Literature Survey on Improving the Discovery of User Search Goals using Click through Data

Parth Patel[1] Jignesh Vania[2]

[1]M.E Student [2]Assistant Professor

[1,2]Department of Computer Engineering

[1,2]L. J. Institute of Engineering & Technology, Ahmedabad, Gujarat, India

Abstract— Search engines play an important role in retrieving and arranging relevant data for various purposes in today's e-world. Internet is highly used to get information and satisfy user needs in these days. But, in real the relevancy of results given by search engines are still disputable because of its results having huge amount of irrelevant and unnecessary results and user information needs are not satisfied on submission of uncertain query to search engine, because different types of user needs different information for a same query fired to a search engine on various facet. So discovery of different user search goals becomes complicated. Web mining is sufficient and powerful research area from which retrieval of relevant information from the web resources can becomes faster and better. To improve user experience and search engine relevance the evaluation and illustration of user search goals can be very useful. Web usage mining is very important in deciding user search goals. Different user search goals can be discovered by analyzing user query logs from various search engines. Different user search goals from click through logs for a query can be clustered and the user feedback sessions can be useful to discover different user search goals and restructure the web search results.

Key words: User Search Goals, Time Sensitive Queries, binning

I. Introduction

The World Wide Web has become an important source of information and services and it is very popular and interactive. The web is enormous, varied and active. As the web is growing very quickly, the users get easily lost in the hectic structure. The basic goal of a search engine is to provide useful information to the users as per their needs. Therefore, retrieving the needs of users and finding their needs have become very important. We can study user's search behavior by search log analysis of the user from the search engine. The data can be billions of queries daily for a popular search engine. By use of client-side plug-ins large amount of browser log data also can be collected. To improve search result mining of this huge amount of search and browser log data is needed. The challenge is to create efficient and effective techniques to clean, process and model the log data. Whenever a user performs a search by firing a query and clicks a URL from the search result the contents of that page are extracted. The combination of clicked url, contents that are extracted and query of the user are stored in the server log. So when the next time user enters the query on the search engine the output is compared with the data in server log and ranking is done accordingly so that users can easily reach the goal what they are looking for.

Extraction of useful information from server logs is web usage mining. It can be used to find out what people are looking for on internet. A click through data is the combination of clicked and unclicked urls from a particular search. So by use of Web Usage Mining the interesting patterns can be discovered and the needs of web based applications can be served better. Usage data extracts the behavior of users browsing on internet. So use of click though data and search query can lead to the interest of users and the goal text of the users by applying different techniques of clustering for different types of data The ranking of pages in result are based on content and keywords.

The goal of search engines is to provide relevant information to the users to cater to their needs. Hence, finding the content of the Web and retrieving the users' interests and needs has become increasingly important now. In web search applications, queries being submitted to search engines are to represent the information needs of users. But, sometimes queries may not be able to exactly represent users' specific information needs because many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same type of query. The system attempts to improve the search results by inferring user search intentions, and removing incorrect or limited information problems.

II. Literature survey

A. Enhancing Search Result Delivery Using Web Content Mining and Web Usage Mining [1]

This paper uses a weighted technique to mine the web content based on the user needs.

– User request: Search engines process user requests and produce search results. These search results are then sent for pre-processing.

– Data Pre-processing : Data pre-processing step improves the quality of data by removing the dirty, incomplete and inconsistencies in the data, thereby improving the efficiency and accuracy

– Parameters Calculation: Parameters like terms frequency, occurrence positions need to be computed

– Page Relevance Computation: The user query is checked with the related words. Every result of the keywords and content words are compared by full word matching. If a match is found then a point is awarded to each words based on their position using weighted technique. Finally all matched keywords and contents words are summarized and normalized so that the total must be less than or equal to 1. At last, the normalized value of each result is sorted in descending order to get the most relevant content for the user query. Re-ordered results are sent back to the user so that the top most page is more relevant for the user query.[1]

The proposed approach is to organize search results by aspect learned from user click through logs. Given an input query the general procedure of the approach is:

- User Query will be pre-processed to identify the root words.
- When any query will be entered for the first time and no matching urls will there in user click-through logs the search results will be displaced by weighted ranking approach in the web content mining.
- If the user query and the particular url will present in the query log then the search results will be displayed according to the rank of the corresponding urls for that query. The proposed method over comes the limited information problem and improves the performance by inferring user search goals.

B. A Novel Approach To Discover User Search Goals Using Clickthrough Data [2]

The proposed approach used in this paper aims to discover user search goals/intents by clustering pseudo documents.

The user clicked links are used as relevance judgments to evaluate search precision since click-through data can be collected in an inexpensive manner; it is possible to do its large scale evaluation. All the clicked and unclicked urls before the last clicked url in a session are considered to be in one feedback session. These visited links represent positive feedback and unclicked links represent negative feedback. These feedback sessions are used to infer the user search goals.

To obtain rich information, each url is enriched with additional text content by extracting the title and snippet.

Psuedo documents are then clustered by k-means algorithm. After applying k means algorithm for clustering, each cluster represent a user search goal or user intention. The web search results are then reorganized based on the discovered user search goals/intents. Average precision and voted average precision are used to evaluate the performance of restructured web results[2].

C. Web Usage Mining and Web Content Mining – A Combine Approach for Enhancing Search Result Discovery [3]

The paper highlights the problems faced for the optimization of the search engine performance. The problems it takes into account are:

- Incomplete or Limited Information Problem: A number of heuristic assumptions are typically made before applying any data mining algorithm; as a result some patterns generated may not be proper or even correct.
- Incorrect Information problem: Even when a web user is lost, the clicks made by the user are recorded in the log, and many mislead future recommendations. It becomes a havoc when a website is badly designed and more people end up visiting such unsolicited pages, making them seem more and more popular.
- Persistence Problem: When a new pages are added to a web site, because they are not visited yet, the recommender system may not recommend them, even though they could be relevant Moreover, the more a page is recommended, the more it may be visited, thus making it look popular and boost its candidacy for future recommendation.
- Incorrect recommendation: Since what user cares about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical.[3]

The basic aim spot lights the organization of search results by aspect learned from user click through logs. The Basic algorithm is:

- User Query will be pre-processed to identify the root words.
- All the feedback sessions (The feedback session is defined as the series of both clicked and unclicked URLs) of the query will be extracted from user click- through logs.
- Resulted feedback sessions will be mapped to pseudo documents.
- User search goals will be inferred by clustering these pseudo documents and depicted with some keywords.
- When any query will be entered for the first time and no matching urls will there in user click-through logs the search results will be displaced by weighted ranking approach in the web content mining.
- Evaluate the performance of restructuring search results.

This system improves the results by inferring user search intentions, thereby removing incorrect or limited information problems.

D. Mining Sub Query Topics from Search Log Data [4]

The paper focuses on clustering algorithm that can effectively leverage the two phenomena to automatically mine the major subtopics of different queries, where each subtopic is to be represented by a cluster containing a number of URLs and keywords. The mined subtopics can then be used in multiple tasks in web search and they are evaluated in aspects of the search result presentation such as clustering and re-ranking. It introduces two phenomena:

1) One Sub Topic per Search:

One subtopic per search (OSS) means that the jointly clicked URLs in a specific search are likely to represent the same subtopic.

2) Subtopic Clarification by Additional Keyword:

The phenomenon can be explained in the following ways:

- Search users are rational users.
- Sometimes users tend to add additional keywords to specify the subtopics in their minds[4].

The short query is referred to as the original query, and longer queries containing the short query as expanded queries. The clicked URLs after searching with the original query and the expanded queries tend to represent the same subtopic. The keywords can also become labels of the subtopic. This is the phenomenon of subtopic clarification by additional keyword.

3) Clustering method used:

- Preprocessing:

All the queries are indexed in the prefix and suffix tree. In the prefix tree, query ‘Q’ and its expanded queries ‘Q+W’ are indexed in a father node and child nodes respectively. Search log data of each query is also stored in its node. False expanded queries are removed from prefix and suffix trees. If a query does not have URL overlap with its expanded queries, then those expanded queries will be viewed as false expanded queries and pruned from the trees. Clustering is performed on the clicked urls for each query and expanded query. The specific algorithm is as follows:

Step 1: Select one URL and create a new cluster containing the URL.

Step 2: Select the next URL *ui*, and make a similarity comparison between the URL and all the URLs in the existing clusters. If the similarity between URL *ui* and URL *uj* in one of the clusters is larger than threshold θ, then move *ui* into the cluster. If *ui* cannot be joined to any existing clusters, create a new cluster for it.

Step 3: Finish when all the URLs are processed.

The output of the clustering process is clusters of URLs for each query and its expanded queries. The clusters which consist of only one URL are excluded. Each cluster represents one subtopic of the query. Further keywords from the expanded queries are extracted and assigned to the corresponding clusters as sub-topic labels. As a result, each cluster not only consists of URLs but also retains keywords as cluster labels. The subtopic popularity can be further estimated from the frequency of clicked URLs in each cluster. Finally, this method can only be applied when there is enough search log data, which is also a drawback for most log mining algorithms. How to apply the approach in tail queries is also an issue that needs to be considered.

E. User Intentions Modeling In Web Application Using Data Mining [7]

For deriving the user intentions two kind of linguistic features in the text are considered: keyword and concept feature.

A keyword feature is a single word extracted from the text, which may be stemmed and stop-word excluded. WordNet is used to extract the concept hierarchy of each keyword and select the most representative one as the concept feature by means of Association Rules. At the keyword level of feature extraction, the text part is parsed such that all words are extracted from the sentences and are stemmed with stop-word excluded.[5] Each keyword is a feature and will be added to the keyword feature set.

For user intention modelling each user action record contains a text part and a tag of action as well as other important information that may reflect the user’s intention. The XML format is adopted when user’s log data is recorded. The obtained intention model for the user is used to predict the user’s intention in the future.

III. CONCLUSION

The paper presents a survey on different aspects of the work done till now in the field of web usage mining and clustering the search results for re ranking based on relevancy. The traditional systems were not using the log data of users in generating search results. But now in recent the methods to improve the results of search by re ranking the results of search using the log data and Clickthrough data and clustering to different data are introduced. And there is lot more improvement can be done by using different clustering methods and different data. Hence, user log data is of high importance. By, understanding the issues faced by the current systems, more improvements can be done in the field of improving search results by clustering the feedback sessions using log data. So, users can find exact information needed as they want very efficiently.

REFERENCES

[1] Ms. Shital C. Patil, Prof. R. R, Keole,“Enhancing Search Result Discovery Using Web Content Mining and Web Usage Mining”, H. P. V. M’s College of Engineering & Technology, IJSRM, Volume 2, Issue 1, Pages 496-500, 2014

[2] Charudatt Mane, Pallavi Kulkarni,”A Novel Approach To Discover User Search Goals Using Clickthrough Data”, IJCSIT, Volume 5(1), 20-24, 2014.

[3] Miss. Shital C. Patil, Prof. R. R. Keole,”Web Usage Mining And Web Content Mining – A Combine Approach for Enhancing Search Result Discovery” IJARCSSE, Volume 3, Issue 10, Oct 2013.

[4] Yunhua Hu, Yanan Qian, Hang Li, Daxin Jiang, Jian Pei. Quinghua Zeng,”Mining Query Subtopics from Search Log Data” SIGIR’12, ACM, August12-16, 2012.

[5] Zheng Chen, Fan Lin, Huan Liu, Yin Liu, Wei-Ying Ma, Liu Wenyin “User Intention Modelling in Web Applications Using Data Mining”,Internet and Web Information Systems, 5, 181 – 191, 2002.