International Journal of Multidisciplinary Allied Research Review and Practices

IJMARRP ISSN No: 2455-1570 (Volume 3, Issue 7, July 2016)

A Novel Technique to Recommendation of Query in Search Engine

Sonam yadav[1] , Dr. Mukesh yadav[2]

[1]M.Tech (computer science and engineering), GITM, (GURGAON), INDIA

[2]Assistant professor, computer science and engineering department, GITM (GURGAON),INDIA

ABSTRACT

Query Logs are important information repositories, which record user activities on the search results. The mining of these logs can improve the performance of search engines. Search engines generally return long lists of ranked pages, finding the desired information content from which is typical on the user end and therefore, search result optimization techniques come into play. The proposed system based on learning from query logs predicts user information needs and reduces the seek time of the user within the search result list. To achieve this, the method first mines the logs using a novel similarity function to perform

query clustering and Finally, search result list is optimized by re-ranking the pages using the proposed formula. The proposed system proves to be efficient as the user desired relevant pages occupy their places earlier in the result list and thus reducing the search space. The paper also presents a query recommendation scheme towards better information retrieval.

Keywords

- Search Engine,
- Query Recommendation,
- Data Mining.

1. INTRODUCTION

In this paper, we have explained the overview of web with its functions, and then we are using the concept of web usage mining that is our minor area of my project. The entire data source and the information are being used with the context of WUM. We have explained the introductory stage starting from the basic terminology followed by the problems that occurred during the query recommendations and after that solution has been also advised to optimize the search engine result. Further, it has been explained the various steps involved in the working my project and finally every chapter has been discussed thoroughly.

1.1Web

The World Wide Web abbreviated as WWW or W3 commonly known as the Web is a system of interlinked hypertext documents accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia, and navigate between them via hyperlinks.

Web developed three essential technologies:

1. A system of globally unique identifiers for resources on the Web later known as Uniform Resource Locator (URL) and Uniform Resource Identifier (URI).

2. The publishing language Hyper Text Markup Language (HTML)

3. The Hypertext Transfer Protocol (HTTP).

1.2 Function

The Internet is a global system of interconnected computer networks. It is one of the services that run on the Internet. It is a collection of text documents and other resources, linked by hyperlinks and URLs, usually accessed by web browsers from web servers. The Web can be thought of as an application running on the Internet.

Viewing a web page on the World Wide Web normally begins either by typing the URL of the page into a web browser or by following a hyperlink to that page or resource. The web browser then initiates a series of communication messages, behind the scenes, in order to fetch and display it.

1.3. Web Mining

It is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types.

1.3.1 Web content mining

Mining, extraction and integration of useful data, information and knowledge from Web page contents.

1.3.2 Web structure mining

It is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, it can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.

2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

1.3.3 Web Usage Mining

The following explanation is minor area of my project. Web Usage Mining (WUM) is a part of Web Mining, which, in turn, is a part of Data Mining. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of data, it involves mining the usage characteristics of the users of Web Applications. This extracted information can then be used in a variety of ways such as, improvement of the application, checking of fraudulent elements etc

It is the process of extracting useful information from server logs and finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data.

The major problem with Web Mining in general and Web Usage Mining in particular is the nature of the data they deal with. With the upsurge of Internet in this millennium, the Web Data has become huge in nature and a lot of transactions and usages are taking place by the seconds. Apart from the volume of the data, the data is not completely structured. It is in a semi-structured format so that it needs a lot of preprocessing and parsing before the actual extraction of the required information.

In Web Usage Mining, data can be collected in server logs, browser logs, proxy logs, or obtained from an organization's database. These data collections differ in terms of the location of the data source, the kinds of data available, the segment of population from which the data was collected, and methods of implementation.

1. Search Engine

Pouter defines a WWW search engine as a retrieval service, consisting of a database describing mainly resources available on the WWW, search software and a user interface also available via WWW.

A positive point about the Internet and its most visible component, the World Wide Web, is that there are hundreds of millions of pages available, waiting to present information on an amazing variety of topics. But the negative point about the Internet is that there are hundreds of millions of pages available, most of them titled according to the whim of their author, almost all of them sitting on servers with cryptic names. We can visit a **Search Engine**.

Search Engine is designed searching the information on World Wide Web. Results are generally presented in a list of result often called SERP’s or Search Engine Result Page. The Internet is a global data communications system. It is a hardware and software infrastructure that provides connectivity between computers. In contrast, the Web is one of the services communicated via the Internet. It is a collection of interconnected documents and other resources, linked by hyperlinks and URLs (Uniform Resource Locator) or (Uniform Resource Identifier) URI which also specifies where the identified Resource is available and the protocol for retrieving it.

The plentiful content of the World-Wide Web is useful to millions. Information seekers use a search engine such as Google, Yahoo to begin their Web activity .On the Internet, a search engine is a coordinated set of programs which searches an index and returns matches to a specified keyword. Search Engine is situated on the computer system connected to Internet.

There are differences in the ways various search engines work, but they all perform three basic tasks:

- They search the Internet or select pieces of the Internet based on important words.
- They keep an index of the words they find, and where they find them.
- They allow users to look for words or combinations of words found in that index.

Search Engine to provide best services regularly index millions of web pages involving a comparable number of distinct terms by employing special software known as Web Crawlers or Spiders to retrieve information on web to prepare up catalog for ready reference. The most important measure for a search engine is the search performance, quality of the results and ability to crawl and index the web efficiently. The primary goal is to provide high quality search results over a rapidly growing World Wide Web.

In a Search Engine, user sends the query. If related query is in indexed pages then page related top query returned to user. If required pages not in indexed pages then query is sent to crawler module. Crawler module sends the query to crawlers. Crawler search pages related to query and send those pages to page repository. And also sends the related link back to crawler module. Crawler module when gets these link, it sorts them according to their relevancy and sends them back to crawler. Crawler processes all the links till the list is empty and adds the results to page repository. The Indexer indexes the stored data in a particular format. Collection analysis module stores the pages on the basis of their utility. Ranking module ranks the retrieved pages according to their relevance. Retrieved results are sent back to user.

2. LITERATURE REVIEW

A literature survey is done to identify different approaches proposed by researchers in order to mine essential features from query log data of search engine.

J. Wenet al. [1] presented a content based similarity measure to cluster similar queries to recommend URLs to frequently asked queries of a search engine by using four notions according to: first, the context of the query; second, common clicked URL's between queries; third, string matching of keywords, and fourth, the distance of the clicked documents in some pre-defined hierarchy. But result of this method generates very sparse distance matrices but this sparsely is diminished using large query logs. Thus string matching features are used to locate similar queries.

O. Zaiane et al. [2] have used content similarity to recommend similar queries using Query Memory, which is a data structure that holds the collective query trace and also extra information pertaining to the queries that would help in measuring similarities between queries. Query trace is a log containing previously submitted queries. The major advantage of this method is that it suggests the queries when user is not satisfied by current search result but sometimes produces irrelevant result and leaves the choice up to user.

S. Cucerzan et al. [5] have presented a click context based method that suggests queries based on mining into post-query browsing behaviors referred as search trails. They utilized user landing pages which are the ending pages of search trails to generate query suggestions. For each landing page of a user submitted query they identify queries from query logs that have these landing pages as one of their top 10 results ad these queries are used for suggestions.

C. Sumathi et al. [9] also proposed a session based approach where the proposed system is based on the users navigational patterns and provide recommendations to satisfy the current users information need. In this method they classify and match an online user based on his browsing interests.

Q. He et al. [7] used a session based novel sequential query prediction approach to grasp a users search intent based on users past query sequence and its resemblance to historical query sequence models mined from massive search engine logs. Differently from previous work where only single preceding query is used for prediction, this work considers variable number of preceding query and effectively captures more complex context information for recommendation. Results show that the sequence-wise approaches significantly outperform the conventional pair-wise ones in terms of prediction accuracy. Thus the work has one fundamental difference from all previous session-based approaches. As all previous work focuses on pair-wise query relations and uses only a single preceding query for query prediction, presented method consider a variable number of preceding queries and effectively capture more complex context information for query recommendation. Moreover, this approach can automatically determine the optimal context length to be used for query prediction.

R. Baeza-Yates et al. [4] explained a method to suggest a list of related queries to user based on a query clustering process. The method not only discovers the related queries, but also ranks them according to a relevance criterion. This notion of query similarity has several advantages that it is simple and easy to compute. On the other hand, it allows relating queries that are worded differently but stem from the same topic, hence capturing semantic relationships among queries.

METHODLOGY

The proposed system works in the following steps

1. Similarity Analyzer

2. Query Clustering Tool

3. Favored Query Finder

4. Sequential Pattern Generator

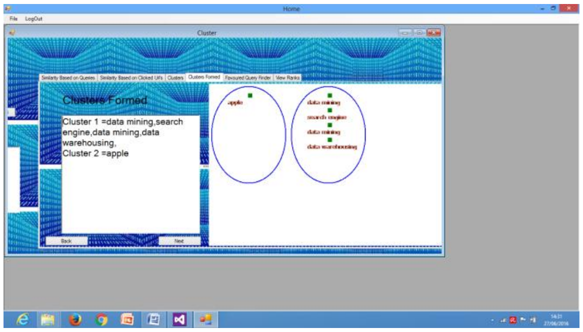
5. Rank Updater

6. Query Recommender

3. PROPOSED WORK

When user submits a query on the search engine interface, the query processor component matches the query terms with the index repository of the search engine and returns a list of matched documents in response. User browsing behavior including the submitted queries and clicked URLs get stored in the logs and are analyzed continuously by the Similarity Analyzer module, the output of which is forwarded to the Query Clustering Tool to generate groups of queries based on their similarities.
Favored Query Finder extracts most popular queries from each cluster and stores them for future reference. The Pattern Generator module discovers sequential patterns of web pages in each cluster. The Rank Updater component works online and takes as input the matched documents retrieved by query processor. It improves the ranks of pages according to sequential patterns which were discovered offline. The Query Recommender guides the user with similar queries with the most famous query highlighted.

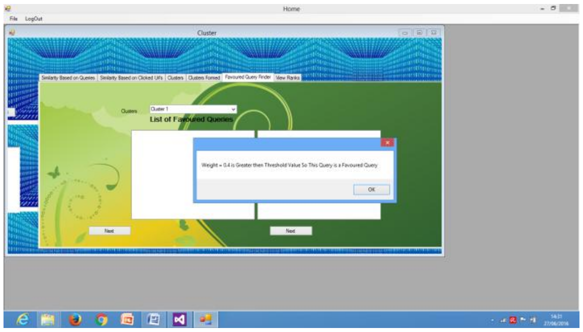
The proposed system works in the following steps
1. Similarity Analyzer
2. Query Recommender
3. Query Clustering



Clusters Formed

Cluster 1 =data mining,search engine,data mining,data warehousing,
Cluster 2 =apple

4. Favored Query Finder

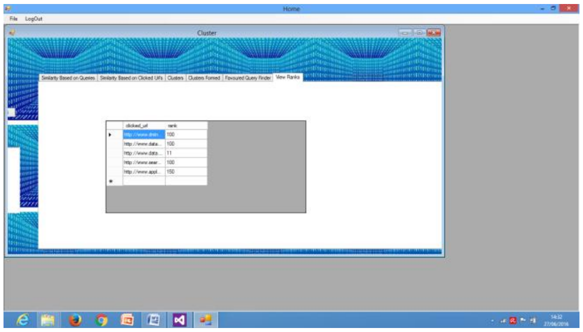


Clusters: Cluster 1

List of Favoured Queries

Weight = 0.4 is Greater then Threshold Value So This Query is a Favoured Query

5. Sequential Pattern Generator
6. Rank updater



clicked_url	rank
http://www.[illegible]	100
http://www.data...	100
http://www.data...	11
http://www.sear...	100
http://www.appl...	150

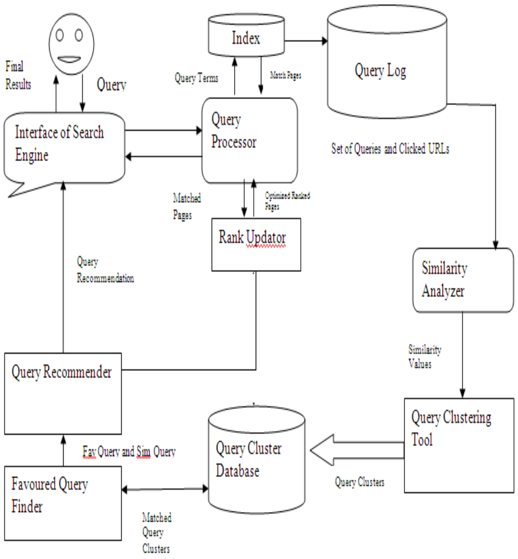


Fig 1. Architecture of Proposed Optimization System

4.1 Improved Rank Updater Algorithm

This module takes its input from the query processor i.e. the matched documents of a user query and an update is applied to modify the rank score of the returned pages. The module operates online at the query time and applies the necessary updates on the concerned documents. The updated documents in question are those which are most frequently accessed by the users and are detected by the Sequential Pattern Generator. The updater works in the following steps:

The popularity from the number of in links and out links is recorded as *Win*(*v* ,*u*) and *W out* (*v*, *u*), respectively .*Win*(*v* ,*u*) given in eq. (3) is the weight of *link*(*v*, *u*) calculated based on the number of in links of page *u* and the number of in links of all reference pages of page *v*.

$$W^{in}{}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} I_p}$$

Where *Iu* and *Ip* represent the number of inlinks of page *u* and page *p*, respectively. *R* (*v*) denotes the reference page list of page *v*. *Wout*(*v*,*u*) given in eq. (4) is the weight of *link*(*v*, *u*) calculated based on the number of outlinks of page *u* and the number of outlinks of all reference pages of page *v*.

$$W^{out}(_{v,u}) = \frac{O_u}{\sum_{p \in R(v)} O_p}$$

Where *Ou* and *Op* represent the number of outlinks of page *u* and page *p*, respectively. *R* (*v*) denotes the reference page list of page *v*.

Considering the importance of pages, the original Page Rank formula is modified in eq. (5) as

$$\mathrm{PR(u)} = (1 - \mathrm{d}) + \mathrm{d} \sum_{v \in B(u)} \mathrm{PR(v)W^{in}{}_{(v,u)}\, W^{out}(_{v,u})}$$

New Formula :

PR(u) =(1-d)+d ∑ PR(v) * W(in)*W(out)*D(v,u)

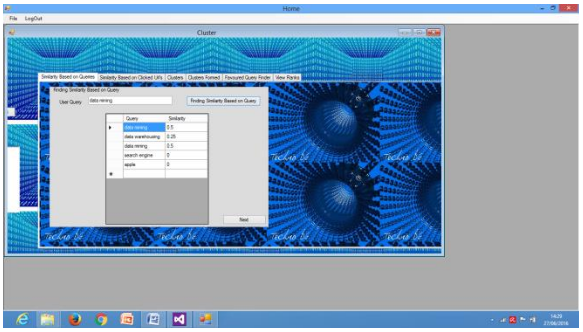
Introduced D in existing formula, D refers here with the number of duplicates

D(v,u) = D (u)/D (v)

Here D(u) and D (p) are the no. of duplicates.

3. RESULTS

A novel approach based on query log analysis is proposed for implementing effective web search with improved page ranking. The most important feature is that the result optimization method is based on users' feedback, which determines the relevance between Web pages and user query words. Since result improvement is based on the analysis of query logs, the recommendations and the returned pages are mapped to the user feedbacks and dictate higher relevance than the pages, which exist in the result list but are never accessed by the user. By this way, the time user spends for seeking out the required information from search result list can be reduced and the more relevant Web pages can be presented.


Home
File
LogOut
Cluster
Similarity Based on Queries
Similarity Based on Clicked Url's
Clusters
Clusters Formed
Favoured Query Finder
View Ranks
Finding Similarity Based on Query
Finding Similarity Based on Query
Next
Techno bd
14:29
27/06/2016

User Query: data mining

Query	Similarity
data mining	0.5
data warehousing	0.25
data mining	0.5
search engine	0
apple	0

The results obtained from practical evaluation are quite promising in respect to improving the effectiveness of interactive web search engines. Further investigation on mining log data deserves more of our attention. Further study may result in more advanced mining mechanism which can provide more comprehensive information about relevancy of the query terms and allow identifying user’s information need more effectively.

REFERENCES

[1] J. Wen,J. Nie, and H. Zhang. 2001. Clustering user queries of a search engine. In Proceedings of the 10th international World Wide Web conference. W3C. pp.162– 168.

[2] O. Zaiane and A. Strilets. 2002. Finding similar queries to satisfy searches based on query traces. In Proceedings of the International Workshop on Efficient Web-Based Information Systems (EWIS), Montpellier, France.

[3] D. Broccolo, O. Frieder, F. Nardini, R. Perego and F. Silvestri.2010. Incremental Algorithms for Effective and Efficient Query Recommendation. SPIRE 2010, LNCS 6393. Pp.13-24. Springer-Verlag Berlin Heidelberg.

[4] R. Baeza-Yates, C. Hurtado and M. Mendoza.2004. Query Recommendation Using Query Logs in Search Engines. LNCS 3268. pp. 588-596. Springer-Verlag Berlin Heidelberg.

[5] S. Cucerzan and R. White. 2007. Query suggestion based on user landing pages. In Proceedings of SIGIR 2007. Amsterdam, Netherland.

[6] Shen Xiaoyan, Cheng Bo, Chen Junliang and Meng Xiangwu. 2008. An Effective Method for Chinese Related Queries Recommendation. In Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing 2008. SNPD ‘08. 6-8 Aug. 2008. pp. 381 - 386.

[7] He, Qi, Dakin Jiang, Zhen Liao, S. Hoi, Kuiyu Chang, Ee-Peng Lim, and Hang Li. 2009. Web query recommendation via sequential query prediction. IEEE 25th International Conference on Data Engineering, 2009. ICDE'09, IEEE, 2009. pp. 1443-1454.

[8][**Nee10**] A. K. Sharma, Neelam Duhan, Neha Aggarwal, Rajang Gupta. Web Search Result Optimization by Mining the Search Engine logs. Proceedings of International Conference on Methods and Models in Computer Science (ICM2CS-2010), JNU, Delhi, India, Dec. 13-14, 2010.

[9] [**Sri96**] Spirant R., and Agawam R. “Mining Sequential Patterns: Generalizations and performance improvements”, Proc. of 5th International Conference Extending Database Technology (EDBT), France, March 1996.

[10] **[Bor98]** A. Birchers, J. Her locker, J. Konstantin, and J. Riel, “Ganging up on information overload,” Computer, Vol. 31, No. 4, pp. 106-108, 1998.

[4] [**Ame2000**] B. Amen to, L. Tureen, and W. Hill, “Does Authority Mean Quality? Predicting Expert Quality Ratings of Web Documents”, In Proceedings of 23th International ACM SIGIR, pp. 296-303, 2000

[5] **[Bee2000]** Beeferman and Berger A., 2000. Agglomerative clustering of a search engine query log. In Proceedings of the 6th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, (August). Acme Press, New York, NY, 407–416.

[6] [**Zha2000**] D. Zhang and Y. Dong, “An Efficient Algorithm to Rank Web Resources,” In Proceedings

of 9th International World Wide Web Conference, pp. 449-455, 2000.

[7] [**Wen01**] J. Went, J. Mie, and H. Zhang. Clustering user queries of a search engine. In Proc. at 10th International World Wide Web Conference. W3C, 2001.

[11][**Bem01**] Bernard J. Jansen and Undo Pooch. A review of web searching studies and a framework for future research. J. Am. Soc. Inf. Sci. Technol., 52(3):235–246, 2001.

[12] **[Ara01**] A. Aras, J. Cho, H. Garcia-Molina, A. Peace, and S. Raghavan, “Searching the Web,” ACM Transactions on Internet Technology, Vol. 1, No. 1, pp. 97-101, 2001

[12] [**Her01**] M.R. Her zinger, “Hyperlink Analysis for the Web,” IEEE Internet Computing, Vol. 5, No.1, pp. 45-50, 2001.

[13] [**Bha02**] K. Bharat and G.A. Michaela, “When Experts Agree: Using Non- Affiliated Experts to Rank Popular Topics,” ACM Transactions on Information Systems, Vol. 20, No. 1, pp. 47-58, 2002.