Web Search Recommendation System Using Concept Based Mining Techniques

V. Jayashree[*1]
Department of Computer Science,
S.N.R. Sons College,
Coimbatore-641046, India

Mr. A. Balasubramanian[2]
Professor and Head,
Department of Information Technology,
S.N.R. Sons College, Coimbatore-641046, India

Abstract: Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Web content mining is the application of data mining techniques to discover patterns from the web. This research paper proposes a method to extend with Concept based mining to prove the terms of accuracy. The main purpose for concept based is to extract previously unknown relationships between Web pages. This allows the user to access the desired information through keyword association and content mining. In this proposed method the concept based web search is more efficient to retrieve the information based on the user interest. The Concept based mining is used to mine the web snippet of the web topics and display the top ranking documents at first in the resulted web pages. Concept based filtering is mainly used to provide the related web document for the suggested query. By using this method the results are displayed in the rank wise in the web pages.

Keywords: Web mining, concept based mining, query suggestion, concept filtering, recommendation system, web snippets.

I.   INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Concept Mining is used to search or extract the concepts embedded in the text document. These concepts can be either words or phrases and are totally dependent on the semantic structure of the sentence. When a new text document is introduced to the system, the concept mining can detect a concept match from this document to all the previously processed documents in the data set by scanning the new document and extracting the matching concepts [5]. In this way, the similarity measure is used for concept analysis on the sentence, document, and corpus levels. These concepts are originally extracted by the semantic role labeler [5] and analyzed with respect to the sentence, document, and corpus levels. The concept based algorithm aims to provide queries that are, at least somewhat, similar to the original [1]. And also is to use query expansion [2], where the goal is to find keywords that, while syntactically different to varying degrees, have the same semantics as the keywords in the original query [3]. Query recommendation can be seen as query expansion where we limit the query universe to those queries that have been previously input by some user [4]. Queries submitted to search engines convey implicit knowledge if we assume that most of the time user actions are meaningful.

Queries can be seen as tags associated to documents clicked by the people making those queries, in a similar way that link anchor text is used as surrogate text of the linked Web pages [5][6].Thus, the matching among these concepts is less likely to be found in non - related documents. If these concepts show matching in unrelated documents, then they produce errors in terms of noise. Therefore, when text document similarity is calculated, the concepts become insensitive to noise. The concept-based mining model can effectively discriminate between non-important terms with respect to sentence semantics and terms which hold the concepts that represent the meaning of the sentence. These are the methods and algorithms used in this proposed research work**.**

II.   LITERATURE REVIEW

Eugene Agichtein et al. [1] proposed Improving Web Search Ranking by Incorporating User Behavior Information and it incorporating implicit feedback can augment other features, improving the accuracy of a competitive web search ranking algorithms. Author explored the utility of incorporating noisy implicit feedback obtained in a real web search setting to improve web search ranking.

Ricardo Baeza-Yates et al. [3] proposed extracting the semantic relations that are implicitly captured in the actions of users submitting queries and clicking answers. Queries can be seen as tags associated to documents clicked by the people making those queries, in a similar way that link anchor text is used as surrogate text of the linked Web pages.

Ricardo Baeza-Yates et al. [4] proposed query logs in search engines. In this method query submitted to search engines and suggests a list of related queries. The related queries are based in previously issued queries, and can be issued by the user to the search engine to tune or redirect the search process.

Georges Dupret et al. [14] proposed a novel query recommendation algorithm. A query is a set of one or more keywords that represent an information need formulated to the search engine. The same query may be submitted several times. Each submission induces a different query instance. A query session consists of one query and the URLs the user clicked on, while a click is a Web page selection belonging to a query session. We also define a notion of consistency between a query and a document.

Luo Si, Rong Jin. [15] proposed Concept-Based Information Retrieval Using Explicit Semantic Analysis. This concept based approach generates new text features automatically, and we have found that high-quality feature selection becomes crucial in this setting to make the retrieval more focused. Concept-based IR represents both documents and queries using semantic concepts, instead of keywords, and performs retrieval in that concept space. This approach holds the promise that representing documents and queries using high-level concepts will result in a retrieval model that is less dependent on the specific terms used. Such a model could yield matches even when the same notion is described by different terms in the query and target Documents, thus alleviating the synonymy problem and increasing recall.

III. METHODOLOGY

Concept based mining method is used in this paper to mine the concept oriented web topics in the network for the user query. Concept based filtering technique is used to ignore the non related web topics from the database. The Concept Based Mining method is motivated by the paper "Concept based mining for Domain-Specific Search" [7]. The Concept based mining is used to display the top ranking documents at first in the resulted web pages and it is more efficient to retrieve the information based on the user interest.Concept based mining is used for document retrieval. Using this method the results are displayed in the rank wise. Content Based Filtering is mainly used to provide only the web topics related to the suggested query. Due to that many web users are benefitted with low accessing time and easy search to find the relevant web data. The proposed Concept based Algorithm with content based filtering is used to improve the efficiency of the web data links and time period with high accuracy.The architecture is shown Fig.1
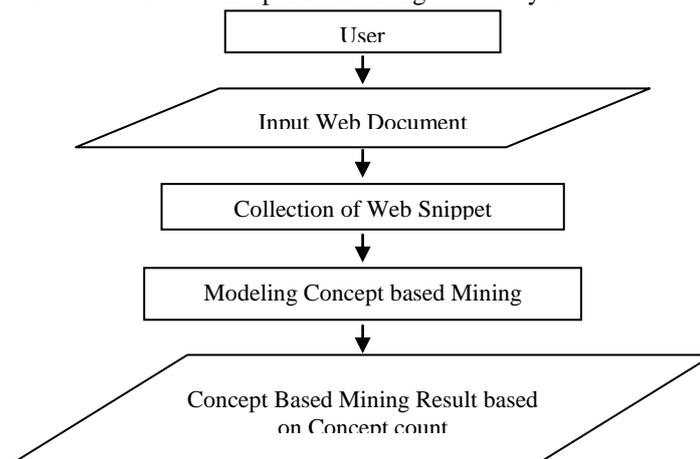

Fig.1. System Architecture

1.1. INPUT WEB DOCUMENT

In the proposed system, the mining of the user query depends on the concept based mining algorithm by using concept based filtering technique. Initially the concept oriented web data links, web snippets and URL Cache links are collected from the network and store it under the basis of domain, category and concept of each web data links in the database of the local server. When the user gives the query it will fetch the web topics along with the web data links, Web snippets and URL Cache Links from the database of the local server. Then consider web topics as input of the web document.

1.2. CONCEPT BASED MINING

Concept Mining is used to search or extract the concepts embedded in the text document. These concepts can be either words or phrases and are totally dependent on the semantic structure of the sentence. When a new text document is introduced to the system, the concept mining can detect a concept match from this document to all the previously processed documents in the data set by scanning the new document and extracting the matching concepts [5].

By using the concept based mining method, when the user gives the query, initially it collect all the web snippets and cache links which appeared in the input web document. The web snippets are splited separately and analyse each web topic snippet and find the occurrences. Concept mining algorithm is used to find the occurrences of the query in each web snippet. The advantages are

➢ Time efficiency through the information retrieval
➢ High efficiency in low cost
➢ Intended recommendation algorithm is scalable to very large data sets.
➢ Display the related concepts based on the user interest
➢ Display the results based on the heat value and concept based results

1.3. CONCEPT BASED MINING ALGORITHM:
➢ **Input:** A set of Web Data Links (W), URL Links (L) and Web Snippets(S)
➢ **Output**: Web Search results Re ranking using Concept Based Mining
➢ **Step 1:** A given set of Query Q={Q1,Q2,Q3….Qn}
➢ **Step 2:** For each keyword fetch web results of the suggested query
➢ **Step 3:** For the collection of web Results apply concept based filtering technique
➢ Web results [i]=web search results[i];
➢ **Step 4:** Find the frequent concept count from the web results.
➢ CC[t]={c1,c2………..cn};
➢ **Step 5:** Finally get Re-ranked Web search results by using concept based mining.

1.4. CONCEPT BASED FILTERING
Concept based filtering is used to avoid the non related web topics for user convenient. By using this technique, user will get the concept oriented web data links. Based on the web snippet count the web data links are ranked and discover the frequently used web snippet in the web page. This method is used to identify the accurate concept for user's query.

IV. PERFRMANCE MEASURES
1.5. CONCEPT BASED COUNT RESULT
By using the concept based mining method, web snippets are mined andanalyse the occurrences of the suggested query. The below table shows the concept count of the web snippets from the web data links.
Concept mining algorithm is used to find the occurrences of the query in each web snippet. Concept based filtering is used to avoid the non-related web topics for user convenient. By using this technique, user will get the concept oriented web data links.

WEB SNIPPET	CONCEPT COUNT
Health Calculators Widgets Animations Videos with 120000 plus	32
programs resources news and events and links to other web sites	26
Read doctor produced health and medical information written for	23
itunes apple com us genre ios medical id6020 mt 8 United	22
Screen reader users click here to turn off Google Instant	22
Google Instant is off due to connection speed Press Enter to search	21
18 May 2012 The A D A M Medical Encyclopedia includes over 4000	21
you to make informed decisions about your health concerns	21
Visit Google s Ads Preferences Manager to learn more or opt out	0
A statutory body regulating medical colleges affiliation new	20

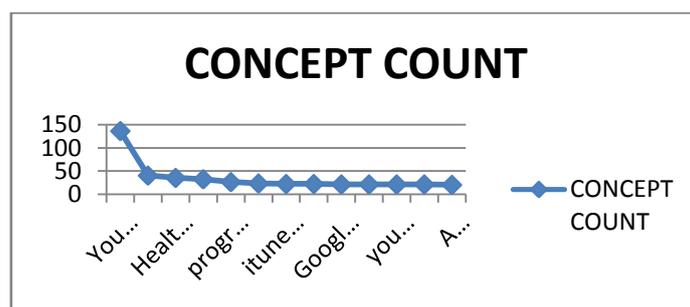Fig.2.Web Snippets and Concept Count

Fig. 3. Graph of Concept Based Count Accuracy

The above figure shows the concept based count accuracy by using the web snippets of the suggested query. By using this concept count accessing time get reduced and its more efficient for the user convenience.

V.   COMPARATIVE ANALYSIS OF QUERY SUGGESTION AND CONCEPT BASED MINING

In the previous work, web data links are mined with the help of heat diffusion technique. By using the technique heat value is calculated and based on heat value web data links are ranked and arranged in the web page.In the proposed system, web snippets are mined along with concept based filtering technique, the concept count and click count is calculated and hybrid. By using the hybrid concept and click count value the web data links are ranked and arranged in the web page.

By comparing Concept based filtering with the heat diffusion technique, the comparison performance has been displayed in the form of chart with the help of web snippet and heat value accuracy. Finally the accuracy of heat diffusion technique and concept based mining showed in the form of graph. Compared with existing method, the concept based method provides high accuracy and reduced the heat value.

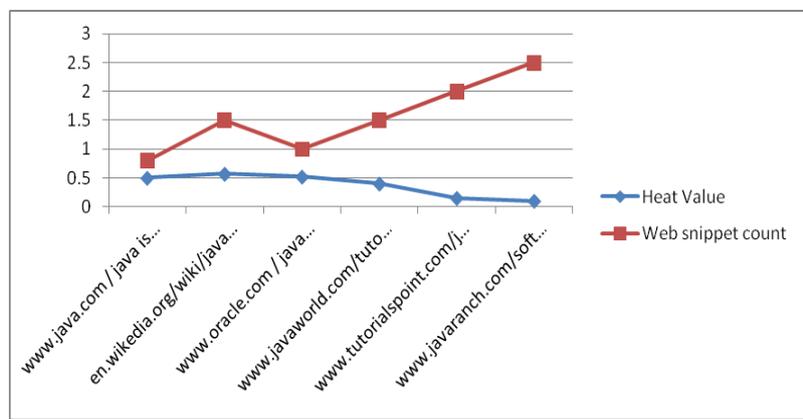| WEB DATA LINKS	HEAT VALUE ACCURACY	WEB SNIPPET ACCURACY
www.java.com / java is oops language Education	0.57	1.5
en.wikedia.org/wiki/java programming language Entertainment	0.52	1
www.oracle.com / java /java se	0.4	1.5
www.javaworld.com/tutorial	0.15	2
www.tutorialspoint.com/java	0.1	2.5
www.javaranch.com/software	0.15	3

Fig.4. Accuracy Improvisation Result

Fig.5. Comparison of heat diffusion and concept based click Count

The above figure shows the comparison performance result of web snippet count and heat value accuracy. Compared with previous work, the concept based method provides high accuracy and reduced the accessing time for the recommended query.

VI.   CONCLUSION

In this research work, concept based mining algorithm and content based filtering techniques are used to retrieve the exact data for the suggested query of the web user from the web server and it will help the web user to satisfy their needs and concise the web search time. It will reduce the time taken for the suggested query and it's used to reduce the computational cost and improves the classification accuracy. Finally it will retrieve the exact dataset for the suggested query. The main use of the concept based mining algorithm is to retrieve the exact data for the suggested query.

VII.   FUTURE WORK

The future work more domains can be added by which retrieval of precise dataset for the given query can be retrieved. If this web recommendation process is replaced by another one is suggested for future research.

REFERENCE

[1] E. Agichtein, E. Brill, and S. Dumais, "Improving Web Search Ranking by Incorporating User Behavior Infomation," SIGIR '07:Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 19-26, 2006.

[2] E. Auchard, "Flickr to Map the World's Latest Photo Hotspots," Proc. Reuters, 2007.

[3] R. TiberiBaeza-Yates and A. Tiberi, "Extracting Semantic Relations from Query Logs," KDD '07: Proc. 13th ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining, pp. 76-85, 2007.

[4] R.A. Baeza-Yates, C.A. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc.Current Trends in Database Technology (EDBT) Workshops, pp. 588-596, 2004.

[5] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," KDD '00: Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 407-416,2000.

[6] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Neural Computation.

[7] J.S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI), 1998.

[8] Christian Safran "A Concept-Based Information Retrieval Approach for User-oriented Knowledge Transfer"

[9] P.A. Chirita, C.S. Firan, and W. Nejdl, "Personalized Query Expansion for the Web," SIGIR '07: Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 7-14, 2007.

[10] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, "Query Expansion by Mining User Logs," IEEE Trans. Knowledge Data Eng., vol. 15, no. 4,pp. 829-839, July/Aug. 2003.

[11] J. Xu and W.B. Croft, "Query Expansion using Local and Global Document Analysis," SIGIR '07: Proc. 19th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 4-11,1996

[12] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating Query Substitutions," WWW '06: Proc. 15th Int'l Conf. World Wide Web, pp. 387-396, 2006.

[13] G. Linden, B. Smith, and J. York, "Amazon.com Recommendations: Item-to-Item Collaborative Filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, Jan./Feb. 2003.

[14] Georges Dupret and Marcelo Mendoza "Automatic Query Recommendation using Click-Through Data"

[15] Luo Si, Rong Jin "flexible mixture model for collaborative filtering" Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003), Washington DC, 2003.