Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Pattern Based Geographical Indexing of Text Documents

[1]Anurag Soni*, [2]Diamond Jonawal
[1]M.Tech (CTA) Research Scholar
[1, 2] RGTU University, Bhopal, India

Abstract— As the internet users are increasing day by day with large amount of digital data. Geographical based tagging of their digital data such as audio; video, images is new area of exploration for the researchers. Geographical information is attached with the data in the form of longitude and latitude values. In this work, textual data is geo Indexed with the flicker website image tags. Most of the previously done work is done on the basis of term based approach, but this work introduces pattern based geo-indexing of the textual data with the help of Ripley's method. Results show that proposed work of pattern based approach is more efficient as compared to the previous researches.

Keywords - Disputant classification, document analysis, decision support systems, text mining.

I. INTRODUCTION

With the increase in the digital data on the servers, different types of storage schemes are developed. Conventional database consist of information storage in the form of attribute values whereas in spatial database some of the geographical information is attached. So, searching can be done on the basis of geographical location using name of the country, state, city, town, etc. Different dimensions can be used for searching of information. This include new field of searching or re-ranking where spatial based query is introduced and searching is done on the basis of spatial features of the data.

Execution of these spatial queries requires appropriate construction of data structure. Geotagging is the process of adding geographical identification metadata to various media. The metadata usually consists of latitude and longitude coordinates and other location–related information, such as the geocoded place names, data sources, etc. Geotagging is used primarily with photos; and has been applied to other media such as videos. Various ways to extend geotagging to text documents have been researched, proposed and demonstrated with little application development having occurred. This project demonstrates a method to apply geotagging and visualization of geotags to e-Book documents.

Geotagging of text presents problems related to scanning for possible geographical references (*placenames* or *toponyms*) located within a body of text, the disambiguation of the candidate placenames that were found, and the geocoding of the disambiguated placenames. Typically, annotations containing metadata resulting from the geotagging process are inserted into the text body, and appear as hyperlinks when the text is viewed. When a geotag is clicked, a web mapping or visualization service displays the geocoded location along with the metadata that is included in the geotags.

Within the annotation process, the geotagger offers an automated and interactive method of visualizing the geographic references via Google Maps and saving the metadata derived from the maps in the annotations. Geotag annotations are stored within KML files that can be shared and displayed on Google My Places, emailed or transferred to the other sites such as Facebook. Geotagging of a body of text adds spatial context to its narrative. With geotagging, places and geography can be visualized, adding spatial context to a novel or text book. It is a valuable research tool and a way to explore the cultural history of a place through the linkage of literature and location.

One of the wide range of applications of the text mining is analysis of the documents for the natural language processing (NLP) that whether the document contains information of which category. This is a kind of separation of the document from one category to other. This paper is focused on developing a system where each document in the dataset or input document can be found and then decide the geographical coordinates for that document. This decision is done on the basis of the text present in the article.

II. RELATED WORK

We, in this paper have used pattern based approach for boosting up the performance of the term based method [2]. As metadata created by the Pattern based approach is much effective then term based approach. A sequential pattern has developed from the paragraph, which is sensible as compare to terms as shown in mathematics and a [1] PTM (Pattern Taxonomy Model) is developed, in this model closed sequential terms are combined to form patterns and this model is adopted to classify document.

Different types of text mining techniques are developed in previous works, out of those bags of words (BOW) play very important role. In BOW, words are arranged in the form of keywords and some time their number of repetition are find in terms of TFIDF value and is also store as in [4] and improves performance.

Various words weighting scheme has explained in [2], the BOW approach helps in filtering the important keyword from the vast set of documents. In [1] Term based approach is use for developing the Meta data for mining, this help in understanding the matter. This is done in [3] where synonymous relation is developed among words. Furthermore, the

TagMaps TF-IDF method proposed in [6] reflects an idea for spatially aware term ranking for geotagging twitter messages.

In [8], a generative probabilistic model is used to determine words with a geographic scope within a tweet, and a form of neighbourhood smoothing is employed to refine the estimations. Geotagging of a general web pages is mostly gazetteer based (e.g., [5]), in which case only toponyms are considered.

Kernel density estimation [9] is a popular technique for analysing geographic point data. 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.

III. PROPOSED WORK

In this work, documents are clustered into specific area using geo tag information of the different areas. Figure 1 represents the steps of proposed work.

Step 1: Geo-Tag Dataset Pre-Processing

In this step dataset is taken which contain geographical co-ordinates and textual tags with those coordinates. As there are many other information present in the dataset which need to be filtered out from the co-ordinates and tags. Now consider one log of the dataset shown below, out of various information available in the log, this work requires only two, first is LABEL which provides associated tags, other is COORDINATES which provides longitude and latitude values of those tags.

"{"type": "Feature","id": 385,"properties": {"woe_id": 385, "place_id": "lBQva7.aCZk","place_type": "county","place_type_id": 9,"label": "Hastings, Ontario, Canada","superseded_by": 29375163,},"geometry": {"type": "MultiPolygon", "created": 1292490545,"alpha": 0.0015,"points": 10,"edges": 10,"is_donuthole": 0,"link": {"href": http://farm6.static.flickr.com/5282/shapefiles/385_20101216_6d2600f793.tar.gz",},"bbox":[77.657371520996,44.47831 7260742,-77.562705993652,44.541423797607], "coordinates": [[[[-77.619949,44.478317], [-77.636887,44.494568], [-77.657372,44.541424], [-77.625580,44.497578], [-77.588814,44.496891], [-77.562706,44.515480], [-77.568336,44.493332], [-77.572853,44.493973], [-77.583015,44.495113], [-77.619949,44.478317]]]]}}

$$C[n] = Geo\_Pre\_Process(GD)$$

So two vectors are created, first is for L[n] label and the other is for coordinates C[n] after pre-processing, where n represent number of logs.

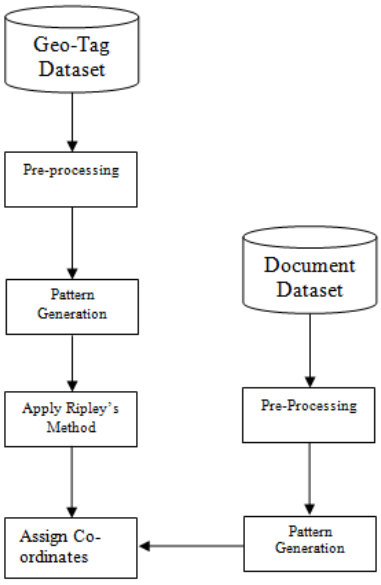

Fig 1: Block diagram of Proposed Work

Step 2: Pattern Generation

In this step patterns are developed from the L[n] vector. In order to find patterns from the vector, similarity between the labels is to be found. This can be understand as let the L[1] is {a, b, g}, L[2] is {c, a, b} then element {a, b} is common between both L[1] and L[2] labels and it is considered as the pattern. So each label number of comparison is evaluated by factorial n times or !n.

Step 3: Ripley's K-Method

In this step each pattern generated from the labels is ranked by using Ripley's k method. Here some sort of distance need to be taken as input for the method. Formula for Ripley's value evaluation for terms score is:

$$S(t) = \frac{\{n(p,q) \mid p,q \in Q_t, d(p,q) < \lambda\}}{N_t^2}$$

In above formula, p and q are patterns from the Q_t pattern set.

D (p, q) is formula for estimating distance between tags position as per geo tags.

λ is the distance threshold

N_t is number of terms patterns present.

Step 4: Document Pre- Processing

Pre-Processing: Text pre-processing consists of words which are responsible for lowering the performance of learning models. Data pre-processing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination, natural language specific stop word elimination and stemming. Stop-words are functional words which occur frequently in the language of the text (for example a, an, the, of etc. in English language). So, they are not useful for classification. Here, we read whole project and put all words in the vector. Now again read the file which contains stop words and then remove similar words from the vector. Once the data is pre-processed then it will become the collection of the words that may be in the vector. For example let one document is taken and its text vector is Rd[] ={a1, f1, s1, a2, s2, a3, a4, f2……..an} and let the stop words collection is S[] = {a1,a2,a3,…….am}. Then the vector obtained after pre-processing is

D[] = {f1, s1, s2, f2,…….fx}.

D[] = Rd[] – D[]

For Example: Rd[] = {'Every',' morning', 'Ram',' study',' for',' two',' hours',' and',' during',' this',' time',' his',' mother',' give',' him',' one',' glass',' milk',' with','bread','jam','in ','breakfast'}

After pre-processing

Now D[] = {'Ram',' hour',' time',' glass',' milk','bread','jam','breakfast'}

Proposed Algorithm

Input: Document DD, Geo_dataset GD

Output: Latitude Lat, Longitude Log

1. GD←Geo_Pre_Process(GD)
2. [Pat_G Geo]←Pattern_generation(GD)
3. R←Ripley_method(PD)
4. DD←Document_Pre_Process(DD)
5. Pat_D←Pattern_generation(DD)
6. Loop 1:n // Number of Pattern in Geo Dataset
7. Loop 1:m // Number of Pattern in Document Dataset
8. If Pat_D[n] = Pat_G[m]
9. Counter[n,m]= Counter[n,m]+1
10. End If
11. End Loop
12. End Loop
13. [Log Lat] ←Assign(Counter, Geo)

IV. EXPERIMENT AND RESULT

Experimental Setup

This section presents the experimental evaluation of the proposed work. All algorithms and utility measures were implemented using the MATLAB tool. The tests were performed on a 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM and running under Windows 7 Professional.

Dataset

Here two data sets have been used. First is geo_coordinates_en geotagged dataset containing flicker image tags. While other is documents used for the tagging purpose, these documents contain information about the spatial place. http://downloads.dbpedia.org/3.7/en/geo_coordinates_en.nt.bz2.

Evaluation Parameter

In order to evaluate results, mean error distance (MED) estimation is required. MED is obtained by finding the geographical distance between two points in terms of longitude and latitude. Obtained values can be put in the mentioned parameter formula to get better results.

Radian_lat = (Lat2 – Lat1)* π/180 // convert degree to radian

Radian_long = (Long2 – Long1)* π/180

Radian_lat1 = Lat1* π/180

Radian_long2 = Long2 * π/180

MED = R*[sin (Radian_lat/2)^2 + sin (Radian_long/2)^2 + cos (Radian_lat1) * cos (Radian_lat2)]

Where R is radius of earth, {Lat1, Long1}, {Lat2, Long2} are coordinates in degree.

Results:

Document tagging is done on the basis of proposed pattern based approach and term based previous work in [10] RTM.

Table 1

Distance between Actual and Proposed Geo Position in miles		
Document	Proposed Work	Previous Work
D1	1.8070	70.5618
D2	2.6287	73.4459
D3	3. 734	63.4459

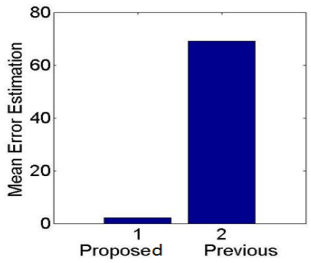

Fig: Graph of Mean Error Estimation at constant grid

Above results shows that as the use of proper threshold of the pattern selection and dictionary it is possible to have values of of MED is 1.807 which is quite good progress done by the proposed algorithm as compare to the previous work in [10], where most of the values are below the average of the results obtained.

V. CONCLUSION AND FUTURE WORK

In this paper it is obtained that a remarkable improvement is done by the proposed work for the identification of the geo coordinates without having any kind of background knowledge or supervised learning. This proposed work shows that the testing produce more effective results from the previous one where 1.807 is the MED obtain. So with the continous updation of the tags this can produce more effective results. There is plenty of work is required to do in this field where one can apply its algorithm such as in different other language as the processing will change most of the steps.

REFERENCES

[1] D.A. Schon and M. Rien, Frame Reflection: Toward the Resolution of Intractable Policy Controversies. Basic Books, 1994.

[2] S. Somasundaran and J. Wiebe, "Recognizing Stances in Ideological Online Debates," Proc. NAACL HLT Workshop Computational Approaches Analysis and Generation Emotion in Text (CAAGET '10), pp. 116-124, 2010.

[3] M. Wasson, "Using leading text for news summaries: Evaluation results and implications for commercial summarization applications" In Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the ACL. Pp.1364-1368. 1998.

[4] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu "Effective Pattern Discovery for Text Mining". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012

[5] [23] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-Where: Geotagging Web Content," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 273-280, 2004.

[6] T. Rattenbury and M. Naaman, "Methods for Extracting Place Semantics from Flickr Tags," ACM Trans. Web, vol. 3, no. 1, pp. 1- 30, 2009.

[7] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu. "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection". IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 42, NO. 3, MAY 2012.

[8] [22]L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, "Spatial Variation in Search Engine Queries," Proc. 17th Int'l Conf. World Wide Web, pp. 357-366, 2008.

[9] [10] B. Silverman, Density Estimation for Statistics and Data Analysis. Chapman & Hall, 1986.

[10] Olivier Van Laere, Jonathan Quinn, Steven Schockaert, and Bart Dhoedt, " Spatially Aware Term Selection for Geotagging". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014

