	86.01

Tables 6.5 and 6.6 provide an overview of these results in terms of error distance between the estimated location for a photo and its true location. These tables confirm the main conclusion from Table 6.4: the use of Dempster-Shafer theory leads to a moderate, but consistent improvement over the baseline, with larger gains to be found at the intermediate levels. Tables 6.7 and 6.8 provide an overview of the results for the same two methods in terms of accuracy at a city level, local administrative unit (LAU) level and country level. The ground truth information for this evaluation was obtained by feeding the real coordinates to the Google Geocoding API [45]. The "Admin" category in the tables corresponds to the *administrative_area_level_1* information provided by Google (i.e. the first-level administrative divisions in a country, such as provinces or states). As the administrative information could not be determined for several photos in the test set and the medoids of several clusters, for the evaluation in Tables 6.7 and 6.8, we have excluded all photos from the test set for which we could not determine the relevant information, as well as all photos which were assigned to a cluster, by any of the methods at any of the granularity levels, with a medoid for which we could not determine the relevant information. This has led to a reduced test set of 32 748 test items (65.49 % of the original test set). The results in Table 6.7 and 6.8 are thus mainly meaningful relative to each other.

To gain a better insight into why the use of Dempster-Shafer theory leads to improved results, we discuss two concrete examples of photos in the test set, where it was needed to look at evidence from other granularity levels to find the correct

Table 6.9: Example assignments of test photos by using Probability – Raw *and* Pign. Prob. – Dempster.

animal zoo wildlife straw colchester mandrill forage foraging

	True location	Estimated location (50 areas)
Probability – Raw	51.8619 0.8267	40.9441 78.9678
Pign. Prob. – Dempster	51.8619 0.8267	51.2189 4.4012

sandals korea toji

	True location	Estimated location (2000 areas)
Probability – Raw	35.1293 127.7567	30.0665 -51.2359
Pign. Prob. – Dempster	35.1293 127.7567	35.2601 128.7594

location. Consider the upper example in Table 6.9. All the tags mentioned in the example were retained at the coarsest granularity level (50 areas). Using the raw probability, this photo was estimated to be in a cluster that represents the North-East of the US, whereas using the pignistic probability correctly assigned it to a cluster in Western Europe. To find the location of this photos, mainly the tags *Colchester* and *zoo* are important, as they clearly suggest that the photo was taken in *Colchester zoo* in the UK. However, at the coarse granularity level of 50 areas, the tag *zoo* will have very little discriminative power, as most of the 50 clusters will contain the location of several zoos. The term *Colchester*, however, will help to find the right cluster, although it leads to an ambiguity: the area containing the UK will definitely contain several occurrences of this tag, but this is also true for the cluster containing the North-East of the US (which contains places called Colchester in VT, CT, NY and IL). Without any further help to make the decision, the baseline system incorrectly assigned it the photo to the US. When looking at the granularity level of 2000 levels, on the other hand, the location becomes obvious: there is only one cluster which a substantial number of occurrences of both *zoo* and *Colchester* (none of the places called Colchester in the North-East of the US has a zoo). The Dempster-Shafer based methods are able to use this evidence from the 2000 area level to find the correct cluster at the 50 area level.

The lower example in Table 6.9 is an illustration of the opposite case, where coarser levels can help us to correctly assign a photo to a cluster at the finer-grained levels. The example concerns a photo taken in South-Korea, which was mistakenly estimated to be in southern Brazil by the baseline, despite the occurrence of the toponym tag *korea* and the apparent lack of ambiguity. After inspecting the training data, we found that the error was due to one cluster (at the 2000 area level) in Brazil with a disproportionate number of occurrences of the tag *toji*, caused by

Table 6.10: Percentage of photos that can be classified at each level of granularity when a fixed accuracy level is imposed (using the probabilities from the language models).

	Acc. (%)	Percentage of photos				
		50	250	500	1000	2000
Probability – Raw	75	100	94	78	72	62
	80	100	78	70	64	52
	85	94	72	62	56	42
	90	84	62	52	44	30
	95	72	46	34	28	14
Probability – Calibrated	75	100	88	78	72	62
	80	100	80	72	66	54
	85	94	72	64	58	46
	90	84	62	54	48	36
	95	74	44	36	32	24

a large number of photos of one user's cat named toji. This tag turned out to be more discriminative than the term *korea* (which occurs in several clusters within Korea), leading to an incorrect decision. At the coarser levels, however, the tag *korea* becomes very discriminative while the tag *toji* loses its importance. In this way, the Dempster-Shafer based methods can use the evidence from the coarser levels to avoid making the mistake at the finest level.

6.6.2 Confidence score reliability

We now turn to the second goal of trying to identify those photos for which the predicted area is most likely to be correct. Being able to identify the "easy" cases from the "hard" cases assists an application in determining the action to be taken: if the application has high confidence in its estimation, it will georeference the photo at hand. Else, when confidence is low, the application does not suggest the location of the photo. Preferably, we want to have a system that is highly accurate in recognizing the "easy" cases. Another way of viewing this task is that we should determine for each photo individually, at which granularity level it is best classified (cfr. the use cases that were outlined in Sections 6.5.2 and 6.5.3). To illustrate this idea, consider the following examples: In the case of a photo tagged with *water wales boats bay cardiff cardiffbay barrage*, the tags unambiguously identify a specific location at a fine granularity level, hence the system should be quite confident in georeferencing such a photo. Secondly, a photo tagged with *france* will not yield a likely locations at the finest granularity level, but at a coarser level of granularity (say, a level at the scale of the European countries), it should become very confident that the photo was taken in the area covering France. Lastly, a photo tagged only with *birthday abby* clearly is a hard case, which is impossible to georeference

Table 6.11: Percentage of photos that can be classified at each level of granularity when a fixed accuracy level is imposed (using Dempster's rule of combination to combine evidence from different granularity levels).

	Acc. (%)	Percentage of photos				
		50	250	500	1000	2000
Plausbility – Dempster	75	100	96	88	80	60
	80	100	88	80	74	52
	85	98	80	74	66	44
	90	88	72	66	58	34
	95	78	62	56	46	0
Belief – Dempster	75	100	96	88	82	66
	80	100	88	82	76	56
	85	98	80	74	68	48
	90	88	72	66	60	36
	95	78	62	56	46	22
Pign. Prob. – Dempster	75	100	96	90	82	66
	80	100	88	82	76	58
	85	98	80	74	68	48
	90	88	72	66	60	36
	95	78	62	56	46	22

even for a human assessor. To determine about which photos' predictions we are confident enough, we can put some threshold on the considered confidence scores. These confidence scores may be probabilities (raw or calibrated), degrees of belief, degrees of plausibility, and pignistic probabilities. In the last three cases, the confidence scores may be evaluated w.r.t. the combined mass assignments resulting from either of the three considered combination rules. The choice of the threshold value allows us to tune the trade-off between having a higher accuracy and having more photos georeferenced.

To assess which method provides the most useful confidence scores, in Tables 6.10–6.13 we show how many photos can be georeferenced when a given level of accuracy is imposed. Comparing the performance of the raw and calibrated probabilities in Table 6.10, we can see that the calibrated probabilities perform consistently better, with the improvement being largest for the finest granularity levels and the highest accuracy thresholds. For instance, at the finest granularity level (2000 clusters), 24% of the photos can be georeferenced with 95% accuracy using the calibrated probabilities, as opposed to only 14% when using the raw probabilities. This means that e.g. if we allow the pignistic probability method to choose 24% of the photos, which it thinks are easiest to georeference, and only require it to georeference these 24%, it will assign a correct cluster to 95% of them. To interpret the meaning of these results, consider an application which suggests

Table 6.12: Percentage of photos that can be classified at each level of granularity when a fixed accuracy level is imposed (using Yager's rule of combination to combine evidence from different granularity levels).

	Acc. (%)	Percentage of photos				
		50	250	500	1000	2000
Plausbility – Yager	75	100	96	88	80	56
	80	100	88	80	72	48
	85	94	80	74	64	42
	90	84	72	64	56	34
	95	74	62	54	46	0
Belief – Yager	75	100	94	88	82	64
	80	100	86	80	74	56
	85	96	78	74	68	46
	90	84	70	64	58	36
	95	64	58	54	46	24
Pign. Prob. – Yager	75	100	94	86	82	64
	80	100	86	80	74	56
	85	94	78	72	66	46
	90	84	70	64	58	36
	95	72	58	54	46	24

a location to users uploading and tagging photos on Flickr, as a way to encourage more people to reveal location-based information, e.g. by showing a map of where the system think the photo was taken. As users will be annoyed if the system is wrong too often, we may need to get it right in, say, 95% in the cases. As this is not possible, by any method, in general (due to there being too many photos with tags that do not reveal any location at all), we can only accomplish this by only making a suggestion to the user when we are confident enough that it is correct. So, given the results in Table 6.10, and assuming that we want 95% of the suggestions we make to be correct, we can only suggest a location in 14% of the cases when using raw probabilities, while we can do it in 24% of the cases using calibrated probabilities. Note that this improvement is mainly due to the better capabilities of the latter method of distinguishing easy cases from hard cases, rather than being (much) better at the actual task of georeferencing.

In Table 6.11, the results of using Dempster's combination rule are presented. A marked improvement over the results from Table 6.10 can be seen, which is largest at the intermediary granularity levels and the higher accuracy thresholds. For instance, at the third granularity level (500 clusters), using Dempster's combination rule and the pignistic probability decision rule, 56% of the photos can be georeferenced with 95% accuracy, as opposed to only 36% for the calibrated probabilities and 34% for the raw probabilities. The best results are found when

Table 6.13: Percentage of photos that can be classified at each level of granularity when a fixed accuracy level is imposed (using Dubois and Prade's rule of combination to combine evidence from different granularity levels).

	Acc. (%)	Percentage of photos				
		50	250	500	1000	2000
Plausbility – Dubois-Prade	75	100	94	88	80	58
	80	100	86	80	72	50
	85	98	80	72	66	42
	90	88	72	64	56	34
	95	78	62	54	46	0
Belief – Dubois-Prade	75	100	94	88	82	64
	80	100	86	80	74	56
	85	98	80	74	68	48
	90	88	72	66	58	36
	95	78	60	54	46	22
Pign. Prob. – Dubois-Prade	75	100	94	86	80	64
	80	100	86	80	74	56
	85	96	78	72	66	48
	90	88	70	64	56	36
	95	76	58	52	44	22

using pignistic probabilities, although the results for degrees of belief are almost identical and the results for degrees of plausibility are similar in most of the cases. Tables 6.12 and 6.13 show the results for respectively Yager's rule and Dubois and Prade's rule. Overall, we may conclude that Dempster's rule provides the best results, followed by Dubois and Prade's rule, and then Yager's rule.

A graphical view on the relation between the number of photos that can be georeferenced and the resulting level of accuracy is provided in Figures 6.4–6.13. These figures provide a clear view of the trade-off in applications between georeferencing a larger percentage of all photos and maintaining a higher accuracy. All the photos in the test set are ranked according to their confidence score (i.e. pignistic probability, belief, or plausibility). As mentioned in the introduction of Section 6.6.2, all the photos whose confidence scores are above a certain threshold would be considered as the "easy" cases. Specifically, for each number of photos n on the X-axis, the accuracy of the n photos with the highest values for this confidence score is reported. First, Figures 6.4–6.8 compare the performance of the three combination rules (using pignistic probabilities), each time also displaying the results for raw and calibrated probabilities. What is particularly noticeable is that the use of calibrated probabilites does not improve the raw probabilities at all for the coarser granularity levels, while at the finest granularity level (Figure 6.8), the calibrated probabilities are essentially as good as the outcome of the

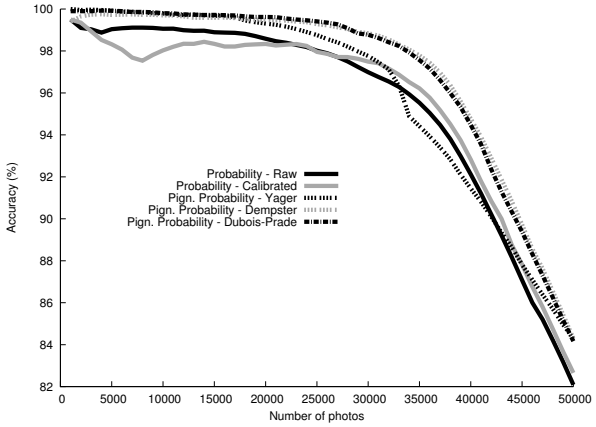


Figure 6.4: Comparing the trade-off between number of georeferenced photos and accuracy for different combination rules, using pignistic probability and 50 clusters.

Dempster-Shafer based approaches. Overall, we can also see that the combination operator being used does not affect the performance in a crucial way. Figures 6.9–6.13 compare the performance of the three decision rules (using Dempster's rule of combination). Here we can clearly see that using degrees of belief or using pignistic probabilities does not substantially change the result. Regarding degrees of plausibility, the results are somewhat mixed. At the finer granularity levels and the left-most part of the graphs, plausibility degrees perform even worse than the baseline. In some sense, this is not surprising, as the idea of plausibility degrees is somewhat at odds with the task of finding those photos for which sufficient location evidence can be found. Indeed, plausibility degrees reflect the compatibility of a given element with available evidence, rather than an amount of support.

6.7 Related work

6.7.1 Finding locations of resources

The task of deriving geographic coordinates for photos has recently gained in popularity (see e.g. [16]). However, to the best of our knowledge, the idea of combining evidence from different granularity levels and the related problem of finding the most appropriate granularity level for a given photo have not been previously considered. In the context of geographic information systems, on the other hand, it is well known that different *scales* may yield different effects on the spatial and thematic resolution of geographic data [12] (e.g. monitoring the earth's surface using satellites with different resolutions).

Most existing approaches are based on clustering, in one way or another, to

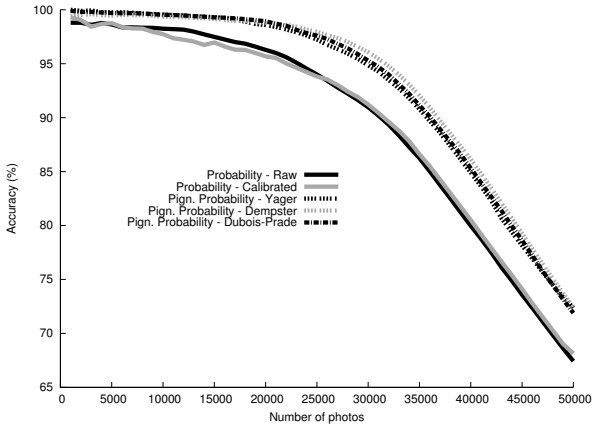


Figure 6.5: Comparing the trade-off between number of georeferenced photos and accuracy for different combination rules, using pignistic probability and 250 clusters.

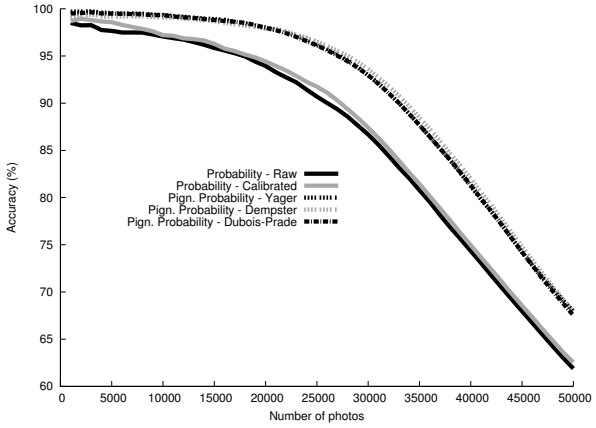


Figure 6.6: Comparing the trade-off between number of georeferenced photos and accuracy for different combination rules, using pignistic probability and 500 clusters.

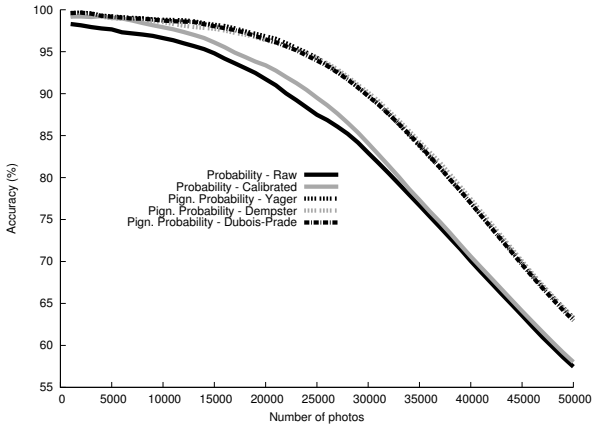


Figure 6.7: Comparing the trade-off between number of georeferenced photos and accuracy for different combination rules, using pignistic probability and 1000 clusters.

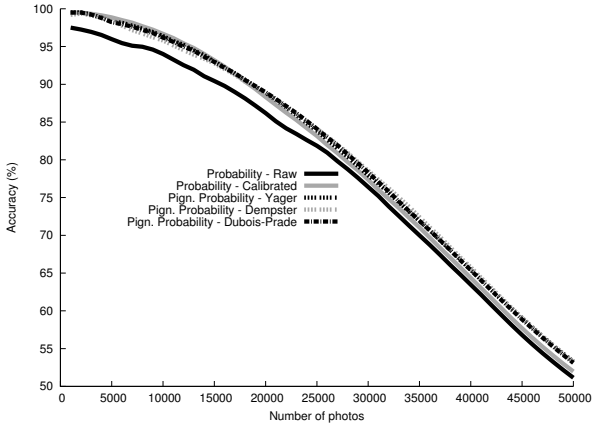


Figure 6.8: Comparing the trade-off between number of georeferenced photos and accuracy for different combination rules, using pignistic probability and 2000 clusters.

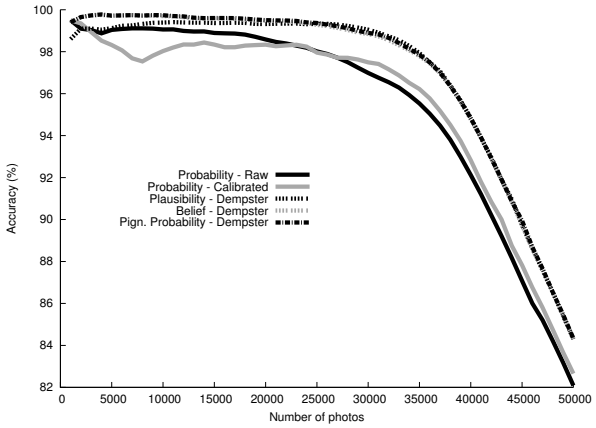


Figure 6.9: Comparing the trade-off between number of georeferenced photos and accuracy for different decision rules, using Dempster's combination rule and 50 clusters.

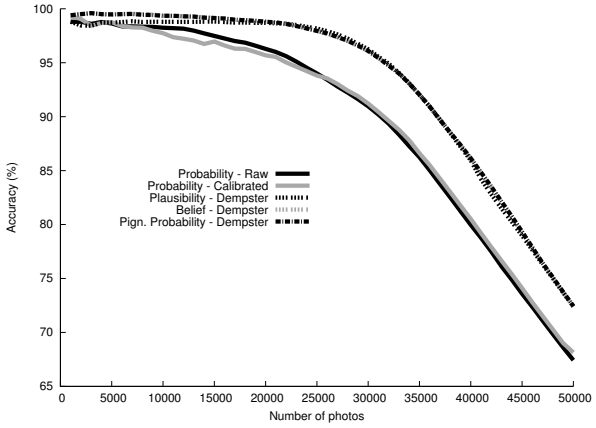


Figure 6.10: Comparing the trade-off between number of georeferenced photos and accuracy for different decision rules, using Dempster's combination rule and 250 clusters.

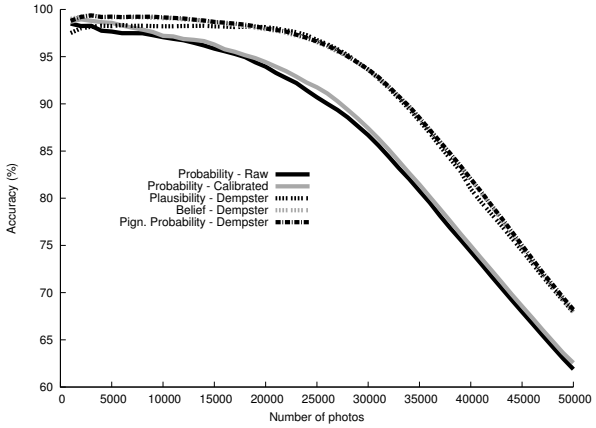


Figure 6.11: Comparing the trade-off between number of georeferenced photos and accuracy for different decision rules, using Dempster's combination rule and 500 clusters.

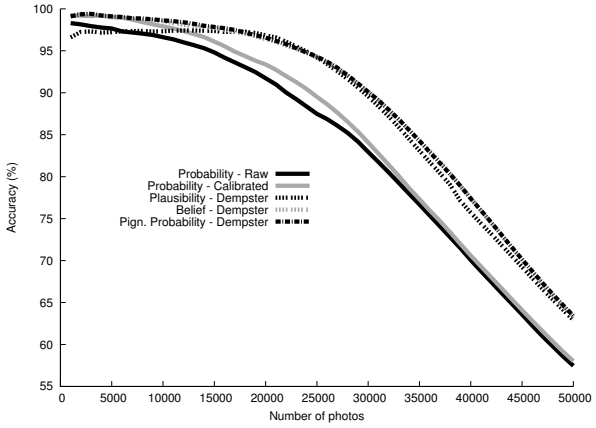


Figure 6.12: Comparing the trade-off between number of georeferenced photos and accuracy for different decision rules, using Dempster's combination rule and 1000 clusters.

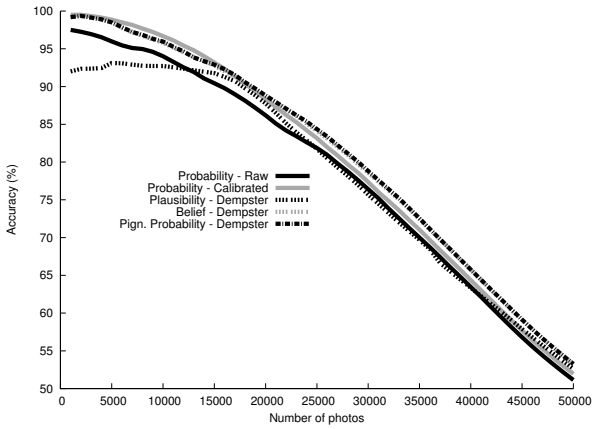


Figure 6.13: Comparing the trade-off between number of georeferenced photos and accuracy for different decision rules, using Dempster's combination rule and 2000 clusters.

convert the task into a classification problem. For instance, in [46] target locations are determined using mean shift clustering, a non-parametric clustering technique from the field of image segmentation. The advantage of this clustering method is that an optimal number of clusters is determined automatically, requiring only an estimate of the scale of interest. Specifically, to find good locations, the difference is calculated between the density of photos at a given location and a weighted mean of the densities in the area surrounding that location. To assign locations to new images, both visual (keypoints) and textual (tags) features were used. Experiments were carried out on a sample of over 30 million images, using both Bayesian classifiers and linear support vector machines, with slightly better results for the latter. Two different resolutions were considered corresponding to approximately 100 km (finding the correct metropolitan area) and 100 m (finding the correct landmark). It was found that visual features, when combined with textual features, substantially improve accuracy in the case of landmarks. In [47], an approach is presented which is based purely on visual features. For each new photo, the 120 most similar photos with known coordinates are determined. This weighted set of 120 locations is then interpreted as an estimate of a probability distribution, whose mode is determined using mean-shift clustering. The resulting value is used as prediction of the image's location.

The idea that when georeferencing images, the spatial distribution of the classes (areas) could be utilized to improve accuracy has already been suggested in [14]. Their starting point is that typically not only the correct area will receive a high probability, but also the areas surrounding the correct area. Indeed, the expected distribution of tags in these areas will typically be quite similar. Hence, if some

area a receives a high score, and all of the areas surrounding a also receive a relatively high score, we can be more confident in a being approximately correct than when all the areas surrounding a receive a low score. Motivated by this intuition, [14] proposes to smooth $P(a|x)$ as follows (using a uniform prior):

$$P^*(a|x) \propto \alpha P(x|a) + (1-\alpha) \cdot \sum_{b \in neigh_d(a)} \frac{P(x|b)}{(2d+1)^2 - 1}$$

where $d > 0$ and $neigh_d(a)$ is the set of all areas that are within distance d of a.

Some Flickr tags are intuitively more important than others in determining the location of a photo. Toponyms in particular are by definition indicative of geographic location. One way of recognizing toponyms is by looking for so-called comma-groups. These are groups of words that are comma-separated, e.g *San Francisco, California, USA*. In this example, there is a clear relationship between the comma-separated values, as San Francisco is a city, located in the state of California, which is in turn one of the states of the USA. As a result, resolution of the toponyms represented by this group reveals an unambiguous geographical reference. Resolution of such comma-groups has been studied by Lieberman in [48].

In addition to georeferencing Flickr photos, several authors have recently focused on finding the location of other web resources such as Twitter posts or Wikipedia pages. For instance, in [49], a probabilistic framework based on maximum likelihood estimation was used to estimate the location of users based on the content of their tweets. In particular, a generative probabilistic model proposed in [50] is used to determine words with a geographic scope within a tweet, and a form of neighborhood smoothing is employed to refine the estimations. For 51% of the users, a location was obtained that is within a 100 mile radius of their true location. Next, [51] looked into georeferencing Wikipedia articles as well as Twitter posts. After laying out a grid over the earths surface (in a way similar to [1]), for each grid cell a generative language model is estimated. To assign a test item to a grid cell, its Kullback-Leibler divergence with the language models of each of the cells is calculated. In [52], it was shown how Wikipedia pages can be georeferenced using language models that are trained from Flickr, taking the view that the relative sparsity of georeferenced Wikipedia pages does not allow for sufficiently accurate language models to be trained, especially at finer levels of granularity.

Interestingly, some recent language modeling approaches have combined the idea of topic models with location-dependent language models. For instance, [54] proposes geographic topic models with the aim of simultaneously capturing linguistic variation across different regions and different topics.

6.7.2 Using locations of resources

When available, the coordinates of a photo may be used in various ways. In [55], for instance, coordinates of tagged photos are used to find representative textual descriptions of different areas of the world. These descriptions are then put on a map to assist users in finding images that were taken in a given location of interest. Their approach is based on spatially clustering a set of geotagged Flickr images, using k-means, and then relying on (an adaptation of) tf-idf weighting to find the most prominent tags of a given area. Similarly, [56] looks at the problem of suggesting useful tags, based on available coordinates. The relevance of a given tag is measured in terms of the number of users that have used it to describe photos located within a certain radius of the current photo's coordinates. A refinement of this method only looks at tags that occur with visually similar photos, which is shown to improve the quality of the proposed tags. Some authors have looked at using geographic information to help diversify image retrieval results [57, 58]. Finally, in [53] GeoSR is presented as a way of measuring the semantic relatedness of Wikipedia articles based on their geographic context, allowing users to explore information in Wikipedia that is relevant to a particular location.

Geotagged photos are also useful from a geographic perspective, to better understand how people refer to places, and overcome the limitations and/or costs of existing mapping techniques [59]. For instance, by analyzing the tags of georeferenced photos, Hollenstein [60] found that the city toponym was by far the most essential reference type for specific locations. Moreover, [60] provides evidence suggesting that the average user has a rather distinct idea of specific places, their location and extent. Despite this tagging behaviour, Hollenstein concluded that the data available in the Flickr database meets the requirements to generate spatial footprints at a sub-city level. Finding such footprints for non-administrative regions (i.e. regions without officially defined boundaries) using georeferenced resources has also been adressed in [9] and [6]. Another problem of interest is the automated discovery of which names (or tags) correspond to places. Especially for vernacular place names, which typically do not appear in gazetteers, collaborative tagging-based systems may be a rich source of information. In [61], methods based on burst-analysis are proposed for extracting place names from Flickr. Finally, note that to some extent, even without geographic coordinates, ontologies, and in particular ontologies of places may be derived from Flickr tags [62].

6.7.3 Evidence theory

Various authors have investigated the use of Dempster-Shafer theory for combining the results of different classifiers [63–66]. However, the aim of using Dempster-Shafer theory in this context is quite different from our aim in this paper. Specifically, these methods mainly use Dempster-Shafer theory for its ability to

represent partial ignorance. For instance, if a given classifier assigns a probability p_i to each class c_i, a belief function may be constructed by choosing $m(\{c_i\}) = f_i$ for some $f_i < p_i$, and $m(C) = 1 - \sum_i f_i$, for $C = \{c_1, ..., c_n\}$ the set of all classes. The value $1 - \sum_i f_i$ can then intuitively be interpreted in terms of confidence in the associated classifier. Note also that all focal elements are then either singletons or the universe, which makes Dempster-Shafer theory sufficiently scalable to deal with large numbers of classes, although sometimes focal elements of the form $C \setminus \{c_i\}$ are also used.

Dempster-Shafer theory has also been widely considered for dealing with the imperfection of real-world geographic information; [67] provides a survey on works using Dempster-Shafer theory in a GIS setting. More generally, we refer to [68] for an overview of different frameworks for handling uncertainty, applied to spatial information.

6.8 Conclusions

We have proposed an approach to georeferencing Flickr photos which combines the evidence provided by different language models using Dempster-Shafer evidence theory. As these language models were trained at different granularity levels, they provide complementary views on the georeferencing process, and implicitly add a spatial dimension to the language models.

The core idea of our approach is to see a probability distribution over coarse areas as a probability distribution over sets of fine-grained areas. Noting that this latter probability distribution corresponds to the notion of a mass assignment from Dempster-Shafer theory, we can connect to the vast amount of work that has already been done on combining evidence. In particular, we have experimented with three well-known combination rules, due to Dempster, Yager, and Dubois and Prade respectively.

After the evidence from the language models has been combined, we end up with a mass assignment that summarizes all available evidence about the location of a given photo. This mass assignment may then be used in different ways: we may use it to select the most likely area at a given granularity level, we may determine the smallest area that contains the true location of the photo with a predefined certainty, or we may simply visualize the evidence after approximating the mass assignment to a possibility distribution. In our experiments, we have focused on the first two of these tasks, as the quality of visual representations is difficult to quantify. In both cases, we have found that our evidence-based approach considerably improves the performance of individual language models.

References

[1] C. B. Jones, A. I. Abdelmoty, D. Finch, G. Fu, S. Vaid, The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing, in: Proceedings of the Third International Conference on Geographic Information Science, 2004, pp. 125–139.

[2] O. Van Laere, S. Schockaert, B. Dhoedt, Towards automated georeferencing of flickr photos, in: Proceedings of the 6th Workshop on Geographic Information Retrieval, 2010, pp. 5:1–5:7.

[3] L. Hollenstein, R. Purves, Exploring place through user-generated content: Using Flickr to describe city cores, Journal of Spatial Information Science 1 (1) (2010) 21–48.

[4] A. Popescu, G. Grefenstette, H. Bouamor, Mining a multilingual geographical gazetteer from the web, in: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009, pp. 58–65.

[5] C. Keßler, P. Maué, J. Heuer, T. Bartoschek, Bottom-up gazetteers: Learning from the implicit semantics of geotags, in: Proceedings of the 3rd International Conference on Geospatial Semantics, 2009, pp. 83–102.

[6] F. Wilske, Approximation of neighborhood boundaries using collaborative tagging systems, in: Proceedings of the GI-Days, 2008, pp. 179–187.

[7] F. A. Twaroch, C. B. Jones, A. I. Abdelmoty, Acquisition of a vernacular gazetteer from web sources, in: Proceedings of the First International Workshop on Location and the Web, 2008, pp. 61–64.

[8] I. Holt, J. Green, Social networks as a future geographical data source, in: Proceedings of the W3C Workshop on the Future of Social Networking, 2009.

[9] S. Schockaert, M. De Cock, Neighborhood restrictions in geographic IR, in: Proceedings of the 30th Annual International ACM SIGIR Conference, 2007, pp. 167–174.

[10] C. B. Jones, R. S. Purves, P. D. Clough, H. Joho, Modelling vague places with knowledge from the web, International Journal of Geographical Information Science 22 (2008) 1045–1065.

[11] M. F. Goodchild, M. J. Egenhofer, K. K. Kemp, D. M. Mark, E. Sheppard, Introduction to the Varenius project, International Journal of Geographical Information Science 13 (8) (1999) 731–745.

[12] M. F. Goodchild, A geographer looks at spatial information theory, in: Proceedings of the International Conference on Spatial Information Theory, Springer-Verlag, 2001, pp. 1–13.

[13] P. A. Longley, M. F. Goodchild, D. J. Maguire, D. W. Rhind, Geographic Information Systems and Science, John Wiley & Sons, 2005.

[14] P. Serdyukov, V. Murdock, R. van Zwol, Placing Flickr photos on a map, in: Proceedings of the 32nd Annual International ACM SIGIR Conference, 2009, pp. 484–491.

[15] M. L. et al., Automatic tagging and geotagging in video collections and communities, in: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, 2011,, pp. 51:1–51:8.

[16] M. Larson, M. Soleymani, P. Serdyukov, V. Murdock, G. Jones (Eds.), Working Notes of the MediaEval Workshop, 2010.

[17] T. Gruber, Collective knowledge systems: Where the social web meets the semantic web, Journal of Web Semantics 6 (1) (2008) 4 – 13.

[18] C. Becker, C. Bizer, Exploring the geospatial semantic web with DBpedia Mobile, Journal of Web Semantics 7 (4) (2009) 278 – 286.

[19] O. Van Laere, S. Schockaert, B. Dhoedt, Ghent university at the 2010 Placing Task, in: Working Notes of the MediaEval Workshop, 2010.

[20] A. Dempster, A Generalization of Bayesian Inference, Journal of the Royal Statistical Society. Series B (Methodological) 30 (2) (1968) 205–247.

[21] G. Shafer, A mathematical theory of evidence, Princeton University Press, 1976.

[22] O. Van Laere, S. Schockaert, B. Dhoedt, Combining multi-resolution evidence for georeferencing Flickr images, in: Proceedings of the 4th International Conference on Scalable Uncertainty Management, 2010, pp. 347–360.

[23] O. Van Laere, S. Schockaert, B. Dhoedt, Finding locations of flickr resources using language models and similarity search, in: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, 2011, pp. 48:1–48:8.

[24] J. Ponte, W. Croft, A language modeling approach to information retrieval, in: Proceedings of the 21st Annual International ACM SIGIR Conference, 1998, pp. 275–281.

[25] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to information retrieval, ACM Transactions on Information Systems 22 (2) (2004) 179–214.

[26] M. D. Smucker, J. Allan, An investigation of Dirichlet prior smoothing's performance advantage, Tech. Rep. IR-445, University of Massachusetts (2005).

[27] P. Bennett, Assessing the calibration of Naive Bayes' posterior estimates, Tech. Rep. CMU-CS00-155, Carnegie Mellon (2000).

[28] B. Zadrozny, C. Elkan, Obtaining calibrated probability estimates from decision trees and NaiveBayesian classifiers, in: Proceedings of the 18th International Conference on Machine Learning, 2001, pp. 609–616.

[29] B. Zadrozny, C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, in: Proceedings of the 8th ACM SIGKDD International Conference, 2002, pp. 694–699.

[30] M. Ayer, H. Brunk, G. Ewing, W. Reid, E. Silverman, An empirical distribution function for sampling with incomplete information, The Annals of Mathematical Statistics 26 (4) (1955) 641–647.

[31] W. Wilbur, L. Yeganova, W. Kim, The synergy between PAV and AdaBoost, Machine Learning 61 (2005) 71–103.

[32] T. Fawcett, A. Niculescu-Mizil, PAV and the ROC convex hull, Machine Learning 68 (2007) 97–106.

[33] P. Smets, Constructing the pignistic probability function in a context of uncertainty, in: Proceedings of the 5th Annual Conference on Uncertainty in Artificial Intelligence, 1990, pp. 29–40.

[34] D. Dubois, H. Prade, On the unicity of Dempster rule of combination, International Journal of Intelligent Systems 1 (2) (1986) 133–142.

[35] F. Klawonn, E. Schwecke, On the axiomatic justification of Dempster's rule of combination, International Journal of Intelligent Systems 7 (5) (1992) 469–478.

[36] L. A. Zadeh, A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination, AI Magazine 7 (2) (1986) 85–90.

[37] P. Smets, R. Kennes, The transferable belief model, Artificial Intelligence 66 (2) (1994) 191 – 234.

[38] R. R. Yager, On the Dempster-Shafer framework and new combination rules, Information Sciences 41 (2) (1987) 93 – 137.

[39] D. Dubois, H. Prade, Representation and combination of uncertainty with belief functions and possibility measures, Computational Intelligence 4 (3) (1988) 244–264.

[40] D. Dubois, H. Prade, Possibility theory: an approach to computerized processing of uncertainty, Plenum Press, 1988.

[41] G. Shackle, Decision, Order and Time in Human Affairs, Cambridge University Press, 1961.

[42] D. Dubois, H. Prade, Fuzzy sets, probability and measurement, European Journal of Operational Research 40 (2) (1989) 135–154.

[43] R. Larson, Geographic information retrieval and spatial browsing, GIS and Libraries: Patrons, Maps and Spatial Information (1996) 81–124.

[44] O. Van Laere, S. Schockaert, B. Dhoedt, Ghent university at the 2011 Placing Task, in: Working Notes of the MediaEval Workshop, 2011.

[45] Google Geocoding API [cited December 6th, 2011].
URL http://code.google.com/apis/maps/documentation/geocoding/

[46] D. J. Crandall, L. Backstrom, D. Huttenlocher, J. Kleinberg, Mapping the world's photos, in: Proceedings of the 18th International Conference on World Wide Web, 2009, pp. 761–770.

[47] J. H. Hays, A. A. Efros, IM2GPS: Estimating geographic information from a single image, in: Proceedings of the 21st IEEE Compuster Society Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[48] M. D. Lieberman, H. Samet, J. Sankaranayananan, Geotagging: using proximity, sibling, and prominence clues to understand comma groups, in: Proceedings of the 6th Workshop on Geographic Information Retrieval, 2010, pp. 6:1–6:8.

[49] Z. Cheng, J. Caverlee, K. Lee, You are where you tweet: a content-based approach to geo-locating twitter users, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010, pp. 759–768.

[50] L. Backstrom, J. Kleinberg, R. Kumar, J. Novak, Spatial variation in search engine queries, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 357–366.

[51] B. Wing, J. Baldridge, Simple supervised document geolocation with geodesic grids, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 955–964.

[52] C. De Rouck, O. Van Laere, S. Schockaert, B. Dhoedt, Georeferencing Wikipedia pages using language models from Flickr, in: Proceedings of the Terra Cognita 2011 Workshop, 2011, pp. 3–10.

[53] B. Hecht, M. Raubal, GeoSR: Geographically explore semantic relations in world knowledge, in: Proceedings of the 11th AGILE International Conference on Geographic Information Science, 2008, pp. 95–114.

[54] J. Eisenstein, B. O'Connor, N. A. Smith, E. P. Xing, A latent variable model for geographic lexical variation, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1277–1287.

[55] S. Ahern, M. Naaman, R. Nair, J. H.-I. Yang, World explorer: visualizing aggregate data from unstructured text in geo-referenced collections, in: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, 2007, pp. 1–10.

[56] E. Moxley, J. Kleban, B. Manjunath, Spirittagger: a geo-aware tag suggestion tool mined from Flickr, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, 2008, pp. 24–30.

[57] L. Kennedy, M. Naaman, Generating diverse and representative image search results for landmarks, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 297–306.

[58] A. Popescu, I. Kanellos, Creating visual summaries for geographic regions, in: IR+SN Workshop (at ECIR), 2009.

[59] M. Goodchild, Citizens as sensors: the world of volunteered geography, GeoJournal 69 (2007) 211–221.

[60] L. Hollenstein, Capturing vernacular geography from georeferenced tags, Master's thesis, University of Zurich (2008).

[61] T. Rattenbury, M. Naaman, Methods for extracting place semantics from Flickr tags, ACM Transactions on the Web 3 (1) (2009) 1–30.

[62] P. Schmitz, Inducing ontology from Flickr tags, in: Proceedings of the Collaborative Web Tagging Workshop, 2006, pp. 210–214.

[63] A. Al-Ani, M. Deriche, A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence, Journal of Artificial Intelligence Research 17 (1) (2002) 333–361.

[64] T. Denœux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, IEEE Transactions on Systems, Man, and Cybernetics 25 (5) (1995) 804–813.

[65] G. Rogova, Combining the results of several neural network classifiers, Neural Networks 7 (5) (1994) 777–781.

[66] L. Xu, C. Suen, Methods of combining multiple classifiers and their applications to handwriting recognition, IEEE Transactions on Systems, Man, and Cybernetics 22 (3) (1992) 418–435.

[67] J. Malpica, M. Alonso, M. Sanz, Dempster-Shafer theory in geographic information systems: a survey, Expert Systems with Applications 32 (1) (2007) 47 – 55.

[68] R. Jeansoulin, O. Papini, H. Prade, S. Schockaert (Eds.), Methods for Handling Imperfect Spatial Information, Studies in Fuzziness and Soft Computing, Springer, 2010.

7

Georeferencing Wikipedia documents using data from social media sources

In this chapter, we evaluate the use of language models trained on Wikipedia, Flickr and Twitter data, individually and in a combined way, for the task of georeferencing Wikipedia documents. In our experimental evaluation, we demonstrate that our language models substantially outperform both classical gazetteer-based methods and language modelling approaches trained on Wikipedia data alone. This supports the hypothesis that social media are an important source of geographic information, which is valuable beyond the scope of individual applications.

⋆ ⋆ ⋆

Olivier Van Laere, Steven Schockaert, Vlad Tanasescu, Bart Dhoedt and Chris Jones

Submitted to ACM Transactions on Information Systems, ACM, November 2012.

Abstract Social media sources such as Flickr and Twitter continuously generate large amounts of textual information (viz. tags on Flickr and short messages on Twitter). This textual information is increasingly linked to geographical coordinates, which makes it possible to learn how people refer to places by identifying

correlations between the occurrence of terms and the locations of the corresponding social media objects. Recent work has focused on how this potentially rich source of geographic information can be used to estimate geographic coordinates for previously unseen Flickr photos or Twitter messages. In this paper, we extend this work by analysing to what extent probabilistic language models trained on Flickr and Twitter can be used to assign coordinates to Wikipedia articles. Our results show that exploiting these language models substantially outperforms both (i) classical gazetteer-based methods (in particular, Yahoo! Placemaker) and (ii) language modelling approaches trained on Wikipedia alone. This supports the hypothesis that social media are important sources of geographic information, which are valuable beyond the scope of individual applications.

7.1 Introduction

Location plays an increasingly important role on the Web. Smartphones enable users around the world to participate in social media activities, such as sharing photos or broadcasting short text messages. In this process, the content that is added by a given user is often annotated with its geographical location (either automatically by a GPS device or manually by the user). As a result, more and more georeferenced content is becoming available on the web. At the same time, due to the popularity of location-based services, the demand for georeferenced content has also become stronger. Applications such as Foursquare[1] or Google Places[2], for instance, allow users to find nearby places of a given type, while applications such as Wikitude[3] provide information about a user's surroundings by using georeferenced Wikipedia articles, among others.

Several authors have investigated how geotagged Flickr photos (i.e. Flickr photos that are associated with coordinates) can be used to estimate coordinates for photos without geotags [1–3]. Although some authors have exploited visual features from the actual pictures, the dominant approach consists of training language models for different geographic areas, and subsequently using these language models to estimate in which area a photo was most likely taken. More recently, similar approaches have been proposed to georeference Twitter messages [4–6] and Wikipedia articles [7, 8]. A key aspect of the aforementioned approaches is that the considered language models are always trained on the type of resources that are georeferenced (e.g. Flickr photos are used to train a system for georeferencing Flickr photos). While this makes sense in the case of Flickr and Twitter, it is less clear whether an approach for georeferencing Wikipedia articles can be truly effective in this way. Indeed, since different users may take photos of

[1] https://foursquare.com/
[2] http://www.google.com/places/
[3] http://www.wikitude.com/

the same places, given a new photo to be georeferenced, it will often be the case that several photos from the same place are contained in the training data. Hence, if we can identify these photos from the training data, accurate coordinates for the new photo can be found. In contrast, given a new Wikipedia article about a given place, there should normally not be any other articles about that place in Wikipedia, implying that at most only approximate coordinates can be inferred (e.g. by discovering in which city the described place is located). On the other hand, there may be georeferenced photos on Flickr of the place, or georeferenced Twitter messages that describe the specific Wikipedia article.

In this paper, our central hypothesis is that a system for georeferencing Wikipedia articles about places can substantially benefit from using sources such as Flickr or Twitter. As the number of georeferenced Flickr photos and Twitter messages is increasing at a fast pace, if confirmed, this hypothesis could form the basis of a powerful new approach for georeferencing Wikipedia articles, and more generally, text documents on the web, continuously improving its performance as more training data becomes available.

The results we present in this paper strongly support our hypothesis. In particular, using a language model trained using 376K Wikipedia documents, we obtain a median error of 4.17 km, while a model trained using 32M Flickr photos yields a median error of 2.5 km. When combining both models, the median error is further reduced to 2.16 km. Repeating the same experiment with 16M tweets as the only training data results in a median error of 35.81 km. When combined with the Wikipedia training data the median error decreases to 3.69 km. Combining all three models results in a median error of 2.18 km, suggesting that while Twitter is useful in the absence of Flickr data, the evidence it provides is superseded by the evidence encoded in the Flickr models.

The remainder of this paper is organized as follows: Section 7.2 summarizes related work in the field of georeferencing textual resources. Section 7.3 describes the different data sources we consider and summarizes the datasets we use in our evaluation. Next, Section 7.4 describes how we estimate and combine language models from Flickr, Twitter and Wikipedia. Our evaluation is discussed in detail in Section 7.5. In Section 7.6 we provide a discussion about our main result. Finally, Section 7.7 states the conclusions and discusses future work.

7.2 Related work

We review two areas of work on georeferencing: gazetteer-based methods in Section 7.2.1, followed by language modelling based methods in Section 7.2.2.

7.2.1 Gazetteer based methods

Gazetteers are essentially lists or indices containing information about a large number of known places, described by features such as geographic coordinates, semantic types, and alternative names. Examples of gazetteers are Yahoo! GeoPlanet[4] and Geonames[5], the latter being freely available for download and containing information about over 8.1 million different entities worldwide. The data contained in a gazetteer is mostly manually selected and reviewed by domain experts and thus tends to be of high quality. However, manual moderation is a time-consuming and cumbersome task. In an effort to address this issue as well as the limited coverage of some gazetteers, [9] proposes a method for automatically constructing a gazetteer from different sources using text mining. [10] produced a gazetteer service that accesses multiple existing gazetteer and other place name resources, using a combination of manual resolution of feature types and automated name matching to detect duplicates. [11] access multiple gazetteers and digital maps in a mediation architecture for a meta-gazetteer service using similarity matching methods to conflate the multiple sources of place data in real-time.

Given access to a comprehensive gazetteer, a natural way to discover the geographic scope of a webpages consists of identifying place names and looking up their coordinates in the gazetteer. In practice, however, this method is complicated by the fact that many place names are highly ambiguous. A well known-example is "Springfield": at least 58 populated places with this name are listed in Geonames. Georeferencing methods using a gazetteer have to cope with this. In [12], gazetteers are used to estimate the locations of toponyms mentioned in text and a geographical focus is determined for each page. During this process, two different types of ambiguities are described: geo/geo, e.g. the previous example of "Springfield", or geo/non-geo, such as "Turkey" or "Bath", which are also common nouns in English. Heuristic strategies to resolve both type of ambiguities are proposed. [13] presents a probabilistic framework that is able to propose additional tags capable of disambiguating the meaning of the tags associated to a Flickr photo. For instance, given the tag "washington", adding "dc" or "seattle" resolves the possible ambiguity. [14] investigated toponym resolution based on the understanding of comma groups, such as the previous example of "Washington, DC", to determine the correct interpretation of the place names. [15] resolve toponyms against a number of gazetteers, and tackle the problem of ambiguity using a number of heuristics based on an in-depth analysis carried out in [16]. In addition to all aforementioned types of ambiguity, place names are sometimes used in a nonspatial sense (e.g. "Brussels" refers to a political entity in a sentence such as "According to Brussels, the proposed measures have been ineffective"). This form of ambiguity can, in principle, be addressed using standard techniques for named

[4] http://developer.yahoo.com/geo/geoplanet/
[5] http://www.geonames.org/

entity recognition (NER), although it is a non-trivial problem.

Another limitation of gazetteer based methods is that people often use vernacular names to describe places, which tend to be missing from gazetteers. For instance, "The Big Apple" is used when referring to "New York City". To cope with this [17] extract knowledge of vernacular names from web sources by exploiting co-occurrence on the web with known georeferenced places.

7.2.2 Language modelling based methods

Over the past few years considerable research has focused on georeferencing Flickr photos on the basis of their tags. The tendency for particular tags to be clustered spatially, and hence to provide strong evidence for the place at which a photo was taken, was studied by [18, 19] who compared alternative clustering techniques and demonstrated the benefits of hybrid approaches. Most existing georeferencing methods exploit the clustering properties in one way or another to convert the georeferencing task to a classification problem. For instance, in [1] locations of unseen resources are determined using the mean shift clustering algorithm, a non-parametric clustering technique from the field of image segmentation. The advantage of this clustering method is that the number of clusters is determined automatically from a scale parameter. To assign locations to new images, both visual (keypoints) and textual (tags) features have been used in [1]. Experiments were carried out on a sample of over 30 million images, using both Bayesian classifiers and linear support vector machines, with slightly better results for the latter. Two different resolutions were considered corresponding to approximately 100 km (finding the correct metropolitan area) and 100 m (finding the correct landmark). The authors found that visual features, when combined with textual features, substantially improve accuracy in the case of landmarks. In [2], the idea is suggested that whenever a classifier suggests a certain area where an image was most likely taken, the surrounding areas could be considered as well to improve the results. Their observation is that typically not only the correct area will receive a high probability, but also surrounding areas will exhibit similar behaviour. This idea was further elaborated on in [20], where the evidence for a certain location from models trained at different levels of granularity is combined using Dempster-Shafer evidence theory to determine the most likely location where a certain photo was taken and to assess the spatial granularity for which this estimation is meaningful. Finally, [3] showed that approaches using classification benefit from a second step, in which a suitable location is determined within the area that was found by the classifier, by assessing the similarity (here the Jaccard measure was used to assess similarity) between the photo to be georeferenced and the photos from the training data that are known to be located in that area. The interest of the research community into this problem resulted in the Placing Task, an evaluation frame-

work focussing on the problem of georeferencing Flickr videos [21], as part of the MedialEval benchmarking initiative[6].

In parallel and using similar techniques, researchers have looked into georeferencing Twitter messages. Due to their limited length, Twitter messages are much harder to georeference than for instance web-pages. For example, when an ambiguous term occurs, it is less likely that the surrounding words will provide sufficient context for accurate disambiguation. However, as tweets are rarely posted in isolation, previous messages from the same user can be exploited as context information. Following such a strategy, [4] show that it is possible to estimate the geographical location of a Twitter user using latent topic models, an approach which was shown to outperform text regression and supervised topic models. [5] propose a method to determine the city in which a Twitter user is located (among a pre-selected set of cities). Each city is modelled through a probabilistic language model, which can be used to estimate the probability that the user's tweets were written by a resident of that city. While this baseline model only found the correct city for 10% of the users, substantial improvements were obtained when using a term selection method to filter all terms that are not location-relevant, leading to a 49.8% accuracy on a city scale. [6] train language models over geotagged Twitter messages, and rely on Kullback-Leibler divergence to compare the models of locations with the models of tweets. The results show that around 65% of the tweets can thus be located within the correct city (among a pre-selected set of 10 cities with high Twitter usage) and around 20% even within the correct neighbourhood (in this case, within the spatial scope of New York only). In comparison, the effectiveness of gazetteer based methods for georeferencing Twitter messages was found to amount to 1.5% correctly georeferenced messages on the neighbourhood scale (in this experiment Yahoo! Placemaker was used).

When it comes to georeferencing Wikipedia documents, the work of [7] is of particular interest. After laying out a grid over the Earth's surface (in a way similar to [2]), for each grid cell a generative language model is estimated using only Wikipedia training data. To assign a test item to a grid cell, its Kullback-Leibler divergence with the language models of each of the cells is calculated. Results are also reported for other approaches, including Naive Bayes classification. The follow-up research in [8] improved this method in two ways. First, an alternative clustering of the training data is suggested: by using k-d trees, the clustering is more robust to data sparsity in certain clusters when using large datasets. Indeed, most of the datasets are not uniformily distributed and using a grid with equal-sized cells will ignore the fact that certain parts of the world can be covered quite densely or sparsely with training data, depending on the location. In this paper, we use k-medoids clustering for a similar purpose. A second improvement is that instead of returning the center of the grid cell, the centre-of-gravity is returned

[6] http://www.multimediaeval.org/

of the locations of the Wikipedia pages from the training data that are located in the cell. The significance of this latter improvement is confirmed by our earlier results in [3], in the setting of georeferencing Flickr photos, and is described in Section 7.4.5.

In this paper, we will investigate the use of mixed data sources to georeference Wikipedia documents. The approaches outlined above indeed all use the same type of information for training and test data. First efforts in this area include our previous work [22] where a preliminary evaluation has been carried out of the effectiveness of georeferencing Wikipedia pages using language models from Flickr, taking the view that the relative sparsity of georeferenced Wikipedia pages does not allow for sufficiently accurate language models to be trained, especially at finer levels of granularity. In addition some evaluations have been carried out that use data from multiple sources. Finally, for the task of georeferencing Flickr photos, [23] introduce the idea of using evidence from Twitter messages by the same user within a given time interval around the time stamp of the photo.

7.3 Datasets

We will evaluate our techniques using two test collections of Wikipedia articles. The first test set, discussed in detail in Section 7.3.1, is used to compare our approach against earlier work in [7] and [8], but has a number of shortcomings. For this reason, we constructed a second test set of Wikipedia documents, as described in Section 7.3.2. Our training data will consist of Wikipedia articles, in addition to Flickr photos and Twitter messages, as detailed in Sections 7.3.3 to 7.3.5.

7.3.1 Wing and Baldrigde (W&B) Wikipedia training and test set

The training and test data from [7] has been made available on the TextGrounder website[7]. Using this dataset enables us to compare the results reported in [7] and [8] to the results we obtain using our approach. The dataset originates from the original English-language Wikipedia dump of September 4, 2010[8], which was preprocessed as described in [7], and divided into 390 574 training articles and 48 589 test articles. In [8] a slightly modified version of this dataset has been used. Accordingly, we filtered the dataset for the 390 574 Wikipedia training documents and 48 566 Wikipedia test documents that have been used in [8].

However, this test set has a number of shortcomings:

- No distinction is made between Wikipedia articles that describe a precise location on the one hand (e.g. the Eiffel tower), and Wikipedia articles whose

[7] http://code.google.com/p/textgrounder/wiki/WingBaldridge2011
[8] http://download.wikimedia.org/enwiki/20100904/enwiki-20100904-pages-articles.xml.bz2

geographic scope cannot reasonably be approximated by a single coordinate, such as large geographical entities (e.g. rivers, trails or countries) or Wikipedia lists (e.g. "List of shipwrecks in 1964"), on the other hand.

- To create the ground truth, the Wikipedia dump used was filtered for pages that mention a geographical coordinate, while the page itself has no explicitly assigned coordinates. As an example, for the article on "List of shipwrecks in 1964"[9], the ground truth location was set to 44°12'N 08°38'E, which is mentioned in the article in relation to October 14, 1964, the day the ship *Dia* sank south of Savona, Italy.
- As part of the preprocessing considered by [7], all information about word ordering has been removed from the original document. This seriously disadvantages our method which relies on n-grams, because Flickr tags often correspond to the concatenation of several terms.

We have therefore also evaluated our method on a newly crawled test collection, as discussed next.

7.3.2 The Wikipedia spot training and test set

Constructing a dataset from raw dumps of Wikipedia pages requires pre-processing as these pages contain fragments of markup language that are not relevant in this context. On the other hand, certain markup codes provide meaningful information that we would like to keep, such as captions of links to files, images or tables. Our pre-processing script converts the example raw Wikipedia fragment:

```
[[Image:Abbotsford Morris edited.jpg|thumb|300px|right|
Abbotsford in 1880.]] '''Abbotsford''' is a [[historic
house]] in the region of the [[Scottish Borders]] in
the south of [[Scotland]], near [[Melrose]], on the
south bank of the [[River Tweed]]. It was formerly the
residence of [[historical novel]]ist and [[poet]],
[[Walter Scott]]. It is a Category A [[Listed
Building]].
```

to the following text: "Abbotsford in 1880. Abbotsford is a historic house in the region of the Scottish Borders in the south of Scotland, near Melrose, on the south bank of the River Tweed. It was formerly the residence of historical novelist and poet, Walter Scott. It is a Category A Listed Building".

To construct the test set, we downloaded the DBPedia 3.7 "Geographic Coordinates" English (nt) Wikipedia dump[10], containing the geographical coordinates

[9] http://en.wikipedia.org/wiki/List_of_shipwrecks_in_1964

[10] http://downloads.dbpedia.org/3.7/en/geo_coordinates_en.nt.bz2

and Wikipedia ID's (e.g. "Abbotsford_House") of 442 775 entities. From these, we retained the 47 493 documents whose coordinates are located within the bounding box of the United Kingdom. The raw XML version of these documents have been obtained by posting the (encoded) ID's against Wikipedia's `Special:Export`[11] function.

Wikipedia contains numerous documents that are hard to pinpoint to a precise location, discussing for example architectural styles, schools of thought, people or concepts. As we consider techniques for estimating precise coordinates, it is useful to restrict the evaluation to articles that have a limited spatial extent, such as landmarks, buildings, schools, or railway stations. Although DBPedia lists the coordinates of the documents, it does not provide any information on the "type" or "scale" of the coordinates. However, this information can be extracted from the XML documents by scanning for the Wikipedia coordinate template markup (i.e. `{{coord*}}`) and parsing its contents. After extracting this information, we have further filtered the dataset, keeping only the documents whose coordinates either refer to a location of type "railwaystation, landmark or edu"[12] (being the only types that refer to spots), or have a reported scale of 1:10000 or finer.

The result is a set of 21 839 Wikipedia test documents. This dataset, along with the pre-processing script, has been published online[13]. To make this set compatible with the W&B training set, we removed any occurrences of our test documents from the W&B training data, resulting in a training set of 376 110 Wikipedia documents. This reduced training set is used whenever our "spot" test set is used. When evaluating the W&B test set, we still use the full training set.

Note that, while the spot dataset only contains Wikipedia articles that are located within the bounding box of the UK, our method does not exploit this information. The restriction on the UK is motivated by the possibility of future work, which could consider additional country-specific evidence, such as local news articles.

7.3.3 Flickr training set

In April 2011, we collected the meta-data of 105 118 157 georeferenced Flickr photos using the public Flickr API. We pre-processed the resulting dataset by removing photos with invalid coordinates as well as photos without any tags. For photos that are part of bulk uploads, following [2] we removed all but one photo. This resulted in a set of 43 711 679 photos. Among these photos, we extracted only those that reported an accuracy level of 12 at least, which means that the geographical coordinates of the photos we use are accurate at a city block level. This

[11] http://en.wikipedia.org/wiki/Special:Export

[12] For a full list of Wikipedia GEO types, see http://en.wikipedia.org/wiki/Wikipedia:GEO#type:T

[13] Our pre-processing script, along with the original XML and processed test set are made available online at https://github.com/ovlaere/georeferencing_wikipedia

final step resulted in a set of 37 722 959 photos, of which 32 million photos served as training data for this paper.

7.3.4 Twitter training set

Twitter provides samples of the tweets published by its users[14]. We monitored the "Gardenhose" stream using the `statuses/filter` API method in combination with a bounding box parameter covering the entire world. This allowed us to track only Twitter messages with a geographical coordinate attached to them. Doing so for a period from March to August 2012 resulted in a dataset of 170 668 054 tweets.

In order to avoid an unfair bias in the number of word occurrences at certain locations caused by a single user, we aggregated all tweets from a given user at the same location into a single document. The resulting document is represented as a set of terms, i.e. multiple occurrences of the same term at the same location by the same user are only counted once. For example:

```
52.135978 -0.466651  Olympic torch http://t.co/q3yNthcj
52.135978 -0.466651  Olympic torch http://t.co/wZUH4a5B
52.135978 -0.466651  Olympic torch http://t.co/M9Tm6Ow0
52.135978 -0.466651  Olympic torch http://t.co/HWqiTDZy
52.135978 -0.466651  Olympic torch http://t.co/2ovhQdPu
52.135978 -0.466651  Olympic torch http://t.co/iIRvEe5C
52.135978 -0.466651  Olympic torch http://t.co/hO8PAsf1
```

then becomes:

```
52.135978 -0.466651 Olympic torch http://t.co/q3yNthcj
http://t.co/wZUH4a5B http://t.co/M9Tm6Ow0
http://t.co/HWqiTDZy http://t.co/2ovhQdPu
http://t.co/iIRvEe5C http://t.co/hO8PAsf1
```

Next, we only retained those documents in which at least one hashtag (e.g. #empirestatebuilding) occurs, further reducing the dataset to 18 952 535 documents. In this paper we used a subset of 16 million of these documents as training data.

7.3.5 Data compatibility

Further pre-processing was needed to arrive at a meaningful combination of Wikipedia, Flickr and Twitter data. For example, while Wikipedia documents contain capitalized words, the Flickr tags are all lowercase and moreover often correspond to the concatenation of several words, e.g. photos on Flickr may be tagged as "empirestatebuilding". This has implications in two steps of our approach:

[14] https://dev.twitter.com/docs/streaming-apis/streams/public

1 the estimation of a language model from Flickr or Twitter data while test documents are taken from Wikipedia. (Section 7.4.3).

2 the comparison of similarity between a test document and training items from the selected area are compared, in the procedure from Section 7.4.5.

7.3.5.1 Wikipedia documents and Flickr data

To make the Wikipedia test data compatible with the Flickr training data, we can "translate" the documents to Flickr tags. This can easily be achieved by converting the Wikipedia test articles to lowercase, and scanning for terms or concatenations of up to 5 consecutive terms that correspond to a Flickr tag from the training data.

7.3.5.2 Wikipedia documents and Twitter documents

To facilitate comparison between Wikipedia test data and Twitter training data, we convert all terms to lowercase and for each of the occurring hashtags, we remove the leading "#" sign. Again, we scan the Wikipedia documents for terms or concatenations of up to 5 consecutive terms that correspond to any term occuring in the Twitter training data, as especially hashtags may correspond to the concatenation of several terms.

7.4 Estimating locations using language modelling

Probabilistic (unigram) language models have proven particularly effective to estimate the location of textual resources [2, 7, 8]. In this section we will detail the adopted approach, which is based on the algorithm outlined in [3]. The fundamental addition to this method consists of the fact that the models were trained using a combination of Wikipedia, Flickr and Twitter data. This implies two modifications to the approach outlined before:

1 There is need for a way to combine different language models

2 The last phase of our approach involves assessing the similarity between the item to be georeferenced and the items in the training set. This means that we need a way of measuring the similarity between e.g. a Wikipedia article and a Flickr photo.

Our approach consists of two main steps. First, we treat the problem of estimating the location of an unseen document $\mathcal{D}$ as a text classification problem. To this end, the coordinates appearing in the training data are clustered into k distinct areas a, that make up the clustering $\mathcal{A}_k$. After clustering, a feature selection procedure is applied aimed at removing terms that are not spatially relevant (e.g.

removing tags such as *birthday* or *beautiful*). In particular, we select a vocabulary $\mathcal{V}$ of m features. Given a specific clustering into $\mathcal{A}_k$ and the vocabulary of features $\mathcal{V}$, language models for each cluster can be estimated. By assessing which of these language models has most likely generated the terms in $\mathcal{D}$, we can determine the area $a \in \mathcal{A}_k$ that is most likely to contain the location of $\mathcal{D}$. In the second step, once an area a has been chosen, we estimate the location of $\mathcal{D}$ as the location of the training item from area a that is most similar to $\mathcal{D}$. Next, we discuss each of these steps in more detail.

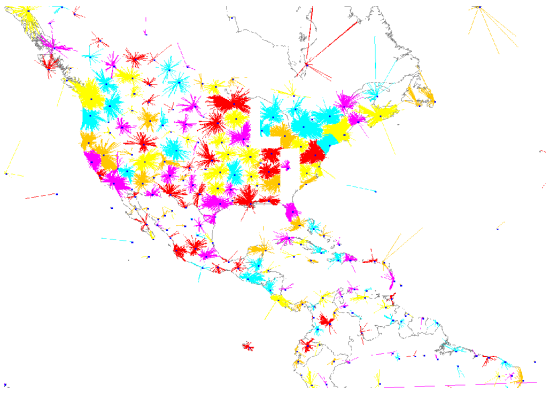
7.4.1 Clustering

To cluster the training data, we have used the k-medoids algorithm, which is closely related to the well-known k-means algorithm but is more robust to outliers. Distances are evaluated using the geodesic (great-circle) distance measure. Other authors have used a grid-based clustering or mean-shift clustering, but experiments in [24] have shown k-medoids to be better suited for this task. A grid clustering ignores the fact that certain grid cells contain much more information than others, allowing more precise location estimations in that part of the world. Mean-shift clustering has a similar issue, and results in clusters which are all of approximately the same scale, independent of the amount of training data that is available for that region of the world. In contrast, k-medoids yields smaller clusters when the data density is higher and larger clusters when data is sparser. Figures 7.1(c) to 7.1(b) illustrate this difference, which is clearly visible when looking at California and New York.

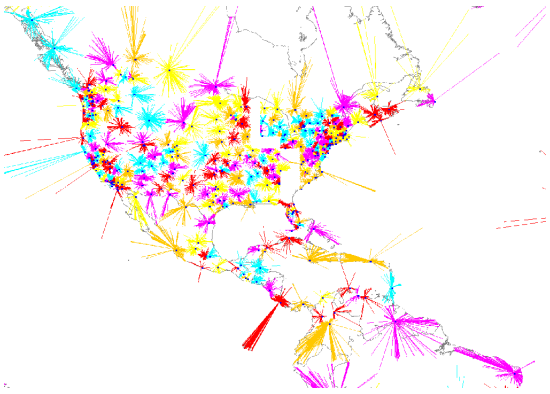
7.4.2 Feature selection

We adopted the *geographic spread filtering* method presented in [25], which has proven to outperform other traditional feature selection techniques at the task of georeferencing [24]. The method determines a score that captures to what extent the occurrences of a term are clustered around a small number of locations. The geographical spread score is calculated as follows:

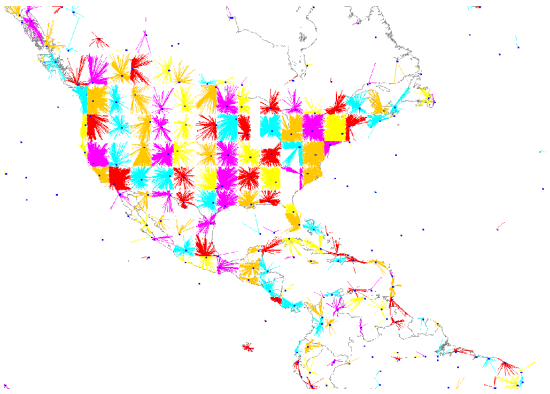
Place a grid over the world map with each cell having sides of 1 degree latitude and longitude
for each unique term t in the training data **do**
  for each cell $c_{i,j}$ **do**
    Determine $|t_{i,j}|$, the number of training documents containing the term t
    if $|t_{i,j}| > 0$ **then**
      for each $c_{i',j'} \in \{c_{i-1,j}, c_{i+1,j}, c_{i,j-1}, c_{i,j+1}\}$, the neighbouring cells of $c_{i,j}$, **do**
        Determine $|t_{i',j'}|$
        if $|t_{i',j'}| > 0$ and $c_{i,j}$ and $c_{i',j'}$ are not already connected **then**



(a) Mean-shift clustering



(b) *k*-medoids clustering

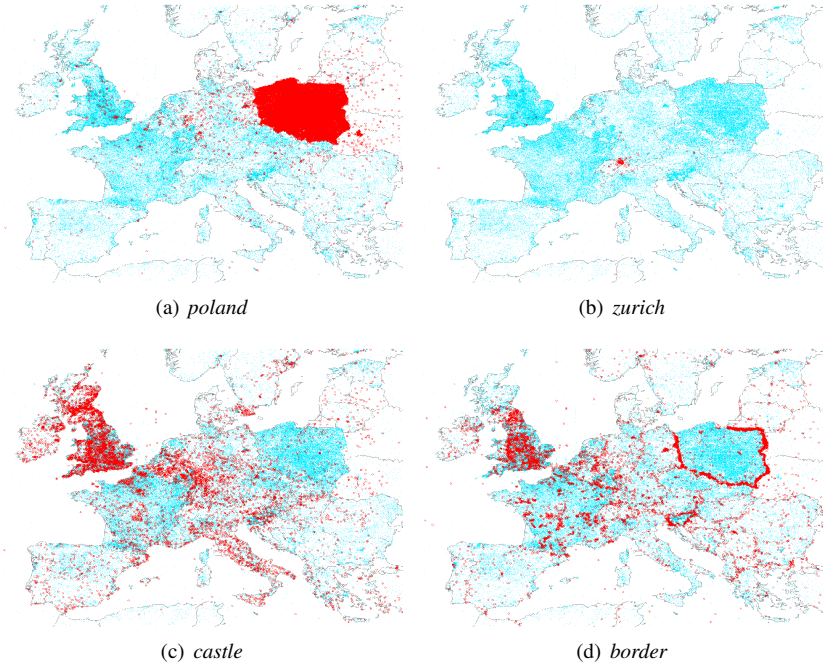


(c) Grid clustering

Figure 7.1: Comparison of three different clustering algorithms on the same subset of data.

Connect cells $c_{i,j}$ and $c_{i',j'}$
end if
end for
end if
end for
$count$ = number of remaining connected components
$score(t) = \frac{count}{\max_{i,j} |t_{i,j}|}$
end for

In the algorithm, merging neighbouring cells is necessary in order to avoid penalizing geographic terms that cover a wider area. The smaller the score for a term t, the more specific its geographic scope and thus the more it is coupled to a specific location. Figures 7.2(a) to 7.2(d) illustrate both terms with a high and low geographical spread score.



(a) *poland* (b) *zurich*

(c) *castle* (d) *border*

Figure 7.2: Examples of occurrences (highlighted in red) in the Wikipedia training data of two terms with a low geographical spread, poland *and* zurich*, and two more general terms with a high spread,* castle *and* border.

7.4.3 Language modelling

Given a previously unseen document $\mathcal{D}$, we now attempt to determine in which area $a \in \mathcal{A}_k$ it most likely relates. We use a (multinomial) Naive Bayes classifier, which has the advantage of being simple, efficient, and robust. Results from [2] have shown good performance for this classifier. Specifically, we assume that a document $\mathcal{D}$ is represented by a collection of term occurrences $\mathcal{T}$. Using Bayes' rule, we know that the probability $P(a|\mathcal{D})$ that document D was taken in area a is given by

$$P(a|\mathcal{D}) = \frac{P(a) \cdot P(\mathcal{D}|a)}{P(\mathcal{D})}$$

Using the assumption that the probability $P(\mathcal{D})$ of observing the terms associated with document $\mathcal{D}$ does not depend on the area a, we find

$$P(a|\mathcal{D}) \propto P(a) \cdot P(\mathcal{D}|a)$$

Characteristic of Naive Bayes is the simplifying assumption that all features are independent. Translated to our context, this means that the presence of a given term does not influence the presence or absence of other terms. Writing $P(t|a)$ for the probability of a term t being associated to a document in area a, we find

$$P(a|\mathcal{D}) \propto P(a) \cdot \prod_{t \in \mathcal{T}} P(t|a) \tag{7.1}$$

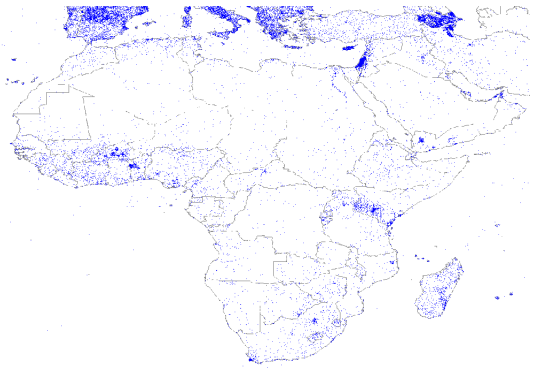
After moving to log-space to avoid numerical underflow, this leads to identifying the area a^* where $\mathcal{D}$ was most likely taken by:

$$a^* = \arg\max_{a \in \mathcal{A}} \left(\log P(a) + \sum_{t \in \mathcal{T}} \log P(t|a) \right) \tag{7.2}$$

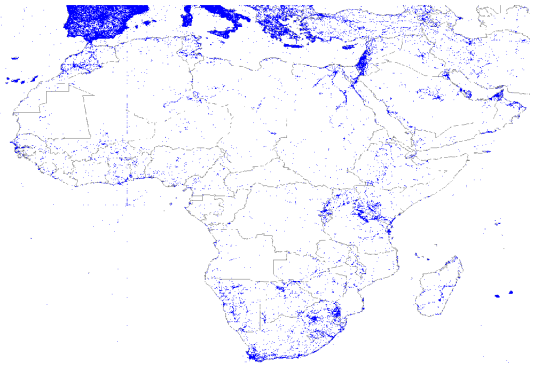
In Equation (7.2), the prior probability $P(a)$ and the probability $P(t|a)$ remain to be estimated. In general, the maximum likelihood estimation can be used to obtain a good estimate of the prior probability:

$$P(a) = \frac{|a|}{N} \tag{7.3}$$

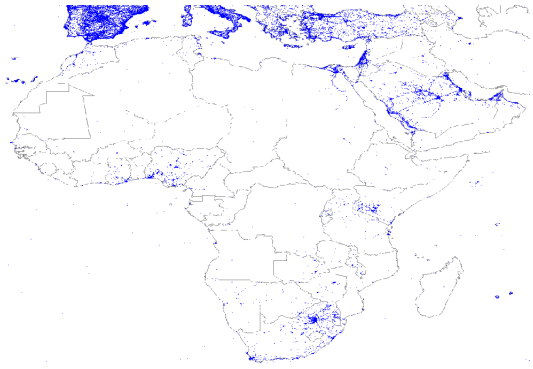
in which $|a|$ represents the number of training documents contained in area a, and N represents the total number of training documents. This reflects the bias of the considered source. For instance, all things being equal, a photo on Flickr has more likely been taken in Western Europe than in Africa. In our setting, in which test data are Wikipedia articles and training data may be taken from Flickr, Twitter and Wikipedia, the justification for the maximum likelihood estimation may appear less strong. However, it should be noted that the geographic bias of Flickr, Twitter and Wikipedia is quite similar, as Figures 7.3(a) to 7.3(c) show, illustrating the



(a) Wikipedia data (390K training items)



(b) Flickr data (32M training items)



(c) Twitter data (16M training items)

Figure 7.3: A qualitative comparison of the data coverage of the different sources of training data over Africa.

coverage of our datasets over Africa. In other contexts, where test items may have a different geographic bias, a uniform prior probability could be more appropriate.

To avoid estimating unreliable probabilities, when only a limited amount of information is available, and to avoid a zero probability when $\mathcal{D}$ contains a term that does not occur with any of the documents from area a in the training data, smoothing is needed when estimating $P(t|a)$ in Equation (7.1). Let O_{ta} be the number of times t occurs in area a. The total term occurrence count O_a of area a is then defined as follows:

$$O_a = \sum_{t \in \mathcal{V}} O_{ta} \tag{7.4}$$

where $\mathcal{V}$ is the vocabulary that was obtained after feature selection, as explained in Section 7.4.2. When using Bayesian smoothing with Dirichlet priors, we have ($\mu > 0$):

$$P(t|a) = \frac{O_{ta} + \mu\, P(t|\mathcal{V})}{O_a + \mu} \tag{7.5}$$

where the probabilistic model of the vocabulary $P(t|\mathcal{V})$ is defined using maximum likelihood:

$$P(t|\mathcal{V}) = \frac{\sum_{a \in \mathcal{A}} O_{ta}}{\sum_{t' \in \mathcal{V}} \sum_{a \in \mathcal{A}} O_{ta}} \tag{7.6}$$

For more details on smoothing methods for language models, we refer to [26].

7.4.4 Combining language models

To combine language models estimated from different sources $\mathcal{S}$, e.g. $\mathcal{S} = \{Wikipedia, Flickr, Twitter\}$, (7.2) can be modified to include weight factors λ_{model}:

$$a^* = \arg\max_{a \in \mathcal{A}_k} \Big(\sum_{model \in \mathcal{S}} \lambda_{model} \cdot \log(P_{model}(a|\mathcal{D})) \Big) \tag{7.7}$$

The area a maximizing expression (7.7), using the probabilities produced by all the different models in $\mathcal{S}$, is then chosen as the area that is most likely to contain the given test document $\mathcal{D}$. The parameters λ_{model} can be used to control the influence of each model on the overall probability for a given area a. In particular, if a given model is less reliable, e.g. because it was trained on a small amount of training data or because the training data is known to be noisy (e.g. many tweets talk about places that are not at the associated location of the user), λ_{model} can be set to a small value.

In practice, we compute the models in memory. This makes it unfeasible to store the probabilities for each of the k areas for each test document and for each

of the language models, at the same time. To cope with this, we compute each model separately and store the top-100 areas with the highest probabilities for each test document $\mathcal{D}$ in the given model. By doing so, probabilities $P_{model}(a|D)$ for certain areas $a \in \mathcal{A}_k$ will be missing in Equation (7.7), which we estimate as follows:

$$P^*_{model}(a|\mathcal{D}) = \begin{cases} P_{model}(a|\mathcal{D}) & \text{if } a \text{ in top-100} \\ \min_{a' \text{ in top-100}} P_{model}(a'|\mathcal{D}) & \text{otherwise} \end{cases}$$

7.4.5 Location estimation

We consider three different ways of choosing a precise location, once a suitable area a has been found.

7.4.5.1 Medoid

The most straightforward solution is to choose the location of the medoid m_a, defined as:

$$m_a = \underset{x \in \textit{Train}(a)}{\arg\min} \sum_{y \in \textit{Train}(a)} d(x, y) \tag{7.8}$$

where *Train*(a) represents the set of training documents located in area a and $d(x, y)$ is the geodesic distance between the locations of documents x and y. This comes down to the idea of selecting the location of the training document that is most centrally located among all documents in a. While this method is rather straightforward, it can still give reasonable location estimates when the number of clusters k is sufficiently large.

7.4.5.2 Jaccard similarity

Another solution consists of returning the location of the most similar training document in terms the Jaccard measure:

$$s_{jacc}(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

where we identify a document with its *set* of terms, *without* considering feature selection. Using feature selection here would be harmful as there may be rare terms (e.g. the name of a local restaurant) or terms without a clear geographic focus (e.g. *castle*) that could be very helpful in finding the exact location of a document.

7.4.5.3 Lucene

A third and final solution is to use Apache Lucene. The fact that Jaccard similarity does not take multiple occurrences of a given feature into account is not an issue when considering Flickr tags. However, when the test and/or training data consists of Wikipedia documents, this could potentially be a shortcoming. Also, [27] have shown that Lucene can be effective in finding similar Flickr photos as well. To find the training document in area a that is most similar to $\mathcal{D}$, we use Lucene search with its default scoring mechanism[15].

7.5 Experimental evaluation

7.5.1 Methodology

In this section, we discuss the results of experiments addressing the research questions stated in Section 7.1. In Sections 7.5.2 and 7.5.4, we establish baseline results for both of the Wikipedia test sets. To this end, we georeference the test documents using only language models trained using other Wikipedia documents. Subsequently, we evaluate the results when using language models only trained using Flickr or Twitter data. After describing the baseline approach, we discuss the effect of combining different language models in Sections 7.5.3 and 7.5.5. Sections 5.6 to 5.9 provide detailed insights in the results. Finally, in Section 7.5.10, we compare the results of our method on both test sets against Yahoo! Placemaker, which is a gazetteer-based service for georeferencing arbitrary web documents.

Baseline approach The approach outlined in Section 7.4 requires several parameters, including the number of features to select and a parameter controlling the amount of smoothing. A detailed analysis of the influence of each of these parameters is beyond the scope of this paper, as a detailed study was conducted in [24]. To focus on the core research questions of this paper, we have therefore fixed the following parameters:

- the number of features used by the feature selection algorithm (Section 7.4.2) was set to 250 000 features for the Wikipedia training data, and 150 000 features for the Flickr and Twitter training sets.
- the smoothing parameter μ, used for the Bayesian smoothing with Dirichlet priors in the language models (Section 7.4.3), was set to 15 000.

We evaluate the results of the experiments using the following metrics:

[15] For details on this scoring function, we refer to http://lucene.apache.org/core/3_6_1/api/all/org/apache/lucene/search/Similarity.html

1 The **accuracy** of the classifier for the given clustering. This is given by $\frac{P}{P+N}$ where P is the number of test documents that have been assigned to the correct cluster and N is the number of documents that have not.

2 For each test document, the distance error is calculated as the distance between the predicted and the true location. The **median error** distance is used as an evaluation metric. This allows us to observe, using a single value, the overall scale of the errors made for a given test collection.

3 From the aforementioned error distances, we also calculate the percentage of the test items that were predicted within 1 m, 10 m, 100 m, 1 km, 10 km, 100 km and 1000 km of their true location, which we refer to as **Acc@Kkm**, with K being the threshold distance in kilometer.

7.5.2 Baseline results for the W&B dataset

Table 7.1 presents the baseline results on the W&B dataset (Section 7.3.1). The optimal results are highlighted in light-blue. The values for the approach taken by [8] are gathered by parsing and evaluating the log files as provided by the authors. These results were obtained using a *k-d*-based clustering of the training data and finding the cluster which is most similar to the Wikipedia document in terms of the Kullback-Leibler divergence.

Overall, the approach taken by [8] achieves better results at the higher error distance thresholds (most notably at 10 km and 100 km), whereas our approach achieves better results at the lower thresholds (most notable at 0.1 km and 1 km), both when using the same training data from Wikipedia and when using training data from Flickr. This difference with [8] can be explained as follows. By returning the centre-of-gravity of the area that was found by the classifier, [8] takes a rather cautious approach, as the centre is reasonably close to most elements in the area. Our method, on the other hand, tries to identify the exact location within an area; cases for which this is successful explain why we do better at the lower thresholds and cases for which this step fails are partially responsible for the worse results at the higher accuracy levels. Differences in the clustering method and the use of Kullback-Leibler instead of Naive Bayes may also lead to some changes in the results. For example, when using fewer clusters, more emphasis is put on the similarity search step which in general is more errorprone. This effect may explain why using 50000 clusters yields better results than using 2500 clusters at the 1 km and 10 km thresholds for the Wikipedia training data.

Interesting to see in Table 7.1 is that the highest Acc@0.1 km and Acc@1 km values are obtained using a language model trained using 32M Flickr photos, with the difference at 1 km being especially pronounced. This result is all the more remarkable because the Flickr model cannot be used to its full potential given that

Table 7.1: Comparison between the results from [8] and our framework from Section 7.4 when trained using Wikipedia, Flickr and Twitter documents separately (W&B dataset). The different k-values represent the number of clusters used while the maximal values across all three models in the table are highlighted for each of the different accuracies, as well as the minimal median error.

Wikipedia	Roller et al	$k = 2500$	$k = 10000$	$k = 25000$	$k = 50000$
Median Error	13.36 km	22.25 km	19.26 km	19.13 km	19.58 km
Accuracy	N/A	64.18%	49.02%	35.72%	26.31%
Acc@0.001 km	0.1%	1.1%	1.06%	1.03%	0.99%
Acc@0.01 km	0.1%	1.15%	1.12%	1.09%	1.05%
Acc@0.1 km	0.16%	1.58%	1.58%	1.55%	1.48%
Acc@1 km	3.53%	5.62%	6.05%	6.28%	6.34%
Acc@10 km	42.75%	32.42%	35.58%	36.19%	36.01%
Acc@100 km	86.54%	79.34%	80.1%	79.01%	77.77%
Acc@1000 km	97.42%	95.73%	95.6%	94.97%	94.21%
Flickr 32M		$k = 2500$	$k = 10000$	$k = 25000$	$k = 50000$
Median Error		51.14 km	48.94 km	50.77 km	53.32 km
Accuracy		44.26%	29.29%	20.64%	15.22%
Acc@0.001 km		0.02%	0.02%	0.02%	0.03%
Acc@0.01 km		0.21%	0.2%	0.19%	0.16%
Acc@0.1 km		2.61%	2.39%	2.14%	1.88%
Acc@1 km		11.25%	10.15%	9.18%	8.4%
Acc@10 km		26.26%	26.75%	25.94%	25.11%
Acc@100 km		62.78%	63%	62.24%	61.2%
Acc@1000 km		88.6%	87.78%	87.09%	86.35%
Twitter 16M		$k = 2500$	$k = 10000$	$k = 25000$	$k = 50000$
Median Error		350.58 km	406.7 km	427.58 km	469.68 km
Accuracy		24.02%	14.61%	9.72%	7.14%
Acc@0.001 km		0%	0.01%	0%	0%
Acc@0.01 km		0.01%	0.01%	0.02%	0.02%
Acc@0.1 km		0.04%	0.07%	0.11%	0.16%
Acc@1 km		0.66%	1.32%	1.71%	1.94%
Acc@10 km		8.56%	11.82%	12.69%	12.74%
Acc@100 km		36.31%	36.01%	35.34%	34.55%
Acc@1000 km		61.05%	59.23%	58.36%	57.05%

the W&B dataset only supports the use of unigrams (see Section 3.1). Finally, even though the results from using a model trained on 16M Twitter documents are worse than the two other models, it is noteworthy that is still allows to locate 1.94% of the Wikipedia documents within 1 km of their true location.

7.5.3 Combining language models using training data from social media (W&B dataset)

7.5.3.1 Wikipedia + Flickr + Twitter

Figure 7.4 shows the result of combining the language models from Wikipedia, Flickr and Twitter, using $\lambda_{flickr} = 0.5, \lambda_{twitter} = 0.15$[16]. The graphs consist of two parts. On the left, we start with a pure Wikipedia model (*Wiki*) and combine this model with different Flickr models trained using a gradually increasing amount of training Flickr photos (up to 32M) (F_{1M} to F_{32M}). In the center of the graphs, where the shaded area begins, we start with the $Wiki + F_{32M}$ model and continue to combine with language models from Twitter trained using up to 16M documents (T_{1M} to T_{16M}). The location estimate returned for each test document is the location from the most similar training item overall (i.e. a Wikipedia document[17], a Flickr photo or a Twitter document) in the cluster selected by the classifier. As for the results in Table 7.1, the Jaccard similarity is used for this purpose. As before, results are evaluated on the W&B test data and the number of clusters is varied from 2500 to 50000.

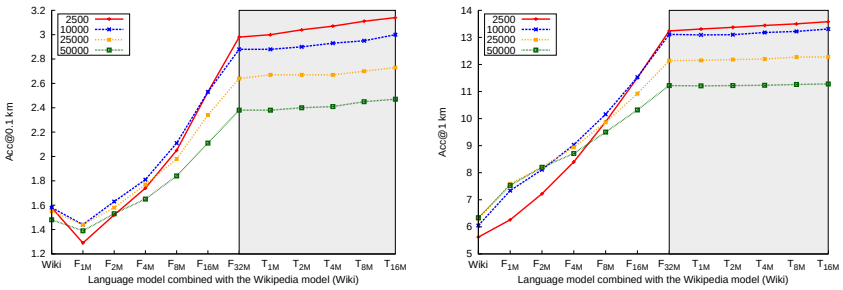
The combination $Wiki + F_{32M}$ in Figure 7.4(a) only shows an increase of 1.4%, which is somewhat dissappointing. We assume this is partially due to the fact that not all test documents from the W&B dataset correspond to a spot. For instance, it does not make sense to estimate an exact coordinate for a test document such as "Sante Fe Trail"[18].

As Figure 7.4(b) to Figure 7.4(d) show, the optimal number of clusters (k) to use depends on the accuracy level we aim for. Using a smaller number of clusters combined with a fairly large amount of training data substantially improves the results for the 1 km threshold. This is due to a trade-off between the classification and similarity search step: using fewer clusters means that the similarity search becomes more important in estimating a good location, a process that is facilitated when more training data is available. By increasing the amount of training data, the similarity search is more likely to find a good match, especially when using millions of Flickr photos. Chances of finding such a similar document in a training set of Wikipedia documents are small due to the fact that no two Wikipedia

[16] A detailed discussion of the influence of these parameter values follows in Section 7.5.7.

[17] In fact, we only use the title of Wikipedia documents during similarity search. We will come back to this in Section 5.9.

[18] http://en.wikipedia.org/wiki/Santa_Fe_Trail



(a) Acc@0.1 km.

(b) Acc@1 km.

(c) Acc@10 km.

(d) Acc@100 km.

Figure 7.4: Percentage of the test documents located within different error distances on the W&B test set, when combining the language model from Wikipedia with Flickr (on the left side) and subsequently with Twitter models (in the shaded area) trained over an increasing amount of information and for different numbers of clusters k.

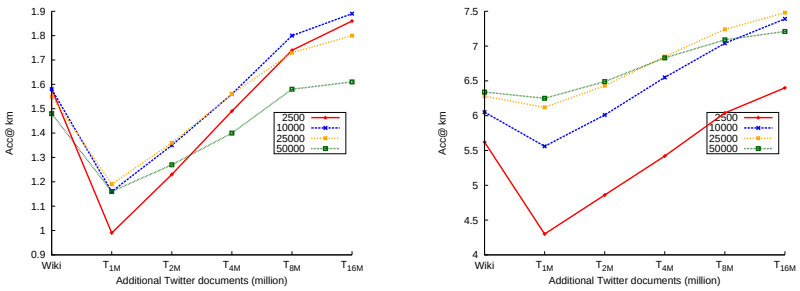
documents should cover the same topic.

All the graphs also show a deterioration of the results when extending the Wikipedia training data with 1M Flickr photos. With only 1M Flickr photos, the language model is apparently not sufficiently reliable to improve the results from the Wikipedia model.

Looking at the right side of the graphs, it seems that the Twitter data is nearly obsolete: only minor improvements are achieved. It should however be noted that the number of georeferenced tweets made available each day is substantially larger than the number of georeferenced Flickr photos, which offers opportunities to training language models from hundreds of millions of tweets, which would likely allow for a more substantial contribution.

7.5.3.2 Wikipedia + Twitter

Using a similar configuration as the previous experiment, we combine the Wikipedia language model with Twitter models trained over up to 16M documents. The results are shown in Figure 7.5.



(a) Acc@0.1 km. (b) Acc@1 km.

Figure 7.5: Percentage of the test documents located within error distances of 0.1 km and 1 km on the W&B test set, when combining the language model from Wikipedia with Twitter models trained over an increasing amount of information and for different numbers of clusters k.

As Twitter documents are generally less informative than the tags associated to Flickr photos, the deterioration on the results when using too few training documents is even more pronounced in Figure 7.5(a) than it was in Figure 7.4(a). Still, when sufficient Twitter data becomes available, significant improvements[19] can be obtained in comparison with only using Wikipedia training data.

[19]To evaluate the statistical significance, we used the sign test as the Wilcoxon signed-rank test is unreliable in this situation due to its sensitivity to outliers. The results are significant with a p-value $< 2.2 \times 10^{-16}$.

7.5.4 Baseline results for the spot dataset

Figure 7.4 showed that adding Twitter and especially Flickr has the potential to substantially improve the results. However, as we discussed in Section 7.3.5, the W&B test data ignores word ordering, which is a disadvantage for our approach because Flickr tags and Twitter terms may correspond to concatenations of terms in a Wikipedia document. Therefore, and also in view of the shortcomings described in Section 7.3.1, we propose an evaluation based on another test set.

We establish the baseline results using the spot dataset consisting of 21 839 Wikipedia test documents, in combination with a filtered training set consisting of 376 110 Wikipedia documents, as described in Section 7.3.2. Table 7.2 depicts the results of our framework, using the same parameter settings as for Table 7.1. Again, the maximal values across all three models in the table are highlighted for each of the different accuracies, as well as the minimal median error.

As could be expected given the nature of the test data, the accuracies presented in Table 7.2 are much higher than those for the W&B test set in Table 7.1. A relatively large fraction of the documents can be localized within 1 km of their true location (35.73% as opposed to 11.25%). Using the Flickr model results in a median error of 2.44 km, compared to 4.64 km for the Wikipedia model. This Flickr model outperforms the two other models at the classification accuracies and at all threshold accuracies except Acc@0.001 km. Again, the results from the Twitter model are worse, except for the fact that 8.28% of the test set can be localised within 1 km of their true location.

7.5.5 Combining language models using training data from social media (spot dataset)

7.5.5.1 Wikipedia + Flickr + Twitter

Similar to the experiment carried out on the W&B dataset in Section 7.5.3, we combine the language models obtained from Wikipedia, Flickr and Twitter and evaluate using the spot test collection of 21 839 Wikipedia documents. The results are presented in Figure 7.6.

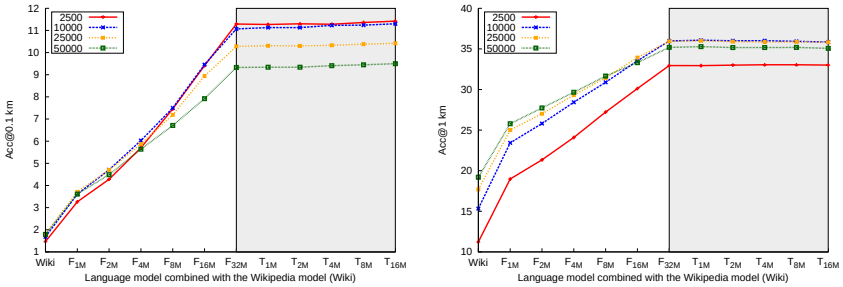
Overall, the relative performance of the different configurations in Figure 7.6 is qualitatively similar to the results for the W&B test set in Figure 7.4, although the magnitude of the improvements is much higher. Given this better performance of the Flickr models, Twitter does not seem to be helpful at all anymore.

7.5.5.2 Wikipedia + Twitter

Figure 7.7 presents the results of combining the Wikipedia language model with Twitter models trained over different amounts of data. In contrast to Figure 7.5,

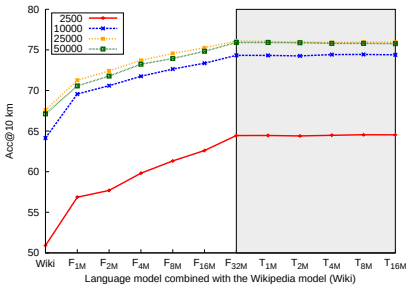
Table 7.2: Comparison of the results from our framework from Section 7.4 when trained using Wikipedia, Flickr and Twitter documents (spot dataset). The different k-values represent the number of clusters used while the maximal values across all three models in the table are highlighted for each of the different accuracies, as well as the minimal median error.

Wikipedia	$k = 2500$	$k = 10000$	$k = 25000$	$k = 50000$
Median Error	9.68 km	5.86 km	4.64 km	4.17 km
Accuracy	70.11%	57.99%	46.21%	36.32%
Acc@0.001 km	0.34%	0.34%	0.33%	0.33%
Acc@0.01 km	0.38%	0.39%	0.38%	0.38%
Acc@0.1 km	1.47%	1.68%	1.81%	1.79%
Acc@1 km	11.23%	15.33%	17.7%	19.2%
Acc@10 km	50.91%	64.15%	67.58%	67.12%
Acc@100 km	93.02%	93.15%	91.38%	90.03%
Acc@1000 km	98.02%	98.34%	97.77%	97.35%
Flickr 32M	$k = 2500$	$k = 10000$	$k = 25000$	$k = 50000$
Median Error	3.7 km	2.6 km	2.44 km	2.5 km
Accuracy	76.97%	63.24%	51.6%	42.35%
Acc@0.001 km	0.06%	0.07%	0.05%	0.06%
Acc@0.01 km	0.95%	0.94%	0.92%	0.84%
Acc@0.1 km	11.52%	11.5%	11.05%	10.49%
Acc@1 km	33.24%	35.45%	35.73%	35.17%
Acc@10 km	63.4%	71.32%	72.44%	71.29%
Acc@100 km	96.47%	96.47%	95.98%	95.48%
Acc@1000 km	98.84%	98.77%	98.63%	98.5%
Twitter 16M	$k = 2500$	$k = 10000$	$k = 25000$	$k = 50000$
Median Error	25.21 km	24.47 km	29.57 km	35.81 km
Accuracy	43.03%	26.83%	18.3%	13.36%
Acc@0.001 km	0%	0%	0%	0%
Acc@0.01 km	0.01%	0.01%	0.01%	0.02%
Acc@0.1 km	0.15%	0.24%	0.23%	0.33%
Acc@1 km	3.14%	6.21%	7.75%	8.28%
Acc@10 km	29.52%	36.98%	36.07%	33.39%
Acc@100 km	72.66%	69.69%	66.7%	64.17%
Acc@1000 km	94.91%	94.23%	93.1%	92.88%

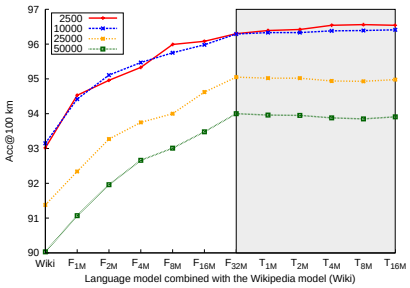


(a) Acc@0.1 km.

(b) Acc@1 km.

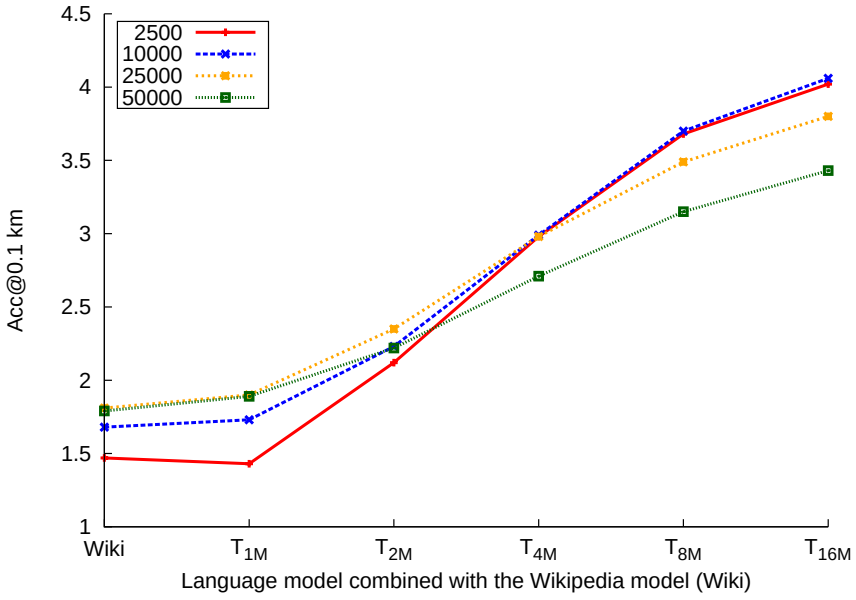


(c) Acc@10 km.

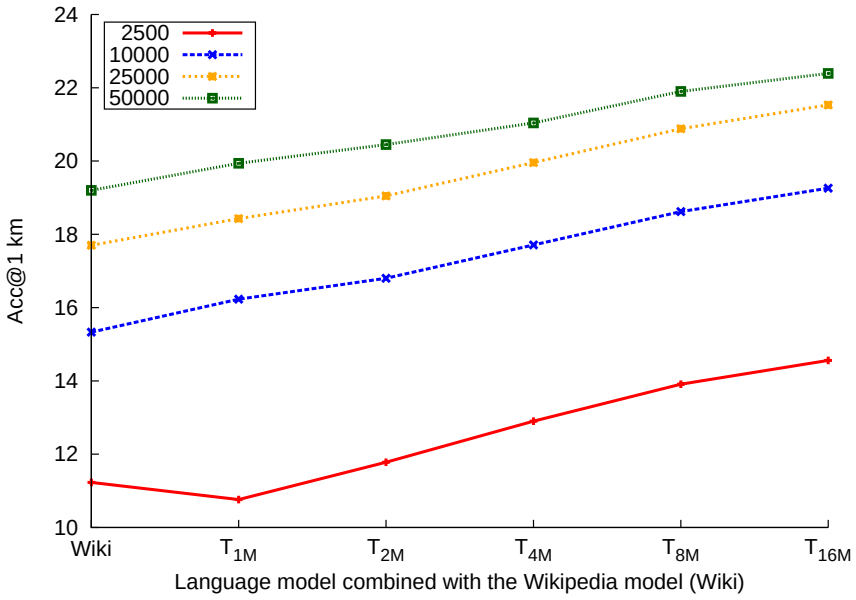


(d) Acc@100 km.

Figure 7.6: Percentage of the test documents located within different error distances on the spot test set, when combining the language model from Wikipedia with Flickr (on the left side) and subsequently with Twitter models (in the shaded area) trained using an increasing amount of information and for different numbers of clusters k.



(a) Acc@0.1 km.



(b) Acc@1 km.

Figure 7.7: Percentage of the test documents located within error distances of 0.1 km and 1 km on the spot test set, when combining the language model from Wikipedia with Twitter models trained using an increasing amount of information and for different numbers of clusters k.

Table 7.3: Comparison of the number of tokens in each of the different training sets (before and after feature selection (FS)). The number of unique tokens is reported, along with the total number of token occurrences, before and after feature selection (see Section 7.4.2).

Dataset	# items	Unique tokens	Total before FS	Total after FS
Wikipedia	390 574	2 817 660	151 325 949	53 134 473
Flickr	1 000 000	563 707	8 395 186	4 829 997
	2 000 000	972 484	17 163 282	8 705 356
	4 000 000	1 732 867	35 597 819	14 667 027
	8 000 000	3 087 690	71 395 087	25 474 723
	16 000 000	5 362 086	143 592 337	44 930 446
	32 000 000	9 269 494	279 109 442	79 968 463
Twitter	1 000 000	2 678 380	18 184 767	7 256 169
	2 000 000	4 667 761	35 581 577	13 796 968
	4 000 000	8 055 391	69 235 192	26 335 231
	8 000 000	13 823 337	136 203 621	51 779 462
	16 000 000	23 077 992	264 632 000	99 964 037
TwitterHashtags	1 000 000	454 884	1 514 359	466 028
	2 000 000	805 521	3 083 544	989 408
	4 000 000	1 428 268	6 188 443	1 937 579
	8 000 000	2 532 145	12 298 065	3 765 776
	16 000 000	4 529 912	24 132 042	6 770 206

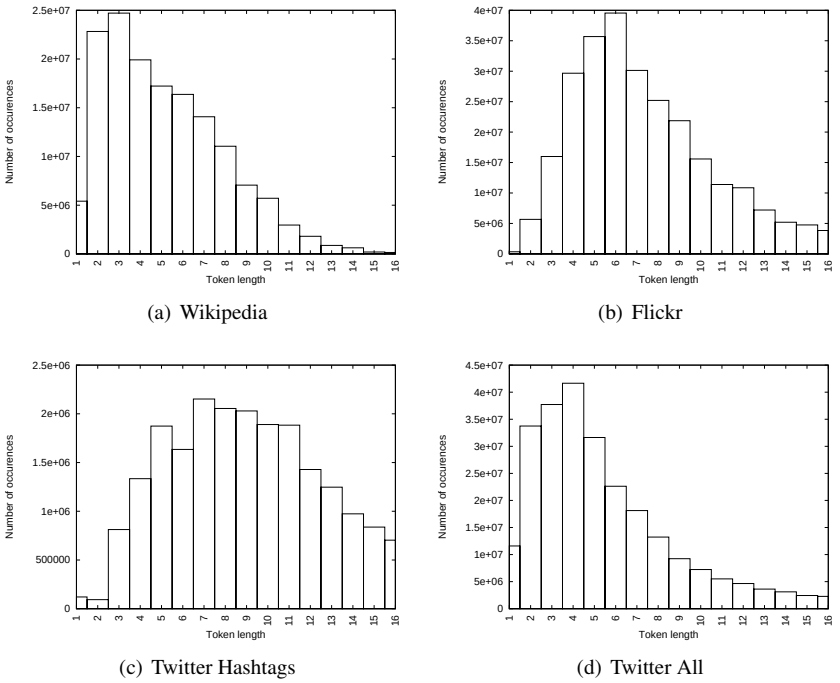
these graphs clearly demonstrate that improvements can be obtained at error margins of 1 km and below by extending the Wikipedia model with only Twitter data. This is remarkable given the difference in structure between a Wikipedia training document and a Twitter message. Also, the deteriorating effect for small amounts of training data is only slightly noticed when using $k = 2500$ clusters.

7.5.6 Training data analysis

It may seem that, by adding for example 32 million Flickr photos to the training data, we are increasing the number of training items by an order of magnitude. However, the amount of textual information that is actually added is comparable to the initial Wikipedia training data, as can be seen in Table 7.3. This is because a Wikipedia training document generally provides a significantly larger amount of textual information (mean of $\simeq$387 tokens) compared to a Flickr training photo (mean of $\simeq$8 tokens). A similar argument holds for Twitter documents with a mean of $\simeq$16 tokens. Table 7.3 provides further details on the unique tokens (words) that occur in the datasets, the total number of tokens in the initial datasets, and the number of tokens that remained after feature selection (see Section 7.4.2).

In addition to our standard Twitter dataset, we included the *TwitterHashtags* variant in Table 7.3, which consists of only the hashtags encountered in the Twitter document. As can be seen from the table, the number of token occurrences is significantly reduced in this dataset, with a mean of $\simeq$0.4 tokens per document. We have omitted the results of this variant in the previous sections, as this dataset produces similar results as the standard Twitter dataset, as can be seen in Table 7.4. This is interesting by itself, as the amount of information used to achieve those results is less than 7.5% of the original Twitter dataset.

Figure 7.8 further summarises some characteristics of the training data, comparing the length of tokens in the different training sets. Note that the mode in Figures 7.8(b) and 7.8(c) is higher than in Figures 7.8(a) and 7.8(d), which is consistent with the idea that tags are more descriptive and therefore likely to be longer, and the view that tags often are concatenation of several words. The latter point is more pronounced in the case of Twitter than in Flickr, as the distribution in Figure 7.8(c) is skewed more towards higher token lengths. The slight difference between Figures 7.8(a) and 7.8(d) in the proportion of tokens of lengths 2 and 3 may be due to the tendency to omit determiners in tweets.



(a) Wikipedia

(b) Flickr

(c) Twitter Hashtags

(d) Twitter All

Figure 7.8: Histograms of the distribution of the word length (up to 16 characters) for the different sources of information, without feature selection.

Table 7.4: Comparison of the percentage of the test documents located within error distances of 0.1 km and 1 km on the spot test set, when combining the language model from Wikipedia with Twitter models, containing all terms and only Hashtags, trained using an increasing amount of information and for different numbers of clusters k.

k	2500	10000	25000	50000	2500	10000	25000	50000
Acc@0.1 km	Twitter				TwitterHash			
Wiki	1.47%	1.68%	1.81%	1.79%	1.47%	1.68%	1.81%	1.79%
T_{1M}	1.43%	1.73%	1.90%	1.89%	1.41%	1.73%	1.88%	1.88%
T_{2M}	2.12%	2.23%	2.35%	2.22%	2.10%	2.19%	2.32%	2.18%
T_{4M}	2.98%	2.99%	2.98%	2.71%	2.94%	2.99%	2.91%	2.62%
T_{8M}	3.68%	3.70%	3.49%	3.15%	3.65%	3.69%	3.43%	3.01%
T_{16M}	4.02%	4.06%	3.80%	3.43%	4.00%	4.02%	3.74%	3.28%
Acc@1 km	Twitter				TwitterHash			
Wiki	11.23%	15.33%	17.70%	19.20%	11.23%	15.33%	17.70%	19.20%
T_{1M}	10.76%	16.23%	18.43%	19.94%	10.73%	16.15%	18.27%	19.87%
T_{2M}	11.78%	16.80%	19.05%	20.45%	11.76%	16.72%	18.88%	20.34%
T_{4M}	12.90%	17.71%	19.96%	21.04%	12.88%	17.69%	19.84%	20.91%
T_{8M}	13.91%	18.62%	20.88%	21.90%	13.97%	18.59%	20.70%	21.61%
T_{16M}	14.56%	19.26%	21.53%	22.39%	14.52%	19.20%	21.34%	22.25%

Table 7.5: Comparison of the results when using different n-grams on the spot dataset. The language model was obtained by combining the Wikipedia, Flickr F_{32M} and Twitter T_{16M} models ($k = 10000$, $\lambda_{flickr} = 2, \lambda_{twitter} = 0.1$).

	1-gram	2-gram	3-gram	4-gram	5-gram
Accuracy	67.05%	69.71%	69.90%	69.92%	69.90%
Median Lucene	3.22 km	2.97 km	2.98 km	2.98 km	2.98 km
Median Similarity	2.31 km	2.05 km	2.03 km	2.02 km	2.02 km

7.5.7 Influence of the λ_{model} parameters when combining different models

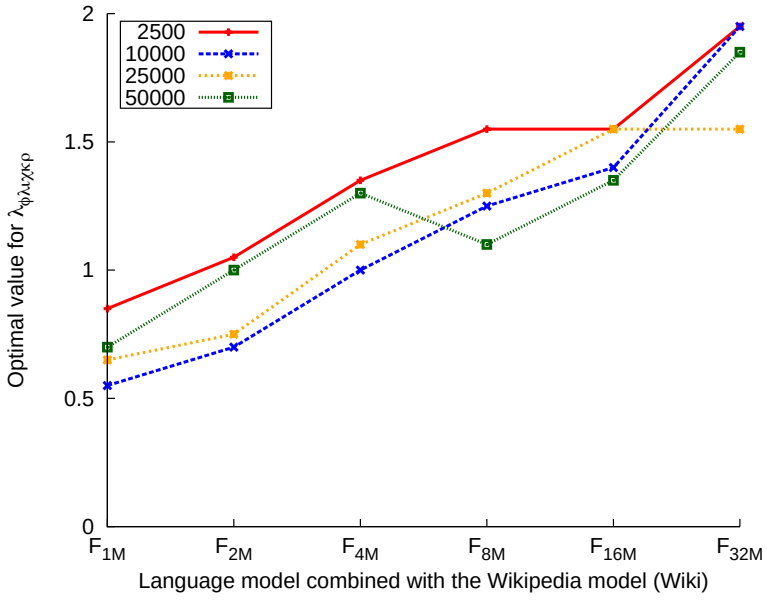
As outlined in Section 7.4.4, the parameter λ_{model} which weighs the different models in Equation (7.7) can play an important role in the results. In Figure 7.9(a), we show, on the spot dataset, for each datapoint the λ_{flickr} value that is optimal when combining the Wikipedia model with each of the Flickr models. As can be expected, the models obtained by using a larger amount of training data prove to be more reliable, allowing to increase the weight λ_{flickr}. The accuracy value for $k = 2500$ at F_{1M} is 75.71% while it increases to 82.15% at F_{32M}.

Figure 7.9(b), shows for each datapoint the $\lambda_{twitter}$ value that was optimal when combining the Wikipedia+F_{32M} model with each of the Twitter models. Unsurprisingly, the $\lambda_{twitter}$ values are low, even for a relatively large amount of training data. For $k = 2500$, it seems that the results become more reliable for more training data. The accuracy value for $k = 2500$ at T_{1M} is 78.58% while it only increases to 79.01% at T_{16M}.

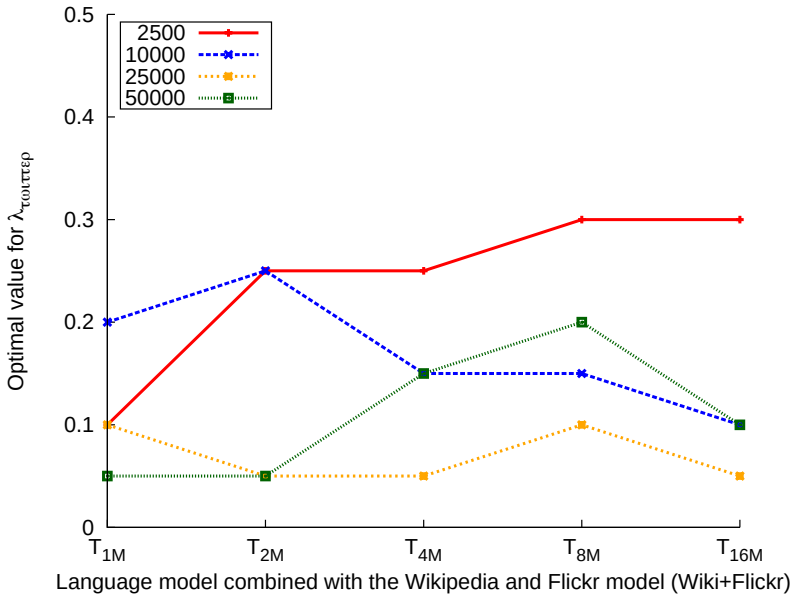
7.5.8 n-grams and similarity search

Table 7.5 illustrates the impact of concatenating words from the Wikipedia training documents to make them compatible with the Flickr and Twitter training data. In this table, we compare the performance of our method when concatenations are not allowed, or limited to a fixed number of consecutive words. We used the spot test set for this table, while the language model was obtained by combining the Wikipedia, Flickr F_{32M} and Twitter T_{16M} models ($k = 10000$, $\lambda_{flickr} = 2, \lambda_{twitter} = 0.1$). The results present both the Lucene similarity and Jaccard similarity to obtain the location estimates for the test documents. As can be seen from the table, allowing longer sequences of words to be concatenated yields higher accuracies and lower median errors, for both similarity methods. In all the experiments for this paper, we used $n = 3$ as the effect of longer sequences does not seem to influence the results substantially.

Table 7.5 shows that the median errors obtained using Jaccard similarity are



(a) Optimal λ_{flickr} for the spot test set



(b) Optimal $\lambda_{twitter}$ for the spot test set

Figure 7.9: Comparing the optimal values for λ under different configurations

Table 7.6: Comparing the results of retrieving the most similar training item using Lucene and Jaccard similarity. These results are shown, using the combined Wikipedia + Flickr (32M) + Twitter (16M) language model and $k = 10000, \lambda_{flickr} = 0.5, \lambda_{twitter} = 0.15$, for both the W&B (left) and spot (right) test set.

	W&B test set		Spot test set	
	Lucene	Jaccard	Lucene	Jaccard
Median Error	16.37 km	17.03 km	3.28 km	2.37 km
Accuracy	50.89%		66.87%	
Acc@0.001 km	0.55%	0.21%	0.18%	0.07%
Acc@0.01 km	0.75%	0.42%	0.84%	0.91%
Acc@0.1 km	2.71%	3.00%	8.38%	11.3%
Acc@1 km	10.64%	13.31%	29.65%	35.85%
Acc@10 km	39.62%	39.71%	74.92%	74.39%
Acc@100 km	82.15%	81.86%	96.44%	96.41%
Acc@1000 km	96.37%	96.34%	99.03%	99.04%

lower than when using Lucene. Table 7.6 compares using Lucene and Jaccard similarity in more detail. These results are based on the combined Wikipedia + Flickr (32M) + Twitter (16M) language model and $k = 10000, \lambda_{flickr} = 0.5, \lambda_{twitter} = 0.15$. Results for both the W&B (left) and spot (right) test set are reported, while the best results for both datasets are highlighted. As can be seen in the table, the results are somewhat mixed.

7.5.9 Similarity search: full content vs. title only

In many cases, the title of a Wikipedia document about a place will be the name of that place. If enough training data from Flickr is available, photos about that place will often be in the training data, and we may try to match the title of the Wikipedia page to the photos in the training data, ignoring the body of the document. Table 7.7 shows the result of using only the page titles for the Jaccard similarity search, compared to using the full document. It should be noted that in the classification step, the full document is used in both cases. The results have been obtained using the combination of the Wikipedia and the F_{32M} Flickr model ($\lambda_{flickr} = 0.5$). We observe the change in median error and Acc@0.001km and Acc@1km, as these are the values that are mainly influenced by the similarity search, whereas the results for the thresholds above 1 km are mostly influenced by the performance of the classifier. As can seen in the table, we observe a substantial improvement when restricting to the title of a Wikipedia page for both datasets. For all the experiments in this paper, the similarity search was carried out using only the Wikipedia page title.

Table 7.7: Comparison between using full wikipedia documents and using titles during similarity search

W&B test set (48 566 items)	$k = 2500$	$k = 10000$	$k = 25000$	$k = 50000$
Median Error Title only	22.43 km	17.14 km	16.85 km	17.4 km
Median Error Full	24.8 km	19.84 km	18.83 km	18.76 km
Acc@0.001 km Title Only	0.17%	0.23%	0.27%	0.31%
Acc@0.001 km Full	0.52%	0.54%	0.58%	0.59%
Acc@1 km Title Only	13.24%	13.11%	12.14%	11.22%
Acc@1 km Full	3.31%	4.39%	5.23%	5.68%
Spot test set (21 839 items)	$k = 2500$	$k = 10000$	$k = 25000$	$k = 50000$
Median Error Title only	3.54 km	2.34 km	2.17 km	2.16 km
Median Error Full	9.26 km	5.40 km	4.00 km	3.31 km
Acc@0.001 km Title Only	0.07%	0.08%	0.06%	0.06%
Acc@0.001 km Full	0.18%	0.20%	0.21%	0.27%
Acc@1 km Title Only	32.95%	35.98%	35.94%	35.19%
Acc@1 km Full	8.65%	13.75%	17.73%	21.01%

7.5.10 Comparing the results to Yahoo! Placemaker

In this section we investigate how the performance of our method relates to the performance of the high-quality information that is available in existing gazetteers. In particular, we compare the result of our combined model (Wikipedia, Flickr and Twitter, $\lambda_{flickr} = 0.5, \lambda_{twitter} = 0.15$), and Yahoo! Placemaker, a freely available webservice capable of georeferencing documents and webpages. Placemaker identifies places mentioned in text, disambiguates those places and returns the centroid for the geographic scope determined for the document. It is important to note that this approach uses external geographical knowledge such as gazetteers and other undocumented sources of information. Placemaker was not able to return a location estimate for all of the documents in our test sets. For this reason, we removed those documents from the evaluation. For the W&B test set, 43 246 documents remained, and 21 265 documents for the spot test set. The results are presented in Tables 7.8 and 7.9 while the optimal results are highlighted. The location estimates for our results are again obtained by using the Jaccard similarity.

In both the tables, the alternative approaches considerably outperform Yahoo! Placemaker, especially in the median error distance. The rather low Acc@10km and Acc@100km for Placemaker on the W&B dataset can be explained by the absence of word ordering in the test set. Placemaker makes use of for example toponym resolution and named entity recognition which is no longer possible. However, when evaluating over the spot dataset, which does not suffer from this drawback, the performance is still poor.

Table 7.8: Comparison of Yahoo! Placemaker, [8] and our approach on the W&B test set (43 246 items).

	Placemaker	Roller et al	$k = 2500$	$k = 10000$	$k = 25000$	$k = 50000$
Median Error	194.13 km	13.17 km	21.79 km	16.86 km	16.51 km	16.94 km
Accuracy	N/A	N/A	67.24%	63.91%	63.31%	62.14%
Acc@0.001 km	0.00%	0.10%	0.16%	0.21%	0.26%	0.30%
Acc@0.01 km	0.05%	0.10%	0.38%	0.42%	0.46%	0.48%
Acc@0.1 km	0.26%	0.15%	3.11%	3.00%	2.72%	2.49%
Acc@1 km	1.75%	3.49%	13.52%	13.23%	12.20%	11.16%
Acc@10 km	9.08%	42.99%	35.92%	39.69%	39.99%	39.50%
Acc@100 km	37.65%	87.32%	80.80%	82.74%	82.11%	80.76%
Acc@1000 km	79.84%	97.69%	97.12%	97.06%	96.50%	95.86%

Table 7.9: Comparison of Yahoo! Placemaker and our approach on the spot test set (21 265 items).

	Placemaker	$k = 2500$	$k = 10000$	$k = 25000$	$k = 50000$
Median Error	28.9 km	3.58 km	2.36 km	2.19 km	2.18 km
Accuracy	N/A	79%	76.92%	76.63%	75.97%
Acc@0.001 km	0.00%	0.07%	0.08%	0.07%	0.06%
Acc@0.01 km	0.03%	0.91%	0.91%	0.84%	0.70%
Acc@0.1 km	0.27%	11.39%	11.28%	10.36%	9.44%
Acc@1 km	4.20%	33.00%	35.86%	35.81%	35.03%
Acc@10 km	27.97%	64.52%	74.46%	76.01%	75.81%
Acc@100 km	74.63%	96.68%	96.55%	95.14%	94.07%
Acc@1000 km	97.75%	99.15%	99.14%	98.81%	98.56%

Table 7.10: Example Wikipedia training documents with unexpected values for their geographical coordinates

Wikipedia name	Latitude	Longitude	Reason
Medusae_Fossae_Formation	-5.0	213.0	On Mars
Quetzalpetlatl_Corona	68.0	357.0	On Venus
Pele_(volcano)	-18.7	-255.3	On Jupiter's moon Io

7.6 Discussion

In addition to general classification errors made by our framework, errors that could potentially be avoided by using more training data, we also noted the following particular issues.

7.6.1 Extraterrestrial coordinates

One of the first anomalies we encountered when processing the Wikipedia training data from the W&B dataset is that certain coordinates had values beyond the expected ranges of latitude ($[-90, 90]$) and longitude ($[-180, 180]$). Table 7.10 provides examples of this. As can be seen from this table, this concerns coordinates that refer to celestial bodies other than the earth. A closer inspection of the training set revealed over 1000 of these extraterrestrial coordinates.

7.6.2 Automated error detection of coordinates

In the spot test set, there is a document about the "Erasmushogeschool Brussel" [20]. The system reported an error of 616.01 km when predicting the location of this test document. Closer inspection revealed that the ground truth for this Wikipedia page was incorrect, and our predicted location was actually the correct location for the place. In particular, the coordinates were reported as `50.7998 N 4.4151 W` instead of an *eastern* longitude which is likely to be due to a manual error.

This example suggests an idea to automatically detect errors in coordinates. If one or multiple sources in which we are highly confident claim that a document is located somewhere else than the current coordinates state, the framework could automatically correct the Wikipedia page. In the spot test collection, we detected three such errors, of which two have since been corrected on Wikipedia (as can be observed in their editing history): "Erasmushogeschool Brussel", which still has the incorrect coordinates online, "Monmouth Hospital"[21] and "Barryscourt Castle"[22].

[20] http://en.wikipedia.org/wiki/Erasmushogeschool_Brussel
[21] http://en.wikipedia.org/wiki/Monmouth_Hospital
[22] http://en.wikipedia.org/wiki/Barryscourt_Castle

7.6.3 Exact matches

Following the idea that no two Wikipedia documents cover exactly the same topic, we would expect not to find any two documents sharing the exact same coordinates. However, looking at the results of Tables 7.1 and 7.2, there are a number of test documents that can be georeferenced to the exact correct location. After manually assessing these cases, we can divide the exact matches into the following categories:

- **Generic coordinates**: Generic coordinates are assigned to different pages that have something in common. For instance, the Wikipedia pages for *Liberia* (in the training data), the West-African country, and its capital *Monrovia* (in the test data), have the same coordinates. The reason for this is that the coordinates in the W&B dataset are obtained by processing the Wikipedia dump data and the coordinate of Monrovia is the first one mentioned in the raw page of Liberia. A similar argument holds for the pages of *Geography of Albania* (test) and *Albania* (training).

- **Identical concepts known by multiple names**: Certain training and test documents actually describe the same location. Apart from concepts known by different names, this can also be due to a change of name over time. This results in duplicates that are sometimes overlooked by Wikipedia authors. Some examples of changes over time are *Tunitas, California* (training) which is a ghost town that changed its name to into the town of *Lobitos* (test). Another example is the former *Free City of Danzig* (test), now known as *Gdańsk*.

- **Different concept but related coordinates**: This category hosts the most interesting matches. For example, the system managed to determine the location of the *MV Languedoc* (test) by providing the coordinates of the *SS Scoresby* (training). Both ships were torpedoed by the U-48 submarine and sunk in the same location. Another example of items that fall into this category are concepts that have their own Wikipedia page but are actually part of a more well-known concept, such as *Queen Elizabeth II Great Court* (test) as part of the *British Museum* (training) or *Larmer Tree Gardens* (test) that hosts the *Larmer Tree Festival*. [7] also provides a brief discussion of this category of examples.

7.7 Conclusions

In this paper, we have presented an approach to georeferencing Wikipedia documents that combines language models trained over different sources of information. In particular, we combine Wikipedia training data with models trained using

Flickr and Twitter, to account for the fact that the places described in a Wikipedia article may already be described in Flickr or Twitter. Overall, we have found that language models trained from Flickr can have a substantial impact on the quality of the produced geotags. As the number of Flickr photos increases every day, the potential of this method continuously increases, although the law of diminishing returns is likely to apply. For this reason, it may be important to consider a broader set of sources. The results we obtained for Twitter were less encouraging: unless language models are trained using billions of tweets, the use of Twitter does not offer substantial performance benefits. It should be noted, however, that various improvements for Twitter may be conceived. In particular, it may be possible to identify messages that are about the current location of the user (e.g. messages beginning with "I'm at") and training models from such messages may be more effective. As part of future work, we intend to look at other sources, such as local news stories, although exact coordinates are usually not available for such resources. As part of a solution, we may envision a system which returns the name of a neighbourhood, for instance, instead of coordinates. This relates to the challenge, discussed in [20], of finding the most appropriate level of granularity at which to estimate the location of a resource. Given the Wikipedia page for the Tour de France[23], for instance, identifying a precise coordinate does not make much sense. Rather, as system that can identify "France" as the most appropriate location estimate may be used (or a polygon which more or less covers France).

Acknowledgments The authors would like to thank Benjamin Wing, Jason Baldridge and Stephen Roller for providing us their dataset and with the details of their experimental results.

[23] http://en.wikipedia.org/wiki/Tour_de_France

References

[1] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. *Mapping the world's photos*. In Proceedings of the 18th International Conference on World Wide Web, pages 761–770, 2009.

[2] P. Serdyukov, V. Murdock, and R. van Zwol. *Placing flickr photos on a map*. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 484–491, 2009.

[3] O. Van Laere, S. Schockaert, and B. Dhoedt. *Finding locations of Flickr resources using language models and similarity search*. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval, pages 48:1–48:8, 2011.

[4] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. *A latent variable model for geographic lexical variation*. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1277–1287, 2010.

[5] Z. Cheng, J. Caverlee, and K. Lee. *You are where you tweet: a content-based approach to geo-locating Twitter users*. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pages 759–768, 2010.

[6] S. Kinsella, V. Murdock, and N. O'Hare. *"I'm eating a sandwich in Glasgow": modeling locations with tweets*. In Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents, pages 61–68, 2011.

[7] B. Wing and J. Baldridge. *Simple supervised document geolocation with geodesic grids*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 955–964, 2011.

[8] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. *Supervised text-based geolocation using language models on an adaptive grid*. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1500–1510, 2012.

[9] A. Popescu, G. Grefenstette, and P. A. Moëllic. *Gazetiki: automatic creation of a geographical gazetteer*. In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, pages 85–93, 2008.

[10] H. Manguinhas, B. Martins, and J. Borbinha. *A geo-temporal Web gazetteer integrating data from multiple sources*. In Proceedings of the 3rd International Conference on Digital Information Management, pages 146–153, 2008.

[11] P. D. Smart, C. B. Jones, and F. A. Twaroch. *Multi-source toponym data integration and mediation for a meta-gazetteer service*. In Proceedings of the 6th international conference on Geographic information science, GIScience'10, pages 234–248, Berlin, Heidelberg, 2010. Springer-Verlag.

[12] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. *Web-a-where: geotagging web content*. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 273–280, 2004.

[13] K. Q. Weinberger, M. Slaney, and R. Van Zwol. *Resolving tag ambiguity*. In Proceedings of the 16th ACM international conference on Multimedia, pages 111–120, 2008.

[14] M. D. Lieberman, H. Samet, and J. Sankaranayananan. *Geotagging: using proximity, sibling, and prominence clues to understand comma groups*. In Proceedings of the 6th Workshop on Geographic Information Retrieval, pages 6:1–6:8, 2010.

[15] R. Tobin, C. Grover, K. Byrne, J. Reid, and J. Walsh. *Evaluation of georeferencing*. In Proceedings of the 6th Workshop on Geographic Information Retrieval, pages 7:1–7:8, 2010.

[16] J. L. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, School of Informatics, University of Edinburgh, January 2007.

[17] C. B. Jones, R. S. Purves, P. D. Clough, and H. Joho. *Modelling vague places with knowledge from the Web*. Int. J. Geogr. Inf. Sci., 22:1045–1065, January 2008.

[18] T. Rattenbury, N. Good, and M. Naaman. *Towards automatic extraction of event and place semantics from flickr tags*. In Proceedings of the 30th Annual International ACM SIGIR Conference, pages 103–110, 2007.

[19] T. Rattenbury and M. Naaman. *Methods for extracting place semantics from Flickr tags*. ACM Transactions on the Web, 3(1):1–30, 2009.

[20] O. Van Laere, S. Schockaert, and B. Dhoedt. *Georeferencing Flickr photos using language models at different levels of granularity: An evidence based*

approach. Web Semantics: Science, Services and Agents on the World Wide Web, 2012.

[21] A. Rae and P. Kelm. *Working Notes for the Placing Task at MediaEval2012*. In Working Notes of the MediaEval Workshop. CEUR-WS.org, ISSN 1613-0073, online http://ceur-ws.org/Vol-927/mediaeval2012_submission_-6.pdf, 2012.

[22] C. De Rouck, O. Van Laere, S. Schockaert, and B. Dhoedt. *Georeferencing Wikipedia pages using language models from Flickr*. In Proceedings of the Terra Cognita 2011 Workshop, pages 3–10, 2011.

[23] C. Hauff and G.-J. Houben. *Placing images on the world map: a microblog-based enrichment approach*. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pages 691–700, 2012.

[24] O. Van Laere, S. Schockaert, and B. Dhoedt. *Georeferencing Flickr resources based on textual meta-data*. Accepted for publication in Information Sciences, Elsevier, February 2013.

[25] C. Hauff and G.-J. Houben. *WISTUD at MediaEval 2011: Placing Task*. In Working Notes of the MediaEval Workshop, Pisa, Italy, September 1-2, 2011. CEUR-WS.org, ISSN 1613-0073, online http://ceur-ws.org/Vol-807/Hauff_WISTUD_Placing_me11wn.pdf.

[26] C. Zhai and J. Lafferty. *A study of smoothing methods for language models applied to Ad Hoc information retrieval*. In Proceedings of the 24th Annual International ACM SIGIR Conference, pages 334–342, 2001.

[27] F. Krippner, G. Meier, J. Hartmann, and R. Knauf. *Placing Media Items Using the Xtrieval Framework*. In Working Notes of the MediaEval Workshop, Pisa, Italy, September 1-2, 2011. CEUR-WS.org, ISSN 1613-0073, online http://ceur-ws.org/Vol-807/Krippner_CUT_Placing_me11wn.pdf.

8

Conclusions and Perspectives

Stay hungry. Stay foolish.

– Back cover of the 1974 edition of The Whole Earth Catalog

Location-based information has become crucial for today's web applications and services. New geotagged content (i.e. content that has a geographical grounding) is uploaded to the Internet every day. This content has either been tagged automatically, for example by the device on which it was created, or it has been tagged manually by the user, for instance by clicking on a location on a map. To address the growing need for location-based data, research has focused on techniques to automatically geotag content that was uploaded without any assigned coordinate, by exploiting information from available geotagged content. In this dissertation, we focused on the task of georeferencing Flickr photos. Given the fact that there are over 200 million geotagged photos available on Flickr described by tags, we believe this data to be a potentially valuable source of geographical information. If sufficiently rich and accurate information is indeed (implicitly) contained in Flickr data, and if this information can be extracted, it can be exploited to automatically georeference other textual content that has no spatial grounding.

In order to find toponyms in text, the use of a gazetteer is widely adopted. However, when applied in the context of finding geographical entities in Flickr tags, it is less effective. First, the entities that are present in a gazetteer are mainly well-known administrative places, such as cities and towns. Second, there is the problem of vernacular place names: the actual, administrative names of places are often not used by people when referring to those places in social media. Third, the

coverage of a gazetteer is generally limited to a city scale. If the tags of a Flickr photo refer to specific local venues in a city or neighbourhoods, it is unlikely that one would be able to geolocate them using a gazetteer. Along with these three limitations, there is the problem of terms with ambiguous meanings: the limited context information (i.e. the available tags) for Flickr photos makes it hard to disambiguate among the different meanings of a given term using a gazetteer.

To overcome these shortcomings, researchers have found that using language modelling is more effective. This approach converts the problem into a classification problem, in which a language model is used to determine a single area from a disjoint set of areas (the classes) that is most likely to contain the location where the photo was taken. In our work, we have extended this classification approach with a second step. In this subsequent step, we search for training items within the designated area that are most similar to the photo we are trying to locate. Our experimental results show that the overall performance of the system significantly improves due to this second step. It is also this step that allows us to find locations at a sub-city scale. In our evaluation, we were able to reduce the median error, on a given test set, from 15.16 km to 9.23 km.

To obtain the set of areas used for classification, different methods have been proposed in literature of which the most straightforward one is to divide the surface of the earth using a geodesic grid. As an alternative, the mean-shift clustering procedure has been applied in related work as well. We have implemented both these algorithms and compared them with k-medoids. We have shown that k-medoids performs best at this task due to its tendency to produce smaller scale clusters in those areas of the world for which more training data is available.

When comparing the performance of a number of classical feature selection algorithms at the task of georeferencing Flickr photos, we see that Information Gain (IG) performs comparably to a simple heuristic that ranks the tags according to the number of times they occur. Similarly, Dunning's log likelihood measure performs substantially worse than χ^2, while the geospread measure from [1] was found to outperform all others. Using the same evaluation setup as mentioned before, the median error distance over the test set can be reduced from 9.23 km to 5.75 km using this feature selection algorithm. This clearly shows that classical feature selection algorithms fail to select spatially relevant features and suggests the need for new methods that take the spatial nature of the problem into account. To address this need, we have studied the use of kernel density based methods for selecting location-relevant tags from a collection of georeferenced Flickr photos, as well as the use of Ripley K's function from geographical epidemiology. In particular, we studied two spatial smoothing algorithms. The first method uses the divergence between the distribution of the occurrences of a single tag and the overall distribution. A second method uses the entropy value of the distribution of the occurrences of a single tag to measure the extent to which they occur in clusters around certain

points. In our evaluation, we have clearly demonstrated the necessity for good feature selection procedures as the results deteriorate significantly when all features are included in the language models.

A major drawback of existing georeferencing systems is that they will always estimate a precise location, even if there is not enough data available to make an informed decision (e.g. in case a Flickr photo has no tags associated to it). In this work, we proposed an approach that trains language models at multiple levels of granularity and automatically determines the appropriate level at which to georeference a photo. This level is selected adaptively and based on evidence available in the tags to support a decision within a given confidence threshold. During our evaluation, we have shown that this approach quantitatively outperforms the standard language modeling approach. If a certain confidence threshold is set (e.g. 95% accuracy) the system will only georeference those cases of which it is sufficiently confident that its prediction will be good and thus achieve a high accuracy. On the other hand, if the threshold is lowered, more photos will be annotated by the system and, unavoidably, more errors will be made. For instance, our approach using the pignistic probability manages to locate over 28 000 out of 50 000 photos with an accuracy of 95%, while the baseline approach can only achieve this accuracy for 17 000 of the photos.

The central hypothesis of this dissertation was that given a source of geotagged data (e.g. from social media, such as Flickr photos), we can extract models that enable automated geotagging of other textual content. To demonstrate this, we have focused on the use case of georeferencing Flickr photos. Since georeferencing Flickr photos is only useful in a limited number of use cases, we have looked at whether the language models trained using Flickr could be useful in a wider context. In particular, we have shown that a language model trained using data from Flickr is better at georeferencing Wikipedia documents than training a model on Wikipedia data itself. Also, when comparing this approach to a traditional, gazetteer-based method, it yields significantly better results. We obtain a median error distance of 2.18 km over the test collection of 21 839 Wikipedia pages, locating 35.03% of the documents within 1 km of their true location, whereas the gazetteer-based method achieves a median error distance of 28.9 km (locating 4.2% of the documents within 1 km). These findings confirm the potential of exploiting the geographical data that is implicitely present in the wisdom of the masses on social media.

To evaluate our methods on a large scale, we implemented a scalable georeferencing framework capable of handling generic sources of textual data, such as Flickr photos, Twitter messages or Wikipedia documents. The final version of our framework can handle over 64 million training examples and can be used to construct language models that contain over 200 000 classes in combination with 3 000 000 features. A model with 32M training photos, 137 000 classes and

1 000 000 features took only ∼11 hours, on a single 16-core computer, to train and evaluate 48 000 test documents. In most experiments however, the number of classes varies from 500 to 10 000 in combination with 1 000 000 features or less, for which training and evaluation takes between 2 to 15 minutes, depending on the configuration.

The georeferencing framework outlined in this PhD dissertation has been evaluated in the 2010, 2011 and 2012 editions of the MediaEval Placing Task benchmark, for which we received the *"quantum leap award"* in 2010 with a submission that substantially outperformed all other submissions to the task. Our framework is able to accurately assign a geographical coordinate to different sources of textual documents. This has experimentally been confirmed using tagged Flickr photos, Twitter messages, Getty Images photo captions and Wikipedia documents. In some of our evaluations, we have included a comparison to Yahoo! Placemaker as a baseline gazetteer approach, which was significantly outperformed in all of our tests.

We see a number of opportunities for future work. Although the research on georeferencing has advanced over the last couple of years, we have not seen a corresponding progress in applications that use this technology. Our adaptive, multilevel approach to georeferencing provides a good starting point for an application that can automatically and accurately geotag existing Flickr photos. Consider the example of [2], a web application that allows anyone to suggest a location for Flickr photos without a location. As our approach can be set to a certain confidence threshold, it can be configured to only suggest location for those photos in which the system is highly confident, leaving only the harder, ambiguous, cases for the human annotators. Moreover, as the amount of geotagged content increases, our models and thus the system will automatically become better and more accurate (in assessing which photos to process) over time.

Next, in our current approach, all features are weighted equally in the similarity search step. It is clear that not all available features associated with a Flickr photo have an equal importance. Research should be carried out to find similarity measures that better reflect this than the Jaccard measure.

Further developing on the idea of georeferencing Wikipedia documents, we envision the automated geotagging of news articles. Considering the number of news articles found in the archives of online newspapers, having a geographical grounding for each of them would open doors for some exciting new applications. For instance, if someone is buying a house and wants to retrieve all news articles related to burglary in the surroundings of a house that is for sale, he should only have to draw a polygon on a map to retrieve the search results.

Finally, there is the idea of extracting semantic information related to places from social media. Preliminary research on the check-ins of Foursquare users has been carried out in [3] to determine urban areas based on the activity of people in

it. The automated detection of Points-of-Interests (POI's) using social media has been investigated in [4], but can be taken a step further by automatically adding semantic information, such as the type of place. Preliminary research along these lines has been published in [5].

With respect to possible applications of the results of this PhD research, we envision a number of use cases. First of all, our models can be used to automatically geotag textual content, which can then in turn be used in applications that use this information, such as tourist guides that retrieve information based on location information (e.g. retrieve reviews of restaurants nearby). Next, our system can be used to measure the extent to which users of online media reveal their whereabouts by accident. One can think of an application that analyses fragments of text and reports the scale at which it can be located. A logical extension to such an application would be a component that suggests the removal of certain words that are a clear indication of geographical location in order to obfuscate the data (a process that is sometimes referred to as geo-cloaking). The most powerful application would certainly be if one could automatically geotag queries in search engines. This would enable a search provider to localize the results to a certain geographical scope, which would most likely yield high quality search results for the end user. On the other hand, this would also enable the search provider to better adjust advertisements to the geographical context, developing a better business companies would invest in.

References

[1] C. Hauff and G.-J. Houben. *WISTUD at MediaEval 2011: Placing task*. In Working Notes of the MediaEval Workshop, 2011.

[2] A. S. Cope. *Suggestify*. Available from: http://suggestify.appspot.com/ [cited Jan 18th, 2012].

[3] J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh. *The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City*. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 2012.

[4] A. Rae, V. Murdock, A. Popescu, and H. Bouchard. *Mining the web for points of interest*. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pages 711–720, 2012.

[5] S. Van Canneyt, S. Schockaert, O. Van Laere, and B. Dhoedt. *Detecting places of interest using social media*. To appear in Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence, 2012.