Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 11, November 2014, pg.73 – 80

RESEARCH ARTICLE

Kernel Based Tagging Method Using Spatial Paradigm

S.Dinesh[1], K.S.Kannan[2]
PG STUDENT, ASSISTANT PROFESSOR
NPR COLLEGE OF ENGINEERING AND TECHNOLOGY, TAMIL NADU, INDIA
EMAIL: spdc27@gmail.com, saikannan2012@gmail.com

Abstract: The aptitude to select those standings from a given assembly that are most telltale of geographic location is of key reputation in efficaciously addressing this task.This procedure of selecting spatially relevant terms is at present not well understood, and the popular of current systems are based on regular term selection methods.They propose two classes of term assortment methods based on standard geostatistical techniques.To implement the idea of spatial flattening of term existences, consider the use of kernel density estimation (KDE) to model each term as a two-dimensional possibility circulation over the surface of the Earth.Gazetteers have customarily been the main tool to assess the geographic scope of textual properties.Modalities in which geographical data's can be extracted.The nature of various modalities and lay out aspects that are estimated to govern the selections with respect to vision applications.The likeness between the two maps creating one jump to the supposition that the more populous regions frequently invite greater levels of photographic action.

Index Terms— Base Station, Ciphertext, Block Chaining (CBC), Concealed Data Aggregation (CDA), Data Aggregation, Wireless Sensor Networks (WSNs), Symmetric key encryption

1. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amounts of data stored either in data warehouses or other information repositories like database. Data mining has mostly get an important area for research. The term is actually defined misnomer. It is a non-trivial process of identifying valid, novel, potentially and ultimately understandable patterns in data mining. It can be view as the result of the natural evolution of information technology. The database has witnessed an evolutionary path in the development of the following functionalities. Processing of data collection and database creation and data management and advanced data analysis.

Data mining is a step in the process of knowledge discovery consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

The most common one is in the form of a large number of observations (cases). Real world applications usually are large in respect of the number of variables (dimensions) that are represented in the data set. Data mining is also concerned with this side of largeness. Especially in the field of bioinformatics, many data sets compromise only a small number of cases but large number of variables. Secondary analysis implies that the data can rarely be regarded as a random sample from the population of interest and may have quite large selection biases. The primary focus in investigating from a small sample to a large universe, but more likely partitioning the large sample into homogeneous subsets.

Geotagging is the process of recognizing textual references to geographic locations, known as toponyms, and resolving these references by assigning each lat/long values. Typical geotagging algorithms use a variety of heuristic evidence to select the correct interpretation for each toponym. A study is presented of one such heuristic which aids in recognizing and resolving lists of toponyms, referred to as comma groups. Comma groups of toponyms are recognized and resolved by inferring the common threads that bind them together, based on the toponyms' shared geographic attributes.

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 United Kingdom 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 resource, such as slang words, regional product names, and so on.

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., empire state building, Eiffel tower), events (e.g., 911, ironman), slang words, and so on. For each of these types of tags, the distribution of tag occurrences should deviate substantially from that of general tags such as birthday party 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. 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., empire state building, Eiffel tower),events (e.g., 911, ironman), slang words, and so on. For each of these types of tags, the distribution of tag occurrences should deviate substantially from that of general tags such as birthday party 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 abundance of geotagged resources give the possibility of easily extracting a geographical profile of terms. By extending the concept of event in the domain of the image-sharing application there is a need to define a method for employing this raw geographical data. The reason of doing this is the creation of a general retrieval model in order to improve the detection and search of event-related resources.

2. LITERATURE SURVEY

2.1.Georeferencing Wikipedia Pages Using Language Models from Flickr

The task of assigning geographic coordinates to web resourceshas recently gained in popularity. In particular, several recent initiativeshave focused on the use of language models for georeferencing Flickrphotos, with promising results. Such techniques, however, require theavailability of large numbers of spatially grounded training data. Theyare therefore not directly applicable for georeferencing other types of resources, such as Wikipedia pages. As an lternative, in this paper weexplore the idea of using language models that are trained on Flickrphotos for finding the coordinates of Wikipedia pages. Our experimentalresults show that the resulting method is able to outperform popularmethods that are based on gazetteer look-up.

The geographic scope of a web resource plays an increasingly important role forassessing its relevance in a given context, as can be witnessed by the popularityof location-based services on mobile devices. When uploading a photo to Flickr,for instance, users can explicitly add geographical coordinates to indicate whereit has been taken. Similarly, when posting messages on Twitter, information maybe added about the user's location at that time.

Our method, on the other hand, only usesinformation that was obtained from freely available, user-contributed data, inthe form of georeferenced Flickr photos, and uses standard language modelingtechniques.These results suggest that the implicit spatial information that arises fromthe tagging behavior of users may have a stronger role to play in the eld of geographic information retrieval, which is currently still dominated by gazetteerbased approaches. Moreover, as the number of georeferenced Flickr photos isconstantly increasing, the spatial models that could be derived are constantly improving. Further work is needed to compare the information contained implicitlyin such language models with the explicit information contained in gazetteers.

The reason isthat typically there is only one Wikipedia page about a given location, so eitherits location is already known or its location cannot be found by using othergeoreferenced pages. Moreover, due to the smaller number of georeferenced pages and the large number of spatiallyirrelevant terms on a typical Wikipedia page, the process further complicates.

2.2 Exploring Place through User Generated Content: Using Flickr to Describe City Cores

Terms used to describe city centers, such asDowntown, are key concepts in everydayor vernacular language. Here, we explore such language by harvesting georeferencedand tagged metadata associated with 8 million Flickr images and thus consider how largenumbers of people name city core areas. The nature of errors and imprecision in taggingand georeferencing are quantified, and automatically generated precisionmeasures appearto mirror errors in the ositioning of images. Users seek to ascribe appropriate semanticsto images, though bulk-uploading and bulk-tagging may introduce bias. Between 0.5–2%of tags associated with georeferenced images analyzed describe city core areas generically,while 70% of all georeferenced images analyzed include specific place name tags, withplace names at the granularity of city names being by far the most common.

Exploring commonsense notions requires that we have access to descriptions whichsomehow reflect vernacular usage. One data source with considerable potential in GIScienceis so called user generated content (UGC) or, perhaps more specifically, volunteeredgeographic information (VGI). Both UGC and VGI are essentially data uploaded to theweb by individuals, such as YouTube videos, Open Street Map data, or Flickr images andtags. References to location may be stored as (potentially) unambiguous coordinates orreferences to place names in the form of, for example, tags or titles, if users feel that locationis a useful (or practical) way of describing such an information object.

Chief amongst these tag properties appears to be the importance of both vernacularnames specific to individual cities, and more generally, place names, in describinggeoreferenced images. More than 70% of georeferenced images, and up to 35% of tags,included at least one place name tag of some granularity,

reflecting the overall importanceof place names in tagging behavior, and tags at the granularity of cities dominate.

Users appear to have added the same taglist to a whole set of photosuploaded simultaneously. Overall, the relationship between misplaced and mistaggeditems suggests that most users take tagging seriously, but not all of them are willing orable to correctly locate images on a map when georeferencing—in other words people arebetter at describing what they have seen in terms of semantics than they are at assigning anaccurate georeference.

2.3.A Latent Variable Model for Geographic Lexical Variation

The rapid growth of geotagged social mediaraises new computational possibilities for investigatinggeographic linguistic variation. Inthis paper, we present a multi-level generativemodel that reasons jointly about latent topicsand geographical regions. High-level topicssuch as "sports" or "entertainment" are rendereddifferently in each geographic region,revealing topic-specific regional distinctions.Applied to a new dataset of geotagged microblogs,our model recovers coherent topicsand their regional variants, while identifyinggeographic areas of linguistic consistency.In this paper, we present a method foridentifying geographically-aligned lexical variationdirectly from raw text. Our approach takes the formof a probabilistic graphical model capable of identifyingboth geographically-salient terms and coherentlinguistic communities.

One challenge in the study of lexical variation isthat term frequencies are influenced by a variety offactors, such as the topic of discourse. We addressthis issue by adding latent variables that allow us tomodel topical variation explicitly. We hypothesizethat geography and topic interact, as "pure" topicallexical distributions are corrupted by geographicalfactors; for example, a sports-related topic willbe rendered differently in New York and California.Each author is imbued with a latent "region"indicator, which both selects the regional variant ofeach topic, and generates the author's observed geographicallocation.

We develop a model that incorporates two sourcesof lexical variation: topic and geographical region.We treat the text and geographic locations as outputsfrom a generative process that incorporates bothtopics and regions as latent variables.6 During inference,we seek to recover the topics and regions thatbest explain the observed data.

We see this work as a first step towards a unsupervisedmethodology for modeling linguistic variationusing raw text. Indeed, in a study of morphosyntacticvariation, Szmrecsanyi finds that bythe most generous measure, geographical factors accountfor only 33% of the observed variation. Ouranalysis

might well improve if non-geographicalfactors were considered, including age, race, gender,income and whether a location is urban or rural.

We apply mean-field variational inference: a fullyfactoredvariational distribution Q is chosen to minimizethe Kullback-Leibler divergence from thetrue distribution. Mean-field variational inferencewith conjugate priors is described in detail elsewhere we restrict our focus to the issues that areunique to the geographic topic model.

2.4.You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users

Three of the key featuresof the proposed approach are: (i) its reliance purely on tweetcontent, meaning no need for user IP information, privatelogin information, or external knowledge bases; (ii) a classification component for automatically identifying words intweets with a strong local geo-scope; and (iii) a lattice-basedneighborhood smoothing model for refining a user's locationestimate. The system estimates k possible locations for eachuser in descending order of confidence. On average we findthat the location estimates converge quickly, placing 51% of Twitter users within 100miles of their actual location.

Mining this people-centric sensor data promises new personalized information services, including local news summarized from tweets of nearby Twitter users, the targeting of regional advertisements, spreading business information to local customers, and novel location-based applications (e.g., Twitter-based earthquake detection, which canbe faster than through traditional official channels).

To overcome this location sparsity problem, we proposein this paper to predict a user's location based purely onthe content of the user's tweets, even in the absence of anyother geospatial cues. Our intuition is that a user's tweetsmay encode some location-specific content { either specificplace names or certain words or phrases more likely to beassociated with certain locations than othersfor people from Texas). In this way, we can ll-the-gap forthe 74% of Twitter users lacking city-level granular locationinformation.

The content-based approach relies ontwo key refinements: (i) a classification component for automatically identifying words in tweets with a strong localgeo-scope; and (ii) a lattice-based neighborhood smoothingmodel for refining a user's location estimate. We have seenhow the location estimator can place 51% of Twitter userswithin 100 miles of their actual location.

To quantify the impact of an increasing amount of userinformation, we calculate the distribution of Error Distanceand the Average Error Distance across all of the test usersbased on the Local Word ltering location estimator relyingon a range of tweets from 100 to 1000.The error distance distribution, where each point representsthe fraction of users with an error in that range.

2.5.Kernel Density Estimation via Diffusion

We present a new adaptive kernel density estimator based onlinear diffusion processes. The proposed estimator builds on existingideas for adaptive smoothing by incorporating information from a pi-lot density estimate. In addition, we propose a new plug-in bandwidthselection method that is free from the arbitrary normal reference rulesused by existing methods. We present simulation examples in whichthe proposed approach outperforms existing methods in terms of accuracy and reliability.

The aim of this paper is to introduce an adaptive kernel density estimationmethod based on the smoothing properties of linear diffusion processes. Thekey idea is to view the kernel from which the estimator is constructed as thetransition density of a diffusion process. We utilize the most general lineardiffusion process that has a given limiting and stationary probability density. This stationary density is selected to be either a pilot density estimateor a density that the statistician believes represents the information aboutthe data prior to observing the available empirical data. The approach leadsto a simple and intuitive kernel estimator with substantially reduced asymptotic bias and mean square error. The proposed estimator deals well withboundary bias and, unlike other proposals, is always a bona fide probabilitydensity function.

The resulting diffusion estimator unifies many of the existing ideas about adaptivesmoothing. In addition, the estimator is consistent at boundaries. Numerical experiments suggest good practical performance. As future research, the

proposed estimator can be extended in a number of ways. First, we can construct kernel density estimators based on L´evy processes, which will havethe diffusion estimator as a special case. The kernels constructed via a L´evyprocess could be tailored for data for which smoothing with the Gaussiankernel density estimator or diffusion estimator is not optimal. Such casesarise when the data is a sample from a heavy-tailed distribution. Second,more subtle and interesting smoothing models can be constructed by considering nonlinear parabolic PDEs.

The paper introduces an improved plug-in bandwidth selection method that completely avoids the normal reference rules that haveadversely affected the performance of plug-in methods. The new plug-inmethod is thus genuinely "nonparametric," since it does not require a preliminary normal model for the data. Moreover, our plug-in approach doesnot involve numerical optimization and is not much slower than computinga normal reference rule.

3. CONCLUSION

The geo tagging and time-tagging features of QDA Miner are part of its Hyper linking capability so that, in order to attach geographic coordinates or a time stamp to a piece of qualitative information, you must first create a hyperlink. Hyper linking is typically conceived as a way of creating links that allow one to jump to another document or another location in the same document or to access external resources such as a web page or another computer file. In QDA Miner, hyperlinks are also used to store geographic and time coordinates. In the case of geographic information, one may then use such a hyperlink to quickly jump to a mapping tool like Google Earth and view the exact location corresponding to the geographic coordinates. This section will show you how to create those links.

It could not be evaluated if a different weighting of the two (three) dimensions could improve retrieval performance. Connected to the latter point is the question of how many images per dimension should be retrieved and what the appropriate thresholds for retrieval scores (e.g. no images retrieved with a similarity score below 0.5) are. Additionally, different combinational strategies (different Comb strategies, Borda fusion instead of Comb) may prove to provide more relevant retrieval results.

The applications of geographic information in conjunction with roughly categorized as location recognition, object or event recognition, visualization, recommendation, social networking, and mapping. More generally, many of the applications are directed at helping a computer understand one or more15images, based on known relationships that record which objects are likely to be where in the world. Other applications consolidate a large-scale dataset of geo-tagged information to produce maps that indicate where things are in the world.

REFERENCES

[1] P. Serdyukov, V. Murdock, and R. van Zwol, "Placing FlickrPhotos on a Map," Proc. 32nd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 484-491, 2009.

[2] C. De Rouck, O. Van Laere, S. Schockaert, and B. Dhoedt,"Georeferencing Wikipedia Pages Using Language Models fromFlickr," Proc. the Terra Cognita 2011 Workshop, 2011.

[3] T. Rattenbury, N. Good, and M. Naaman, "Towards AutomaticExtraction of Event and Place Semantics from Flickr Tags," Proc.30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 103-110, 2007.

[4] L. Hollenstein and R. Purves, "Exploring Place through User-Generated Content: Using Flickr to Describe City Cores," J. SpatialInformation Science, vol. 1, no. 1, pp. 21-48, 2010.

[5] A. Popescu and G. Grefenstette, "Deducing Trip RelatedInformation from Flickr," Proc. the 18th Int'l Conf. World WideWeb, pp. 1183-1184, 2009.

[6] J. Eisenstein, B. O'Connor, N.A. Smith, and E.P. Xing, "A LatentVariable Model for Geographic Lexical Variation," Proc. Conf.Empirical Methods in Natural Language Processing, pp. 1277-1287,2010.

