IJRRECS/January 2014/Volume-2/Issue-1/1917-1922 **ISSN 2321-5461**

INTERNATIONAL JOURNAL OF REVIEWS ON RECENT ELECTRONICS AND COMPUTER SCIENCE

ADVANCEMENT TOWARDS INFORMATION RETRIEVAL BASED ON THE SEMANTIC WEB

S.Vinod Kumar Raju[1], Senduru Srinivasulu[2]

[1]M.Tech Student, Dept of IT, Sathyabama University, Chennai, T.N, India

[2]Associate Professor, Dept of IT, Sathyabama University, Chennai, T.N, India

ABSTRACT:

The expanding of web makes the increase of number of pages indexed in search engines respectively. Huge amounts of available information as well as a high rate of novel information have been modernized. When a query was submitted by the user to a search engine, a succession of web-snippets will be returned towards the user. The concepts of extracted and click-through data were used to build a graph of Query-Concept bipartite. If a user clicks on the result of search, the concepts which are appeared in the web-snippet of the result of search are correlated to the equivalent query in the graph of bipartite. A hybrid semantic strategy was introduced to estimate the similarity of query on the basis of data of user click-through which is exploited to recognize the interest of user. The technique of query recommendation which is based on the model of TF-IQF initially breaks the URLs into self-regulating tokens by means of splitting every URL string with separators followed by some stopping words are separated.

Keywords: Search engine, Query, URL, Hybrid semantic strategy.

1. INTRODUCTION:

With the huge volume of data, it is more and more complicated to discover relevant information which can convince the requirements of users on the basis of queries of simple search. The queries which are submitted to search engine by means of users have a propensity to be short and

uncertain [4]. It has observed that the average length of queries which are submitted to search engines was merely 2.35 terms. Lots of pages which are retrieved may possibly be inappropriate to the users' requirements because of the unclear queries. Users may possibly not desire to reformulate their queries by using search terms, in view of the fact that it compels extra burden on them all through searching process. A hybrid semantic strategy was introduced to estimate the similarity of query on the basis of data of user click-through which is exploited to recognize the interest of user. A user clicks on the result of search mainly since the web-snippet contains the applicable matter that the user is concerned in. Hybrid semantic similarity scheme is subsequently consists of the subsequent three most important steps such as when a query was submitted by user, concepts or tokens in addition to their associations are mined from web-snippets to put up a bipartite graph; the query likeness is subsequently calculated on the basis of constructed bipartite graph, and a method of hybrid similarity calculation is recommended as shown in fig1; the most comparable queries are recommended on the way to the user for searching enhancement.

The performance of the approach was evaluated by means of the standard measures, specifically precision, recall in addition to F-Measure, which are expansively adopted methods in research area of IR in addition to natural language processing. The technique of concept extraction is inspired by the renowned setback of discovering recurrent item sets in data mining [8]. When a query was submitted by the user to a search engine, a succession of web-snippets will be returned towards the user. If a phrase t_i appears commonly in the web-snippets of a meticulous query q, subsequently ti can be considered as a notion connected to q, for the reason that it coexists in close propinquity with q. The following cut off formula was applied to calculate the interest of t_i with regard to the returned web-snippets occurring from q.

$$cutoff(t_i) = \frac{sf(t_i)}{n} \times |t_i|$$

n is the total number of returned web-snippets, $sf(t_i)$ is web-snippets number holding t_i, and $|t_i|$ is the number of terms within t_i. The concepts of extracted candidate frequently deal out in all of the web-snippets, on the other hand, it is expected that users are concerned towards

browsing and looking for results from the initial page which are returned by search engine [1]. In order to undertake with the difficulty of information insufficiency, an additional way is to make use of features further than URL. It has been noticed that numerous tokens appeared in the URL are momentous, in particular those superior web pages which are clicked by users with superior prospect. The technique of query recommendation which is based on the model of TF-IQF initially breaks the URLs into self-regulating tokens by means of splitting every URL string with separators followed by some stopping words are separated [3]. With the extraction of meaningful tokens, each query can be corresponded to a token vector, specifically TF-IQF vector, which is identical to the accepted model of TF-IDF within IR. The concepts of extracted and click-through data were used to build a graph of Query-Concept bipartite, in which one side of the vertices symbolize exclusive queries, and the other side of vertices matches up to unique notions [9]. If a user clicks on the result of search, the concepts which are appeared in the web-snippet of the result of search are correlated to the equivalent query in the graph of bipartite.

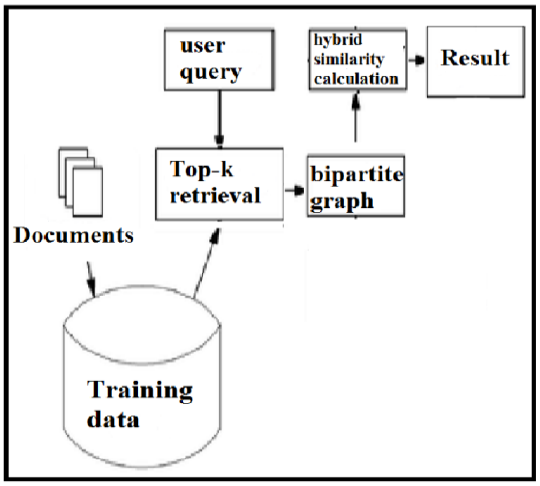

Fig1: An overview of hybrid similarity calculation.

2. AN OVERVIEW OF QUERY RECOMMENDATION ON BASIS OF HYBRID SEMANTIC SIMILARITY:

Implicit feedback and patterns of query can be exploited to systematize web documents, specifically; queries of user can be treated as words of document characteristic to resolve the setback of dictionary [10]. Users can decide to outlook documents as the expansions of query word. Users may possibly not desire to reformulate their queries by using search terms, in view of the fact that it compels extra burden on them all through searching process. In the shown fig2, Doc1, Doc2 represents the Query1, as well as Doc2, Doc4 represents Query3. This method can solve the difficulty of query lacking representation. If the URLs are clicked during the procedure of query

search, regarding these connected URLs, the documents which are clicked to outlook can be measured as the related document [7]. It is instinctive that two queries will contain a superior similarity if they contribute to more terms of synonymous. If the original as well as other queries contribute to the identical or synonymous terms, they are expected to be retrieval appropriate. The recovery of identical queries can get better the superiority of the query expansion. Hence the connected queries will be recovered as the expansion terns of candidate query. During the clicking of URL, it is recognized that the two queries proportionate to the identical URL are extremely appropriate. Genetic algorithm based clustering is exploited to cluster queries and prompts the structure of subtopic which is intended for their users [2]. Genetic algorithm based clustering is a randomized search and technique of optimization which is guided by means of the natural selection principals in addition to heredity. It is resourceful and tough search processes which carries out search in huge landscapes and make available near-optimal elucidations for the idea of an optimization trouble. Each query is primarily measured as an individual point in the space of search [5]. A random grouping of queries is programmed in the form of string, known as chromosome and chromosomes collection is known as a population, and a population of random distributed is created. Three operators of biologically inspired such as selection, crossover in addition to mutation, are used to give in novel child chromosomes [6]. These operators carry on quite a few generations till the termination standard of termination is fulfilled. The tough chromosomes with the superior fitness are chosen to subsequent generation intended for retaining the tremendous seeds.

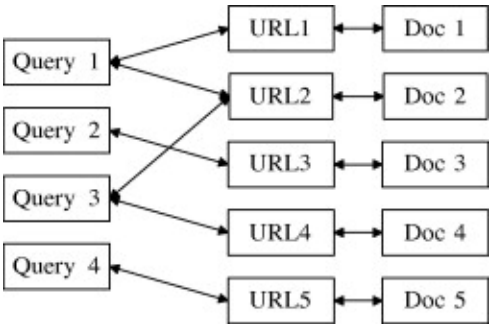

Fig2: An overview of clicked documents

3. RESULTS:

To assess the performance of the propose method, the experimental setup was performed for collecting the necessary data of click-through data. Google was implemented to look for specified test queries intended for collecting the data of

click-through data. The performance of the approach was evaluated by means of the standard measures, specifically precision, recall in addition to F-Measure, which are expansively adopted methods in research area of IR in addition to natural language processing. The given figure specifies that our method of hybrid semantic similarity can attain improved recall and precision concurrently.

4. CONCLUSION:

A hybrid semantic strategy was introduced to estimate the similarity of query on the basis of data of user click-through which is exploited to recognize the interest of user. The technique of query recommendation which is based on the model of TF-IQF initially breaks the URLs into self-regulating tokens by means of splitting every URL string with separators followed by some stopping words are separated. Numerous tokens appeared in the URL are momentous, in particular those superior web pages which are clicked by users with superior prospect. With the extraction of meaningful tokens, each query can be corresponded to a token vector, specifically TF-IQF vector. The performance of hybrid semantic strategy was evaluated by means of the standard measures and was observed that it can attain improved recall and precision.

REFERENCES:

[1] Y.H. Wu, Y.C. Chen, and A.L.P. Chen, "Enabling Personalized Recommendation on the Web based on User Interests and Behaviors," in Proceeding of International Workshop on Research Issues in Data Engineering , RIDE IEEE, 2001

[2] L. Ding, T. Finin, A. Joshi, Y. Peng, R. Pan, and P. Reddivari. Search on the semantic web. IEEE Computer, 10, 2005.

[3] H.J. Kim, and S.G. Lee, "A Semi-Supervised Document Clustering Technique for Information Organization," in Proceeding of the 9th International Conference on Information and Knowledge Management, 2000.

[4] The semantic web: Roles of XML and RDF, Stefan Decker and Sergey Melnik, Frank Van Harmelen, Dieter Fensel, And Michel Klein Jeen Broekstra Michael Erdmann Ian Horrocks, IEEE Internet Computing, October 2000, vol. 15, nr. 3, pgs. 63--74.

[5]Gómez-Pérez, A. and O. Corcho, 2004. Ontology languages for the semantic web. IEEE Intelligent Systems, 17: 54-60.

[6] K. Wang, C. Xu, and B. Liu, "Clustering Transactions Using Large Items," in Proceeding of the 8th International Conference on Information and Knowledge Management, ACM, 1999.

[7] K. Chang, B. He, Z. Zhang (2004, December). "Mining semantics for large scale integration on the

web: evidences, insights, and challenges". ACM SIGKDD Explorations Newsletter, Volume 6, Issue 2. pp. 67-74

[8] B. Mobasher, R. Cooley, and J. Srivastava, "Creating Adaptive Web Sites Through Usage-Based Clustering of URLs," in Proceedings of the Workshop on Knowledge and Data Engineering Exchange, 1999.

[9] L. Ding, R. Pan, T. Finin, A. Joshi, Y. Peng, and P. Kolari. Finding and ranking knowledge on the semantic web. In Proceedings of the 4th International Semantic Web Conference, pages 156–170, 2005

[10] J. Wen, J.Y. Nie, and H.J. Zhang, "Clustering User Queries of A Search Engine," in Proceedings of the 10th International World Wide Web Conference, 2001.