Impact Factor Value: 4.029 ISSN: 2349-7084
International Journal of Computer Engineering In Research Trends
Volume 4, Issue 11, November - 2017, pp. 514-518 www.ijcert.org

A Survey on Keyword Interrogation Implication on Document Vicinity Based on Location

Akshay A. Bhujugade*[1], Dattatraya V. Kodavade [2]

[1]*PG Scholar, Dept. of Computer Science and Engineering, DKTE Society's Textile & Engineering Institute, Ichalkaranji (An Autonomous Institute), 416115, India.*

[2]*Professor, Dept. of Computer Science and Engineering, DKTE Society's Textile & Engineering Institute, Ichalkaranji (An Autonomous Institute), 416115, India.*

akshaybhujugade65@gmail.com[1], dvkodavade@gmail.com[2]

Abstract: - Keyword suggestions are the basic feature of the search engine and it accesses relevant information. The naive user doesn't know how to express their queries; keyword suggestion in web search assists users to access relevant information without any prior knowledge of how to express in queries. The keyword suggestion module can use the current location of a user and retrieve documents which are near to user location. The Euclidean distance is measured for user location and the documents locations. Accordingly the edge weight adjustment is done referring initial K-D graph. The keyword-document graph is used to map the keyword queries and the spatial distance between the resulting documents and the user location. The graph is browsed in random walk with restart, for calculating the highest score for better keyword query suggestion. The paper discusses techniques for the keyword suggestions and also about location-aware keyword query suggestion framework and improved partition based algorithm.

Keywords: Query suggestion, Document proximity, spatial databases.

1. Introduction

Data mining is the activity where mining processes like concept hierarchies, to manage attributes onto various levels of abstraction. One of the fundamental key features of web search engine is keyword suggestion. A web search engine has keyword suggestion module which suggests m number of keywords that takes the user search in the right direction. Sometimes the user is not satisfied with the results of a specific query. If users don't know how to express their queries he can use keyword suggestion which is used in web search so that, it can help the user to access relevant information. Query writing is not easy because queries are short and words are ambiguous. The user may not know how to use the query in web search to get effective result.

Few of the Systems provides location-aware keyword suggestion, such that the suggested keyword should retrieve documents near user location and also relevant to a keyword query. For location and user-supplied keyword query, this requirement emerges due to the popularity of spatial keyword search [2]. The query is supplied as arguments that return the object that is spatially close to the location and textually relevant. For example, suppose the tourist wants hotels which are nearest to their location within 5 miles and search hotel's providing two amenities reviews and prices to compare hotels the tourist can launch fast nearest search query with ranking parameters for the first search for retrieving qualified hotels.

One of the common activity in data mining is searching that motivated to develop the methods which retrieve spatial objects. A spatial object has a spatial data with longitude and latitude of the location. Spatial keyword query is a way of searching for the qualified spatial objects. The spatial-keyword query considers the location of the query issuer and the keyword specified by the user. Considering the both spatial and keyword requirements, the goal of spatial keyword query is to find effectively search results that satisfy the search criteria [3].

The spatial database importance has been reflected in the geometric nature of modeling entities. For example, in map location of restaurants, hospitals, hotels are represented as points, while larger extents like parks, lakes, and landscapes are represented by rectangles. Different functionalities of a spatial database can be used in many different ways in specific contexts. In geography information system, range search can be deployed to find all the restaurants in a certain area. Nearest neighbor retrieval can retrieve all the restaurant nearest to given place. However, some existing techniques of keyword suggestion do not consider the location of users. Users have difficulties in expressing their needs because they do not know correct keywords to search. After submitting the query, the user may not be satisfied with the results.

A weighted keyword-document graph will capture the relevance of keyword query and the distance between the resulting document and the user location. The K-D graph is used for connecting keyword queries with their relevance documents. The weights on edges are adjusted for capturing not only the semantic relevance between the keyword queries but also the spatial distance between the query issuer location and document locations. On K-D graph the random walk with restart (RWR) is applied to measure graph distance. The computational cost of RWR search is high on large graphs. Hence for scaling up RWR search requires pre-computational, but the edge weights of K-D graph are unknown in advance. So there is none of the technique that can accelerate RWR when the edge weight is unknown in advance. The partition-based algorithm is used to addresses this issue, and PA reduces the RWR search cost on the dynamic bipartite graph. The PA divides the keyword queries and the documents into a number of partitions and uses a lazy mechanism that will accelerate RWR search. A partition-based algorithm (PA) that greatly reduces the computational cost of BA.

The paper is organized as follows Section 2 contain the literature survey of keyword suggestion techniques, Section 3 contain architecture of proposed work, Section 4 contain implementation part of LKS framework, and in Section 5 we discussed results, Section 6 we conclude about work done, Section contains 7 future work.

2. Literature Review

R. Baeza-Yates et al. [4] discusses a methodology for a during query firing process search engine suggest m related queries, this related queries are based on queries issued previously, and can issued by user to redirect the search process. This will further improve the notion of interest of the suggested queries and to develop other notions of interest for the query recommender system. Query clustering is done to achieve the semantically similar queries. During clustering process, the use of the content of historical preferences of users is checked. The method also ranks related queries to relevance criteria.

P. Berkhin [5] presented BCA Computes authority weights over the web pages utilizing the web hyperlink structure. In the original BCA, a node distributes its ink aggressively and care only about the nodes with ink greater than ϵ. BCA can be optimized by using lazy updating Mechanism and spatial proximity caching. BCA results in a Bookmark coloring vector. BCA models RWR as a bookmark coloring process.

N. Craswell et al. [6] discussed a search engine which has the ability to record the documents which were clicked for which query. In a weighted graph, RWR (Random Walk with Restart) gives the relevance score of two nodes. RWR specify how closely related the two nodes are in graph. RWR do not scale for large graphs. The Markov random walk is applied to a large click log. The advantage is it will retrieve relevant documents that are not yet been clicked for that query and rank effectively.

H. Cao et al. [7] devised a context-aware query suggestion approach follows two-step offline model learning step and online query suggestion step. Offline model learning step, address data sparseness, queries are summarized into concepts by forming the cluster of click-through bipartite, Then sequence suffix tree is generated from session data for query suggestion model. In online query suggestion step, mapping of query sequence is achieved by capturing search context. By looking up the context in the concept sequence suffix tree this approach suggests queries to the users in context-aware manner.

M. P. Kato et al. [8] mentioned when there is rare or single-term queries input, the search engines should provide better assistance. And according to searcher's current state they should dynamically provide query suggestions. It will further investigate the usage of query suggestion with data sets including user information to propose a query reformulation taxonomy specifically designed for query suggestion classification, and to improve query suggestion functionality based on our insights.

Shuyao Qi et al. [1] devised bookmark coloring Algorithm that computes the RWR based on the top-m query suggestion as a baseline algorithm. BA processes the nodes in the graph in descending order of their active ink. BA only ranks keyword query nodes. The Baseline algorithm has drawbacks such as the no of iterations are more it is time-consuming process. BA is slow for several reasons. Only one node is processed at each iteration. If the number of iteration is more, there is an overhead of maintaining the queue. To improve the performance of BA, the partition Based Algorithm is proposed PA divides the documents and keyword queries in KD graph g into a number of groups. PA adopts a lazy mechanism that accelerates RWR search. As partitions are created the number of iteration are less, it is also time-saving process.

Our proposed Studies shows that the partition based algorithm performance is effective and less time-consuming as that of baseline algorithm. First of all the keyword document graph is needed to be constructed that is the initial K-D graph. Random walk with restart provides a good relevance score between two nodes in a weighted graph. The graph is browsed in random walk with restart, for calculating the highest score for better keyword query suggestion [9] [10] [11]. The K-D graph Gq has two types of nodes that is keyword query nodes and document nodes. Then the user's location is detected and current longitude and latitude are obtained. Each document has its own longitude and latitude. The user longitude and latitude is represented by λ and user query by q, and both are represented by λ_q. Based on the user location the Euclidean Distance is measured for each document di. Based on the location the edge weight adjustment is done. The two directed graph that is keyword queries node to documents node and documents node to keyword queries node is built, and edge weight is adjusted depending upon user's location.

The baseline algorithm has to process each node, so the number of iteration is more due to at each iteration, only one node is processed so as the number of nodes is more than the performance of baseline algorithm decreases. To address this performance issue, the partition based algorithm accelerate the performance by forming the group of nodes as a partition, so the partition of keyword queries and documents are done. By using this algorithm, the system returns the top-m keyword suggestion which is nearest to user location λ.

3. Architecture of Proposed Work

The proposed system as shown in Figure 1 contains the location-based service during the query transaction. The server has to find the user location and match with the corresponding server and retrieve the suggestion in effective manner. The existing system does not consider the rating of the document while suggesting keyword. The star rating of the document is not considered so the proposed system considers the location of query issuer as well as the document rating. In this way, the proposed system can improve performance of keyword suggestion module of web search engine.

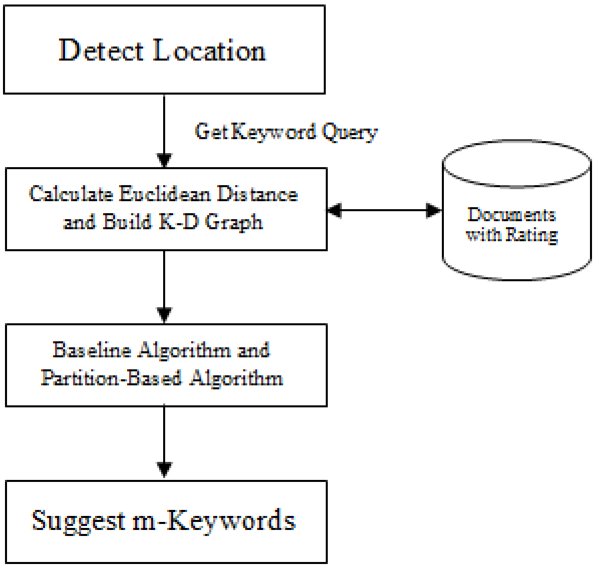
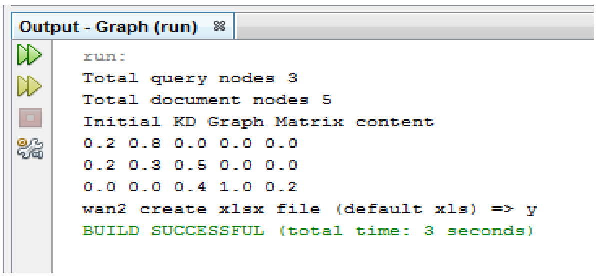

Detect Location
Get Keyword Query
Calculate Euclidean Distance and Build K-D Graph
Documents with Rating
Baseline Algorithm and Partition-Based Algorithm
Suggest m-Keywords

Figure 1. System architecture of keywords suggestion

4. Implementation

The graph is represented in the matrix. The K-D graph has three keyword query nodes and five document nodes as per hard-coded data. When there is no edge from keyword query node to document node the edge weight is set to zero in the matrix. Following Figure 2 shows the initial K-D Graph matrix.



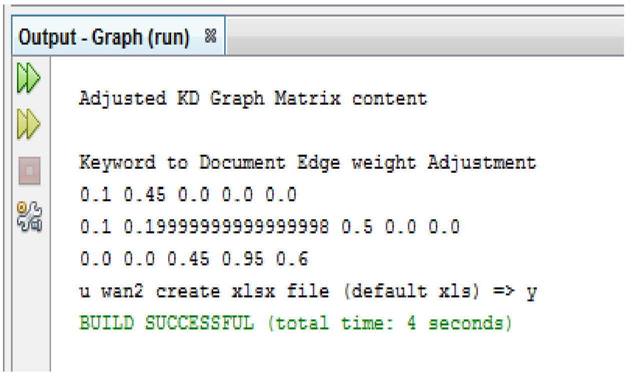
```
run:
Total query nodes 3
Total document nodes 5
Initial KD Graph Matrix content
0.2 0.8 0.0 0.0 0.0
0.2 0.3 0.5 0.0 0.0
0.0 0.0 0.4 1.0 0.2
wan2 create xlsx file (default xls) => y
BUILD SUCCESSFUL (total time: 3 seconds)
```

Figure 2. Initial K-D Graph

Figure 3 shows the adjusted K-D graph from keyword query node to document node which is represented in the matrix using following equation [1].

$$\tilde{\omega}(e) = \beta \times \omega(e) + (1-\beta) \times (1 - dist(\lambda_q, d_j.\lambda) \qquad (1)$$

Where $\omega(e)$ is the initial weight of edge e. $\tilde{\omega}(e)$ is the adjusted edge weight, $dist(\lambda_q, d_j.\lambda)$ is the Euclidean distance between the user's location λ_q and document node d_j and β is set to 1. This adjustment increases the weights of edges of the documents that are near to the user's location.



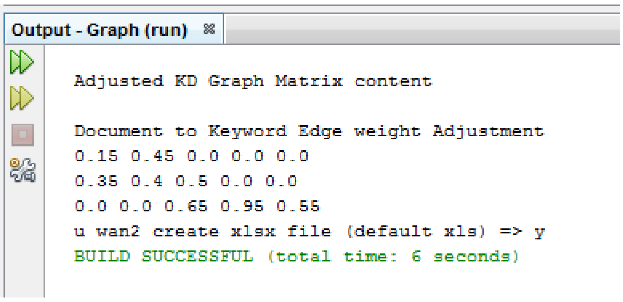
```
Adjusted KD Graph Matrix content

Keyword to Document Edge weight Adjustment
0.1 0.45 0.0 0.0 0.0
0.1 0.19999999999999998 0.5 0.0 0.0
0.0 0.0 0.45 0.95 0.6
u wan2 create xlsx file (default xls) => y
BUILD SUCCESSFUL (total time: 4 seconds)
```

Figure 3. Adjusted K-D Graph from keyword to document

Figure 4 shows the adjusted K-D graph from document node to keyword query node. The adjusted graph is also represented in the matrix using following equation [1].

$$\tilde{\omega}(e') = \beta \times \omega(e') + (1-\beta) \times (1 - mindist(\lambda_q, D(k_i))) \qquad (2)$$

Where $\omega(e)$ is the initial weight of edge e. $\tilde{\omega}(e)$ is the adjusted edge weight, $mindist(\lambda_q, D(k_i))$ is the minimum Euclidean distance of the document node d_j connected to keyword query node k_i with the user location λ_q and β is set to 1.



```
Adjusted KD Graph Matrix content

Document to Keyword Edge weight Adjustment
0.15 0.45 0.0 0.0 0.0
0.35 0.4 0.5 0.0 0.0
0.0 0.0 0.65 0.95 0.55
u wan2 create xlsx file (default xls) => y
BUILD SUCCESSFUL (total time: 6 seconds)
```

Figure 4. Adjusted K-D Graph from document to keyword

5. Results

As the keyword suggestions are based on the location the user locations longitude and latitude are detected, and also each document has its own longitude and latitude. The Euclidean distance is calculated for each document with same user location λ_q, and the edge weight adjustment is done using initial K-D graph.

6. Conclusion

The keyword suggestion module suggests m-keywords based on user location λ_q . The LKS framework suggests keywords considering user location and retrieve document near user location with semantic relevance. We conclude that by doing edge weight adjustment, the weights of edges are increased of those documents which are nearest user location.

7. Future Work

Further, we are going to implement baseline algorithm, partition based algorithm to compute top m keyword suggestion that considers user location.

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] D. Wu, M. L. Yiu, and C. S. Jensen, "Moving spatial keyword queries: Formulation, methods, and analysis," ACM Trans. Database Syst., vol. 38, no. 1, pp. 7:1–7:47, 2013.

[3] Y. Lu, J. Lu, G. Cong, W. Wu, and C. Shahabi, "Efficient algorithms and cost models for reverse spatial-keyword k-nearest neighbor search," ACM Trans. Database Syst., vol. 39, no. 2, pp. 13:1–13:46, 2014.

[4] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query recommendation using query logs in search engines", in

Extending Database Technology, pp.588–596, 2004.

[5] P. Berkhin, "Bookmark-coloring algorithm for personalized pagerank computing," Internet Math., vol. 3, pp. 41–62, 2006.

[6] N. Craswell and M. Szummer, "Random walks on the click graph," in Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval , pp. 239–246, 2007.

[7] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-aware query suggestion by mining click-through and session data," in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 875–883,2008.

[8] M. P. Kato, T. Sakai, and K. Tanaka, "When do people use query suggestion Inf. Retr., vol. 16, no. 6, pp. 725–746, 2013.

[9] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in Proc. 6th Int. Conf. Data Mining, pp. 613–622, 2006.

[10] N. Craswell and M. Szummer, "Random walks on the click graph," in Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 239–246, 2007.

[11] Y. Fujiwara, M. Nakatsuji, M. Onizuka, and M. Kitsuregawa, "Fast and exact top-k search for random walk with restart," Proc. VLDB Endowment, vol. 5, no. 5, pp. 442–453, Jan. 2012.

[12] V. Swathi ,D. Saidulu , B. Chandrakala," Enabling Secure and Effective Spatial Query Processing on the Cloud using Forward Spatial Transformation," International Journal of Computer Engineering In Research Trends.,vol.4,no.7,pp.301-307, July2017.

About the authors

Mr. Akshay A. Bhujugade pursed Bachelor of Engineering from Savitribai Phule Pune University, Pune in year 2016, He is currently pursuing Master of Technology from DKTE's TEI, (An Autonomous Institute), Ichalkaranji, India. His research work focuses on recommendation systems.

Dr. D. V. Kodavade, the Head of Department of Computer Science & Engineering, at DKTE's TEI, (An Autonomous Institute), Ichalkaranji, India. He is a member of the ACM, CSI, IEEE Computer Society. His current research interests includes Artificial Intelligence & Knowledge Based Systems, IoT, Neural Networks, Hybrid Intelligence.