Significance of Search Logs in Crawling

Karishma
Asst.Prof.
J B Knowledge Park,Faridabad
Haryana,India

Abstract

On most sites, users look to on-site Search after they've scanned the page for clues to the content they're seeking. A quality Search experience is critical to making users happy. With clever instrumentation and reporting, you can put together benchmarks for your search that will help identify problems quickly and will make search more efficient and fast. The place to start is the Search Log. How hard is search? How much does personalization help? All the answers lie with Search Logs. Effective organization of search results is critical for improving the utility of any Crawler .In this paper we discuss the significance of search logs on crawlers. Using the contents of Search logs crawler can refine the search and retrieve data more quickly and personalize the search according to user needs and interests.

Keywords
Search logs, Personalization, Crawlers, Spiders.

1. Introduction

The utility of a Crawler (Search engine) is affected by multiple factors [1]. While the primary factor is the soundness of the underlying retrieval model and ranking function, how to organize present search results is also a very important factor that can affect the utility of a search engine significantly. Compared with the vast amount of literature on retrieval models, however, there is relatively little research on how to improve the effectiveness of search result organization. Search engine employs the strategy of ranking the searched results and rank top the most suitable and relevant results [3]. However, when the search results are diverse (e.g., due to ambiguity or multiple aspects of a topic) as is often the case in Web search, the ranked list presentation would not be effective. It is by analyzing search patterns via our logs data search results can be optimized.

2. Search logs

Search engine logs [11] record the activities of Web users, which reflect the actual users needs or interests when conducting Web search. They generally have the following information: text queries that users submitted, the URLs that they clicked after submitting the queries, and the time when they clicked. Search engine logs are separated by sessions. A session includes a single query and all the URLs that a user clicked after issuing the query. A small sample of search log data is shown in Table 1. The process of recording the data in the search log is relatively straightforward. Web servers record and store the interactions between searchers (i.e., actually Web browsers on a particular computer) and search engines in a log

Table 1: Sample entries of search engine logs. Different ID's mean different sessions

ID	Query	URL	Time
1	win zip	http://www.winzip.com	xxxx
1	win zip	http://www.swinzip.com/winzip	xxx
2	time zones	http://www.timeanddate.com	xxxx

file (i.e., the transaction log) on the server using a software application. Thus, most search logs are server-side recordings of interactions. Major Web search engines execute millions of these interactions per day. The server software application can record various types of data and interactions depending on the file format that the server software supports.

Search logs are records of the user requests for information from our index[12]. We can generate and export this information, and then use it as input to our preferred log analysis software or reporting software.

For Web searching, a search log is an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine

These are some examples of the information that search logs can provide:

- What types of queries are users making?
- How fast are users being served?
- Are users getting the results they need?
- Do you need to help users find relevant information by configuring the Related Queries, Key Match, Query Expansion, or One Box features?

On the basis of this information prioritization of links is done and when user further makes query about the related topic more relevant information is retrieved.

3. ANALYSIS OF SEARCH LOGS TO IMPROVISE SEARCH

A number of studies have indirectly compared successful and less successful search strategies by comparing expert and novice searchers in lab studies. Recently, researchers have also begun to use data from search engine logs to identify metrics that are related to users' search success. These studies have provided some promising findings, but the noisiness of the log data makes it hard to determine if the searchers were successful or not and which signals are specific to which kinds of tasks. Next we discuss improvisation of search based on following factors[13].

3.1 Improvisation of Search Based on Content Analysis[7]

One approach of improving search is to filter or rerank search log by checking content similarity between returned web pages and user profiles. User profiles store approximations of user interests. User-issued queries and user-selected snippets/documents are categorized into concept hierarchies that are accumulated to generate a user profile. When the user issues a query, each of the returned snippets/documents is also classified. The documents are reranked based upon how well the document categories match user interest profiles.

3.2 Improvisation of Search Based on Query Analysis

Queries issued can be categorized into clear queries, semiambiguous queries, and ambiguous queries.It can be concluded that quality of search results significantly increased output quality for ambiguous and semiambiguous queries, but for clear queries, one would prefer a common Web search. According to one survey queries can be divided into fresh queries and recurring queries. They found that the recent history tended to be much more useful than the remote history, especially for fresh queries, whereas the entire history was helpful for improving the search accuracy of recurring queries

3.3 Improvisation of Search Based on Hyperlink Analysis[9]

Most generic web search approaches rank importance of documents based on the linkage structure of the web. An intuitive approach of personalized web search is to adapt these algorithms to compute personalized importance of documents. A large group of these works focuses on personalized PageRank. The fundamental motivation underlying PageRank is the recursive notion that important pages are those linked-to by many important pages. This recursive notion can be formalized by the

"random surfer" model on the directed web graph G. A directed edge <p, q> exists in G if page p has a hyperlink to page q. Let O (p) be the outdegree of web page p in G. O(p) is equivalent to number of web pages that linked by page p. Let A be the matrix corresponding to the web graph G, where Aij = 1/O(j) if page j links to page i, and Aij = 0 otherwise. In the random surfer model, when a surfer visits page p, he/she keeps clicking outlinks at random with probability (1-c), and jumps to a random web page with probability c. c is called teleportation constraint or damping factor. The PageRank of a page p is defined as the probability that the surfer visited page p. Iterative computation of PageRank is done as
the following equation:

vk+1 =(1-c)Avk +cu

Here, u is defined as a preference vector, Where |u| = 1 and u(i) denotes the amount of preference for page I when the surfer jumps to a random web page i.The global PageRank vector is computed when there is no particular preference on any pages, i.e., u =[1/n,...,1/n]T. By setting variant preference to web pages, a PageRank vector with personalized views of web page importance is generated. It recursively favors pages with high preference, and pages linked by high-preference page. This PageRank vector is called a personalized PageRank vector (PPV). To accomplish personalized web search, a personalized PageRank is computed for each user based upon the user's preference. For example, web pages in the user's bookmarks are set higher preferences in u. Rankings of the user's search results can be biased according to the user's Personalized PageRank vector instead of the global PageRank.Unfortunately, computing a PageRank vector usually requires multiple scans of the web graph, which makes it impossible to carry out online in response to a user query. Furthermore, when a large number of users employ a search engine, it is impossible to compute and store so many personalized PageRank vectors offline. Many later works make efforts to reduce the computation and storage cost of personalized PageRank vectors.

3.4 Improvisation of Search Based on Community

In most of the above personalized search strategies, each user has a distinct profile and the profile is used to personalize search results for the user. There are also some approaches that personalize search results for the preferences of a community of like-minded users. These approaches are called community-based personalized web search or collaborative web search In a community-based personalized web search, when a user issues a query, search histories of users who have similar interests to the user are used to filter or re-rank search results. For example, documents that have been selected for the target query or similar queries by the community are re-ranked higher in the results list.

4. PERSONALIZING THE SEARCH EXPERIENCE

Rather than providing one centralized search experience for everyone, you can provide different search experiences for different groups of users. Each personalized search experience is based on the interests[4], roles, departments, locations, or languages of the user group.

Users often search for "acme widgets," but they are not all searching for the same results. More typically, when searching for acme widgets:

Engineering staff is searching for design documents and status information
Sales staff is searching for sales forecasts and reports
Customer support is searching for support metrics and update information

With a centralized search experience, some users may find what they are looking for at the top of the results listings while other users might have to view several results before finding what they are looking for. With a personalized search experience[5]:

Each group of users has a unique search experience where results are ranked according to their interests

Users find what they are looking for at the top of the search results.

5. APPROACHES FOR PERSONALIZATION: RELATED WORK

The basic approach for personalization is:

a) Build User model that shows his interest through Click history.
b) Take any learning Strategy for analyzing this History.
c) After analysis ranking mechanism or any other categorical method is applied an search log to personalize search.

To realize personalized web search, search engine needs to make different semantic expansion of the users' queries based on users' personalized information which will affect the users' whole search process. A common method to achieve personalized web search is to construct a user profile model based on user's personal information and then the user profile is applied to influence the user's search results. A variety of methods can be used to construct user profile model according to the personal information collected from users. Different methods not only have different characteristics and performances, but also have different effects on personalized search results. First, the user profile can be described as a set of rules. Second, the user profile can be described as a set of key words. A web page can be determined its correlation with the user interest by calculating the distribution of these key words in it. Third, the user's personalized information is represented as a vector space. After transforming a web page into a corresponding vector in the vector space, the page can be identified if it is relative with the user profile. Fourth, the user profile is represented as a probability table of user interest and its corresponding key words. This table can be used to determine the probability of relevance between user interest and the web page. Vector space model can be flexibly used to express user profile, moreover many information retrieval and machine learning methods are directly based on vector space model.

Three user profile modelling [8] methods are proposed to realize web personalized search. The methods include Rocchio method, k-Nearest Neighbour method and Support Vector Machines method. To measure and compare the search performances of these three methods, a domain dataset is also constructed. Then based on this dataset, the user profile modeling methods are tested. Experimental results [2] show that k-Nearest Neighbor and Support Vector Machines method perform as well. In addition, kNN method has better robustness and can be easy to use. Therefore, kNN method is a better way to construct user profile model for web personalized search.

Another Approach is to personalize the search based on User Behaviour [3]. Because users are difficult to express effective demand and difficult to be analyzed, so these make personalized search engine difficult to collect the users' personalized information, Instead of letting the user to express his needs, let us analyze the uses' behavior from the history use, then finds effective personal information. Therefore, the search system should be able to without user intervention, directly and accurately detect the users' direction of interest, allowing users to look up the information more accurate, more in line with their needs. With the development of network, online advertising has increased from no difference for all users towards the direction of targeted player. Behavior Targeting ad model is generated for users' interest.Tracking the history page of users visited during a certain period, then based on the content, time, frequency, whether set to collection to determine the user's interest. Even under long-term follow-up to confirm the user's long-term interest, short-term interest and the change of the interest, adjust the ads timely, so that these make users always see something they feel interesting. We also can use this idea to the search engine to collect the personalized information. Of course, users' interest often more than one, also not static; so these require us to improve the algorithm continuously to achieve the best possible condition. As the complexity of search engine technology, so it is not enough we collected the user's interest in a similar way, we also need to according to the user's personal interests to give targeted feedback.

There also exist approaches for personalization based on the User preferences, user interest [10] etc.

6. CONCLUSION

We have discussed the effect of using the content of search Logs in crawling .We can utilize Search log data to effectively retrieve information from web. Search results can also be personalized with search log data by collecting User preferences or interest or behavior etc. In closing we can conclude that using search log we can personalize the Search to effective and relevant retrieval of Information from web and there can be various other methods to analyze the search Log.

7. References

[1] A. Kritikopoulos, and M. Sideri, "The Compass Filter: Search Engine Result Personalization using Web Communities," Lecture Notes in Computer Science, v 3169, p 229-240, 2005.

[2] Chunyan Liang, "User Profile for Personalized Web Search" in the proceedings of 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)

[3] Sugiyama K. Studies on Improving Retrieval Accuracy in web Information Retrieval [0]. Tokyo: Nara Institute of Science and Technology, 2004.

[4] Ding Zhenfan, Deng Lei. personalized information recommendation service and personalized search engine (Personalized Information Recommendation Service and Personalized Search Engines). Software space ,2009-12 ,205-206

[5] Lee Aoki, Cui North light. Based on personalized information recommendation services, Web search engine technology Summary of Information . 2007-8,98-101.

[6] Seher, I. Ginige, A. Shahrestani, S.A, "A personalized query expansion approach using context", 3rd IET International Conference on Intelligent Environments, 2007.

[7] Wang, G.T. Xie, F. Tsunoda, F. Maezawa, H. Onoma, A.K., "Web search with personalization and knowledge", IEEE International Conference on Multimedia Software Engineering, 2002.

[8] Hany M. Harb, Ahmed R. Khalifa, Hossam M. Ishkewy, "Personal Search Engine Based on User Interests and Modified Page Rank", July 1 2009.

[9] Fang Liu, Weiyi Meng, "Personalized Web Search for Improving Retrieval Effectiveness", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 1, January 2004.

[10] Li, S.Q. and Han, Z.Y. 2008. Principles & Technique of Personalized Search Engine. Science Press,2008.

[11] Yuan Hong, Xiaoyun He, Jaideep Vaidya, Nabil Adam, and Vijayalakshmi Atluri. Effective anonymization of query logs. In CIKM, 2009.

[12]Michaela G¨otz, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui Xiao, and Johannes Gehrke," Publishing Search Logs – A Comparative Study of Privacy Guarantees" in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING in 2011.

[13] Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately.In WWW, 2009.