IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

Spatiotemporal Keyword Query Suggestion Based On Document Proximity and K-Means Method– A Review

Aju Tom Kuriakose[1], Sobhana N.V[2]

Department of Computer Science and Engineering, Rajiv Gandhi Institute of Technology, Kottayam, Kerala, India[1,2]

Abstract: The commercial web search engines are looking for the efficient keyword suggestions methods for retrieving the relevant information. The submission of the right keyword query determines the right direction of the user's search. Spatial proximity and query results are the crucial factors in the keyword suggestion techniques. The Spatiotemporal suggestion of queries deals with spatial proximity.On the basis of a weighted keyword document graph that maps the relevance of keyword queries semantically with the distance between the location of the user and the documents retrieved. The K-Means method used for retrieving the highest ranked top k objects near to the current location of the user and Time aware query suggestion brings out the most relevant documents based on the temporal proximity. The word sense disambiguation gives an added advantage of getting the local data without the location name ambiguity.

Keywords: Query Suggestions, query optimization, spatial proximity, K-Means Method, Word Sense Disambiguation.

I. INTRODUCTION

The Keyword query suggestion is the most relevant feature of a commercial web search engine. The user may not be satisfied with the search results. It's likely to refine the queries for the search operation by the keyword suggestion module for regulating the user's search towards the desired results. Query session data and click information from query logs makes the keyword suggestion more effective. The location-aware keyword query suggestion helps in retrieving not only the documents related to the user's information needs but also retrieves relevant documents near to the user's location.

The spatiotemporal keyword query suggestion based on document proximity and k-means method have a location aware keyword query suggestion framework based on the weighted keyword- document graph that maps the relevance of the keyword query semantically with user location and spatial distance with the resulting documents. The graph is traversed in a random-walk-with-restart fashion, to select the highest scored keyword queries as suggestions.

A partition-based approach which gives more efficient outcome than the baseline algorithm is used to make the system more scalable. The k- means method retrieves the objects near to the user location and in addition, a time aware query suggestion helps in retrieving the relevant query suggestions in the temporal proximity.

The ambiguity in the location names can be solved by creating place name ontology. The word sense disambiguation solved using a method of ontology generation for Indian place names helps in solving the ambiguity.

II. LITERATURE REVIEW

Shuyao Qi et al. [1] proposed the keyword query suggestion techniques based on the location of the user and the semantic relevance of the document. The spatial proximity factor and the search results are not considered in the existing keyword suggestion techniques. In this paper, a Location-aware keyword query suggestion framework is proposed with challenges of efficiently measuring the keyword query similarity while capturing the spatial distance factor and efficiently computing suggestions. It divides the keyword queries and the documents into partitions and adopts a lazy mechanism that accelerates RWR search in large graphs

In Keyword Query Routing, Thanh Tran et al. [2] introduces a method to route keywords to relevant sources for reducing the high cost of processing for keyword search queries in the sources. Keyword search in this is an intuitive paradigm for searching the linked data sources on the web. They propose a novel method for computing top-k routing plans based on their potentials to contain results for given keyword query. The relevance of routing plans based on the scores at the level of keywords, data elements and the subgraphs that connect these elements are computed using a multilevel scoring mechanism. The routing greatly helps in improving the performance of keyword search without compromising the resulting quality.

T. Miyanishi et al. [3] proposed the Time-aware Structures Query Suggestion (TaSQS) that clusters the query suggestions along with a timeline so as to narrow down the search information from a temporal point of view. Moreover, when a suggested query is being clicked, TaSQS presents the results from bipartite graphs with

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

query-URL and after ranking it accordingly to the click counts within a particular time period. The experiments use data from the commercial search engine log. The time aware clustering and the time aware document ranking features of TaSQS effectively selects the right document and query.

Yang song et al. [4] proposedQuery suggestion by constructing term-transition graphs. It wasa framework for query suggestion that leverages user re-query feedbacks from the search engine log data. They mined user activities where the user modifies a part of the query by adding terms after the initial query, deleting terms from a query or even modifying the query with new terms. A term transition graph based on mined data is formed. Two models with topic level and term level query suggestion were proposed. In the unsupervised Pagerank model based on the topic, it performs a random walk on term-transition graph based on the topics and calculates the Pagerank for each term within a topic. In second model term modifications are treated as documents so that each query reconstructed is treated as an instance for the training. The proposed model capable of suggesting new queries based on the initial search terms.

The subsequent combination of geo-location and documents empowers new sorts of inquiries that check both area vicinity and content significance. Dingming Wu et al. [5] proposes another indexing structure for top-k spatial content recovery. The framework leverages the inverted file for content recovery and the R-tree for spatial vicinity questioning. A few indexing methodologies are investigated inside this structure. The system includes calculations that use the proposed lists for processing area mindful and also district mindful top-k content recovery inquiries, along these lines checking both content pertinence and spatial vicinity to prune the hunting space. Consequences of observational studies with an execution of the system show that the proposition is fit for excellent performance.

To fulfill search necessities on ROIs (regions of interest), J. Fan et al. [6] concentrate on another examination issue, called spatial-textual similarity search: Given an arrangement of ROIs and an inquiry ROI, they locate the comparative ROIs by considering spatial cover and printed closeness. They present a filter and verification framework to register the answers. In the filter step, it creates marks for the ROIs and the question. In the verification step, it confirms the applicants and distinguishes the last replies. Test results on genuine and engineered datasets demonstrate that our strategy accomplishes high performance.

Flickr is one of the largest online image collections, where the photos are being shared and the photos are tagged with the geographical coordinates and other information. The tags help in getting a relationship between the image and the keyword being searched from the description that is included in the tags. I Miliou et al [7] discusses the gain of using spatial and textual information in order to recommend more meaningful tags to users. The geotagged properties of the images or the photos in flicker are analyzed and a novel tag recommendation system based on the location is being proposed. For evaluation purposes, they have implemented a prototype system and exploit it to present examples that demonstrate the effectiveness of proposed methods.

L. Chen et al [8] proposed an inside and out study of 12 cutting edge geo-textual indices. They propose a benchmark that empowers the correlation of the spatial-keyword query performance. They additionally investigate the discoveries acquired while applying the benchmark to the indices, hence revealing new bits of knowledge that may control index determination and in addition further research.

Hasna T.P. et al [9] proposed a methodology to recover m nearest objects fulfilling certain client determined conditions. Nearest keyword search (Keyword cover search) is to query objects, which together get the query keywords and the between objects separation is least. In keyword cover search client's current location is not indicated. Location aware nearest keyword search in spatial data can conquer this issue by giving client's momentum location as info and lead to such a location. Traveling distance is considered to create a path, rather than Euclidean distance. Additionally, the outcomes recovered in keyword spread search may not be constantly attractive to client's decision. This basic disadvantage can be overcome by nearest keyword search based location of the client.

Yunjun et al [10] discuss the problem of k-optimal-location-selection (k-OLS) [10] retrieval in metric spaces. The existing methods are not sufficient because they consider the Euclidean space, and are not sensitive to k. In this paper, they present an efficient algorithm for query processing in metric spaces. The solution implements a metric index structure on the datasets and it enables several pruning rules, it makes use of the reusability of the techniques used and the optimal score calculation, for supporting different types of data. In addition, they extend techniques to tackle two interesting and useful variants, namely, MkOLS [10] queries with multiple or no constrained regions. Extensive experimental evaluation using both real and experimental data demonstrates the pruning rules that are being presented and they are used for to evaluate the performance of the proposed algorithms. A hierarchical agglomerative clustering technique to find the clusters of similar queries and URLs in a query log using the click through data proposed by Beeferman and Berger [11]. The Query – URL bipartite graph that is constructed iteratively constructs the clusters by choosing pairs of most similar queries and URLs. This method is content ignorant and quite expensive.

The query suggestion system proposed by R. Baeza-Yates et al [12], uses a query clustering approach that works on the K-Means clustering algorithm. The recommendation system was kept updated always, so need not be reconstructed from scratch. The aging effect is not an issue as the incremental algorithm is used. Specifying the value

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

of k was the main difficulty in applying the query clustering.

Jeonghee Yi and Farzin Maghoul [13], proposed a click through the graph that considers the clicked page and the query relation. The clustering the given query is being done based on the query and the clicked page relationship. The syntactic and the semantic features of the parameters are considered. Many potentially interesting clusters of queries are excluded as they slightly violate the requirements of complete connectedness by the algorithm.

Yuan Hung et al [14] proposed a Top – K search results for the query clustering based on the similarity of the ranked URL results returned by the search engine and the query is used for the clustering. It has a better scalability feature apart from other query suggestion systems.

Kajal Y. Vyas [15], improved the web search rank based optimization using the search engine query logs. It helped in reducing the user's efforts and the seeking time. The challenge that was faced by this method was specifying the number of clusters. It also has a better scalability feature.

The spatial information needed by the users are often recognized within the document text data and the place to which it is referred to. A new method called gazetteer is proposed by Ivre Marjorie R et al [16]. It is a dictionary that defines the toponymic meaning that is expanded from the list of place names and it includes the geographical features such as the spatial correlation and terms related to a place. An OntoGazatteer [16] that record the connections of places name approach is proposed. The factual and the semantic support provided by the ontological gazetteer solve several common problems in geographic information retrieval. They present the OntoGazatteer and demonstrates its possible applications to a place name disambiguation problem. Also, they present a case study on recognizing and disambiguation in place names within news sources.

In geographical information retrieval, the main aim is to improve the accuracy using the place name disambiguation. The disambiguation task becomes more complex when the given input is too short. Wikipedia, providing the information related to places uses the named entity recognition. The semantic web version of Wikipedia known as the DBpedia[17] provides the well-formed structured, understandable mined information from the Wikipedia articles. In this, Yingjie Hu et al [17] propose a method of combined Wikipedia and DBpedia for disambiguating place names in short texts. The combined method is argued to have a better performance than the methods done alone. The method is generic and it can be made extended to other structured data sets such as Freebase or Wikidata [17].

Takashi Aawamura et al [18] proposed the social network information extracted from the social network sites such as Facebook and Twitter as it is useful and it contains a huge amount of location-specific information. Extracting information are necessary for identifying the location based on features that are embedded in a document. The word sense disambiguation helps in solving the location disambiguation, but it doesn't make use of the location-specific clues [18]. The location-specific clues are being considered in this paper are spatial proximity and the temporal consistency. The effectiveness of these clues is confirmed with experiments on Twitter tweets with GPS information.

Classifying the reviews of the customer on the basis of textual analysis with association rules using ontology is proposed by A. Razia Sultana et al [19]. An Ontology with improved k-means algorithm consolidates the reviews and handle the reviews that denote at least one drawback of the product that is being analyzed. The accuracy level of the recommendation system increased considerably with the use of modified K-Means algorithm. Combined pre-processing and ontology improves the classification accuracy of the customer reviews.

SK Ahammad Fahad et al [20] proposed a k-mean clustering method for improving the clustering in big data. They proposed a method that makes the system less time consuming, more effective and efficient with reduced complexity. The selection of the initial centroid is a crucial factor that determines the resulting clusters in the iterations. In this method, they initially find the centroid and set an interval between the elements so that it will not change their cluster during the iterations. This reduced the workload significantly for very large data sets.

Udaya Raj Dhungana et al [21] proposed a method for disambiguating the correct sense of polysemy word based on the clue word based on the WordNet. Polysemy words and the single sense words are referred as clue words for the related words. The nouns, verbs, adjectives and adverbs organized from conventional WordNet are put in together as synonyms called synsets. The different senses of polysemy word along with the single sense words based on the clue words form a new model of the WordNet. These clue words are used for the disambiguation process and the correct meaning of the polysemy words in the context are got using the Knowledge-based WSD algorithms.

Clayton Fink et al [22] proposed a method for finding the place name that is mentioned in a blog for to determine the author's location. Many of the automated geolocation strategies that use the domain name or the IP address are not enough for exactly finding the author's location. The proposed method gave an accuracy of 63% with a collection of 844 blogs with a known location.

The pieces of information that are spread out in a text are linked together using entities like names within it. The names can be ambiguous and it makes the systems more difficult to understand the names that are syntactically identical are actually the same in their semantic meanings too. Joachim Kleb and Raphael Volz [23] proposed a new approach for disambiguating the names using ontologies. They also make use of natural language text patterns and ranking algorithms for disambiguation. The approach was tested with geographic names and the use of it with a geographic news reader.

The web documents and the user's need for the web documents are increasing day by day and the ambiguity of the person names in the web documents make the search more difficult. The correct identification of a person name entity from a web document is challenging problem. Zhao Lu et al. proposed an Ontology-based approach for Personal Name Disambiguation (named "OnPerDis") [24]. The process has main two steps: initially, they construct person ontology that contains rich conceptual modeling and a large set of instances for the support. They create a temporary instance extract features from the web documents for a given personal name on the web as the second process. The similarity score is then calculated between temporary instance and the person ontology instance. The instance with high similarity score is selected as the best-suited person name. Evaluations were done with two rich real-life datasets for to get a better efficiency.

Padmamala R. [25] proposed a word-level translation scheme for translating the Tamil words to English. This method also shows the word sense disambiguation need in the translation for to arrive at a contextual or a meaningful word when different categories can be applied for to a single word syntactically. The semantic features of the words are analyzed using an ontology that is derived from the sub-categorical features. This ontology-based information retrieval helps in assigning correct meaning for the ambiguous words. A word sense disambiguator and a rule-based syntactic parser have been developed and it was tested on Tamil newspaper websites and the search results were remarkable.

Sachio Hirokawa et al. [26] proposed a system for getting the place names from a blog or a web document. The consumer- generated media as published on the web quotes their individual opinions and experiences. Without knowing the exact place name we cannot correlate the place with the information needed in the blogs. In this method, they propose a hierarchical structure for such operations for getting place names. They used 45553 blog articles about Karatsu area in Saga Prefecture and they related 78 potential place names which have not been appeared in the blogs were extracted. Meaningfully related words were evaluated to an accuracy of 80%.

III.CONCLUSIONS

The methods that are used by the query suggestion systems are discussed and the partition based method of keyword query suggestion using the bipartite graph is concluded as the best method since it can outperform many of the baseline methods used conventionally for the query suggestion. The modified K-Means clustering method helps in getting the top K results out of the suggested queries and the word sense disambiguation for the place names can be solved using the ontology that is created with the location names.The word sense disambiguation for the natural language input such as place name is difficult because it can cause challenges with its semantic evaluation with the ontology. The retrieval of location names from the online web documents is challenging as many users won't reveal their locations due to privacy issues, hence the location has to be parsed and retrieved from the web documents.

REFERENCES

[1] Shuyao Qi, Dingming Wu, and Nikos Mamoulis, “Location Aware Keyword Query Suggestion based on Document Proximity”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 28, NO. 1, JANUARY 2016

[2] Thanh Tran and Lei Zhang, “Keyword Query Routing”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 2, FEBRUARY 2014

[3] T. Miyanishi and T. Sakai, “Time-aware structured query suggestion,” in SIGIR, 2013, pp. 809–812.

[4] Yang Song, Dengyong Zhou, Li-Wei He, “Query suggestion by constructing term-transition graphs,” in WSDM, 2012, pp. 353–362.

[5] Dingming Wu, Gao Cong, Christian S. Jensen, “A framework for efficient spatial web object retrieval”, VLDB J., vol. 21, no. 6, pp. 797–822, 2012.

[6] Ju Fan, Guoliang Li, Lizhu Zhou, Shanshan Chen, and Jun Hu, “Seal- spatial-textual similarity search", proceedings of VLDB Endowment, vol. 5, no. 9, pp. 824–835, 2012.

[7] I. Miliou and A. Vlachou, “Location-aware tag recommendations for Flickr”, in DEXA, 2014, pp. 97–104.

[8] L. Chen, G. Cong, C. S. Jensen, and D. Wu, “Spatial keyword query processing: an experimental evaluation”, Proc. VLDB Endowment, vol. 6, pp. 217–228, 2013.

[9] Hasna T.P., Syed Farook K., “An Enhanced Location Aware Closest Keyword Search in Spatial Data”, International Journal of Current Trends in Engineering & Research (IJCTER) e-ISSN 2455–1392 Volume 2 Issue 4, April 2016

[10] YunjunGao, Shuyao Qi, Lu Chen, BaihuaZheng, Xinhua Li, “On efficient k-optimal-location-selection query processing in metric spaces”, 0020-0255/_ 2014 Elsevier Inc. All rights reserved Y., Gao et al. / Information Sciences 298 (2015) 98–117, http://dx.doi.org/10.1016/j.ins.2014.11.038

[11] D. Beeferman and A. Berger. “Agglomerative clustering of a search engine query log” In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, pages 407 – 416. ACM Press, August 2000

[12] R. Baeza-Yates, C. Hurtado, and M. Mendoza, “Query recommendation using query logs in search engines,” in EDBT, 2004

[13] J. Yi and F. Maghoul, “Query Clustering Using Click through Graph,” Proc. the 18th Int’l Conf. World Wide Web (WWW ’09), 2009

[14] Yuan Hong, Jaideep, Vaidya and Haibing Lu Rutgers University "Search engine query clustering using Top-K Search Results", 2011 IEEE/WIC/ACM International Conferences on Web intelligence and intelligent Agent technology, DOI10.1109/WI-IAT.2011.224

[15] Kajal Y.VYAS, "Improved web search result rank optimization using search engine query log" Journal of information knowledge and research in Computer Engineering ISSN:0975-6760 2012 volume-02, ISSUE-02

[16] Ivre Marjorie R. Machado, Rafael Odon de Alencar, Roberto de Oliveira Campos Jr., Clodoveu A. Davis Jr., “An ontological gazetteer and its application for place name disambiguation in text”, Article in Journal of the Brazilian Computer Society · November 2011, DOI: 10.1007/s13173-011-0044-4 · Source: DBLP

[17] Yingjie Hu, Krzysztof Janowicz, Sathya Prasad, “Improving Wikipedia-based Place Name Disambiguation in Short Texts Using Structured Data from DBpedia”, GIR'14, November 04 2014, Dallas, TX, USA Copyright 2014 ACM 978-1-4503-3135-7/14/11.

[18] Takashi Awamura, EijiAramaki, Daisuke Kawahara, Tomohide Shibata, SadaoKurohashi, “Location Name Disambiguation Exploiting Spatial Proximity and Temporal Consistency”,

Proceedings of SocialNLP 2015@NAACL-HLT, pages 1–9, Denver, Colorado, June 5, 2015.

[19] A. RaziaSulthana and RamasamySubburaj, "An Improvised Ontology-based K-Means Clustering Approach for Classification of Customer Reviews". Indian Journal of Science and Technology Vol 9(15), 10.17485/ijst/2016/v9i15/87328, April 2016.

[20] SK AhammadFahad, Md. MahbubAlam, "A Modified K-Means Algorithm for Big Data Clustering", IJCSET(www.ijcset.net) |April 2016 | Vol 6, Issue 4, 129-132.

[21] Udaya Raj Dhungana, SubarnaShakya, KabitaBaral, Bharat Sharma, "Word Sense Disambiguation using WSD Specific Wordnet of Polysemy Words", International Journal on Natural Language Computing (IJNLC) Vol. 3, No.4, August 2014

[22] Clayton Fink, Christine Piatko, James Mayfield, Danielle Chou, Tim Finin, Justin Martineau, "The Geolocation of Web Logs from Textual Clues", 2009 International Conference on Computational Science and Engineering, 978-0-7695-3823-5/09 $26.00 © 2009 IEEE DOI 10.1109/CSE.2009.584

[23] Joachim Kleb and Raphael Volz, "Ontology-based Entity Disambiguation with Natural Language Patterns", IEEE 2009 Fourth International Conference on Digital Information Management (ICDIM) - Ann Arbor, MI, USA.

[24] Zhao Lu, ZhixianYan, Liang He, "OnPerDis: Ontology-based Personal Name Disambiguation on the Web", 2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)978-1-4799-2902-3/13 $31.00 © 2013 IEEEDOI 10.1109/WI-IAT.2013.28185

[25] Padmamala R. "Word-Level Translation (Tamil - English) with word sense disambiguation in Tamil using OntNet", IEEE 2015 International Conference on Computing and Communications Technologies (ICCCT) - Chennai, India (2015.2.26-2015.2.27), 978-1-4799-7623-2/15/$31.00 ©2015 IEEE

[26] Sachio Hirokawa, Tetsuya Nakatoh, HirotoNakae, Takahiro Suzuki, "Discovery of Implicit Feature Words of Place Name", 2014 IIAI 3rd International Conference on Advanced Applied Informatics, 978-1-4799-4173-5/14 $31.00 © 2014 IEEE DOI 10.1109/IIAI-AAI.2014.122