Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 6, No. 2, May 2015), Pages: 65-72
www.jacr.iausari.ac.ir

Personalization of Search Engines, Based-on Comparative Analysis of User Behavior

Shekoofe Bostan[✉1],Mohamad Ghasemzadeh[2]
1) Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Yazd, Iran
2) Department of Electrical and Computer Engineering, Yazd University, Yazd, Iran
s.bostan@ysrbiau.ac.ir; m.ghasemzadeh@yazd.ac.ir

Received: 2014/05/11; Accepted: 2014/07/14

Abstract

In this research work, the impact of user's behavior on search engine results is discussed. It aims to improvement of search results which leads to the higher satisfaction of users. In other words, we are trying to present a personalized search engine for each user, based on his/her activity and search history. We base our hypothesis that the search history of each user in a specific time frame provides precious information to be used in case of offering customizable and more efficient results to the user. In order to evaluate our research project hypothesis, we designed and implemented an experimental search engine as a web platform. This search engine measures the level of user's satisfaction by bringing putting extra information into account. According to the experimental results, consideration of the user's behavior history has significant effect on the quality of the search results, leading to more satisfaction of the users.

Keywords:*Customized Search Engine, User's behavior, Information Retrieval, Search history*

1. Introduction

Information Retrieval includes a set of standards for storage and indexing information and also information extraction according to the user's query. Information Retrieval as shown in Figure1 can be divided into two areas: User Area and information Area. Also there is a middle area named Retrieval that matches the user's information needs and information documents [1].

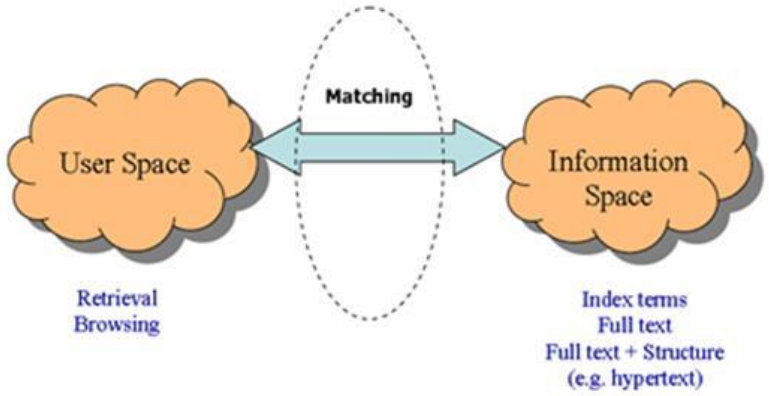

Figure 1. Overview of Retrieval System [1]

The Search Engines are the most important systems in information retrieval but the important issue is that search engines provide results that are most relevant to the user query and Web documents. Also the most important part of search engines is ranking that represent the quality of pages by search engine.

When the user enters a query as shown in Figure2 provided a list of sorted pages to user that more attention is concerned to the results of first page.

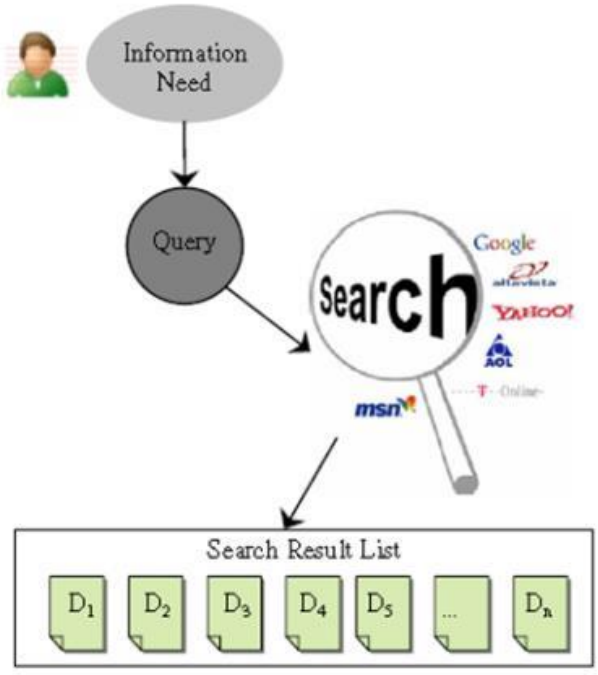

Figure 2. The process of user interaction with the search engine [2]

Various methods are used to information retrieval which are mainly based on the content and structure, also different algorithms are used for this purpose. For example, google used PageRank algorithm that is based on structure and each page ranking is calculated according to the link between pages [3].

2. Search Engine Architecture

Creating a search engine consists of three phases crawling, indexing and searching.

2.1. Crawler

A crawler is a program that visits Web sites and gathering their information to create search engine index. First crawler was designed by Fletcher. After a while, a more advanced version of it was presented. Search engines use different crawlers to achieve fresh and quality pages [4].

2.2. Indexing

In this part, stored documents are processing and indexing. To indexing, there are several ways that the most common method is extracting keywords from the text normalization. The inverted index data structure is a good way for indexing words of documents to optimize the speed of the query. Extracting keywords from the text is really important to represents the document and its contents. According to studies, at least 3.78 billion web pages indexed is exist [5]. Due to the high volume of web pages Web documents are divided into several indexed, the results are then merged [6].

2.3. Search

This section constitutes the core of the search engine because It includes models for information retrieval and ranking algorithms. In fact, this section is important in the efficiency of the search engine, according to these algorithms Page Rank is calculated and documents are sorted. In the other hands, when user enters the query, this session task is to search in stored documents in previews session and ranking document based on its content [7].

3. Related Work

Initially, the Internet was included ftp sites to only download or upload files. The first application of Internet search tools was Archie that named Father of all search engines. Archie was built by Alan Emtage in 1990 in Montreal McGill University [8].In 1993, to retrieve documents and texts in gopher servers, veronica was introduced in university of Nevada [9]. Google in 1997 and yahoo in 1994 was developed at Stanford University.Around 2000, Google's search engine rose to prominence by used PageRank algorithm [10]. Yahoo was providing search services based on Inktomi's search engine but switched to Google's search engine in 2004[11]. In 2005, Google began personalizing search results for each user, Depending on their history of previous searches, Google crafted results for logged in users. In fact, Google Custom Search allows anyone to create their own search engine. Search engines used user profiles to provide personalized search results. When a user is logged into a Google account and start to search in google search engine, all of his searches are recorded into Google's web History. Then, whenever the user attempts to search, the search results are not only based on his query, But according to the pages the user has seen in the past and they were interested,To provide a more relevant results. On December 4, 2009, Personalized Search was introduced to all users of Google, such as those who are not logged into their Account [12].

4. Mahak dataset

In this paper, the standard Mahak[13] dataset available from the research team of Tehran University and Iran Telecommunication Research Center was used. To prepare the Persian Mahak dataset for our platform, all documents in were collected and pre-processed. Document’s content such as address, title, anchor, body and other sections of the document are extracted and stored with TREC standard. The Unicode used for the data storage is utf8 as far as the character set of the documents is Farsi (Persian Language). Each document is given a unique number that represents the document. This dataset consists of several parts, which in this study WebIR files containing XML files are used.

5. Data Preparation

In this study, the test was analyzed thousands of pages of documents. Figure3 shows the data preparation process.First of all words in this text are extracted. All extra characters are removed from the text. Finally, based on the distance, words are recognized and based on page numbers, words are indexed.

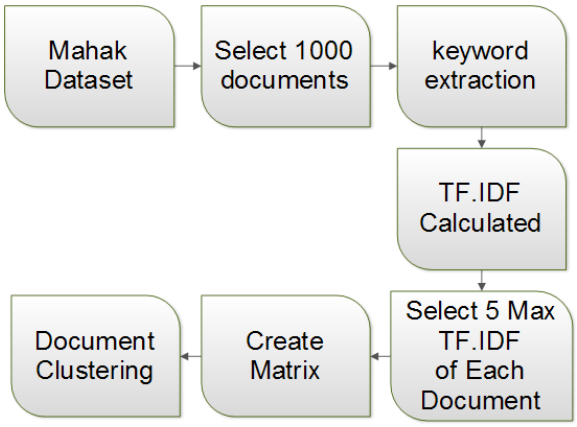

Figure 3. Data preparation

In order to calculate the weight of the words, TF.IDF model was used. Each word represents the contents of a document. After calculating the weight of the words within each document, According to Vidal et al. [14] words are arranged in descending, then 5 first word in each document, is stored as keywords. Reiterative words are removed, finally the vector of Words is computed for all documents. In order to clustering of documents, various techniques are available. The most common approach is to use K-Means algorithm. K-Means clustering is based on minimum distance of each data center. At this point number of clusters to 50 has been considered.

6. Laboratory Search Engine

In this Research, create and design of search engine is considered. As shown in figure 4 we need a user interface between user and documents. To designing this user interface, ASP.net programming is used. Each user creates a user name and password to register. In other words, user has to login to be able to search in this system. We use this method to be able to tracing all users to determine their behavior for better search result. All users' query will be stored in each session. When user types a query, the search begins. At this point, user's query words will be searched in document files to represent the results to the user. In order to better understand the search process, Results of each step is presented to the user as a search model. In this study, five search models, starting from the principle concept of search (plain search) and ending to the personalized search based on the user behavior, are presented to the user. Each model follows the characteristics of the previous model and its own extra features. This progressive design aims to show the differences, applicability and advantages of having different features and characteristics contributing in the search process to better understanding the elements leveraging towards higher satisfaction of users for personalized search based on behavior.

Figure4. BSN search engine

6.1. First search model

The purpose of this model is presented of results returned by a search engine without any ranking method. User behavior History has not been used yet. This model is very simple in which the basic search concept (plain search) is implemented. It just find documents that includes user's query.

6.2. Second search model

In this method, in addition to the user's query words, Behavioral history is also used. So in each search query, last four words of User History table is extracted, then with the user's current query words are searched in the table Info. This method returns documents more than previous methods.

6.3. Third search model

In this method clustering the documents in the file are used. In the other hands, after user enters his/her query, According to previous methods, Info files are searched and Documents that contain the keywords are determined. Then cluster for each document is identified. Finally, all documents within that cluster are extracted. It is due to existence of similar documents in each cluster.

6.4. Fourth search model

To improve on the third model, this model by adding four recent query of user continues. In this model, in addition to the user's current query, last four words of User History is used. Documents that contain the keywords are determined. Cluster for each document is identified. Finally, all documents within that cluster are extracted. The

results of this search model, in comparison with other models proposed the highest numbers of documents.

6.5. Advanced Search model

In this model at first query keyword are searched on info table, then cluster for each document is identified. All documents within that cluster are extracted. All the words in the documents are extracted. Then a new table of document id and the number of occurrences of each word comes from the user's recent queries. Finally, Total repetition of words for each document as document ranking is calculated. In order to rank documents and relevant documents based on the priority, the documents are sorted in descending order according to their rank. Search results of this model is more limited than other models and related.

To evaluate the proposed model Asked 10 friends to register as a user in the system. Duration of the process was considered a week. In order to evaluate the proposed models after each Search survey will be activated. as shown in Figure5 Users must see the results of each model, and rate the model by number between 0 to 10 to show their satisfaction of each models.

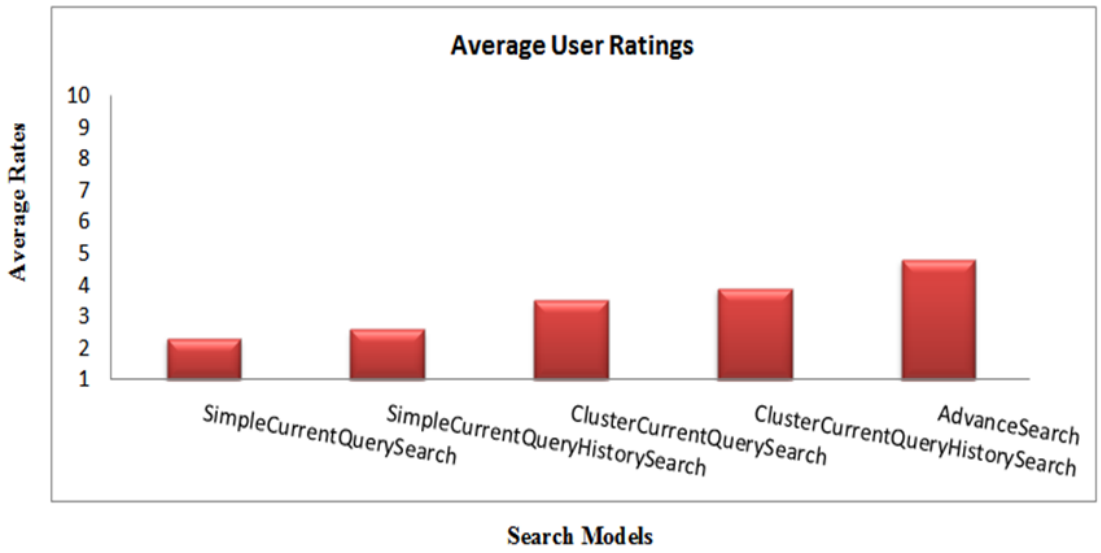

Figure5. User Rating

Also, to verify the accuracy of existing models, we used the precision measure in Figure6 that Calculate the number of relevant documents retrieved divided by the total number of retrieved documents can be obtained.

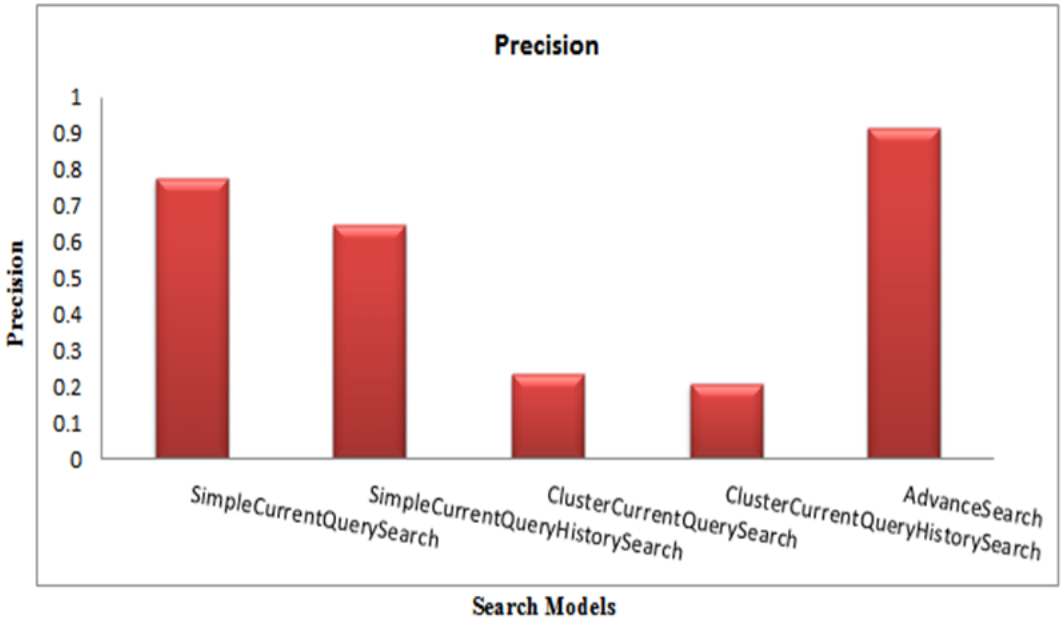

Figure6. Precision

7. Conclusion

In this paper the design and implementation of a search engine laboratory has been mentioned that it can be used in different studies. In fact, this search engine can serve as a tool for those interested in the field of information retrieval and search engine to better understanding of the features and parameters contributing in the search engines based on the query history and user behavior. This study examined the impact of user behavior to improve search results in a time series routine. Experimental results indicate that considering users' interests in Search and Customized Search Engine, Can achieve results more closely to the needs of each user. Though in some cases the strategy shows lower precision due to the fact that not always the search queries are consistently followed in a time series, out final model shows both high precision and user ratting comparing to the other models. Eventually, the results released that considering the user behavior and different features such as cluster analysis on the document content (not only the title or the Meta tags in the header) can reveals better performance which results to the higher user satisfaction. This study can be also highlighting that not always structural and routine search procedures but also heuristic search models can be also offer high efficiency in the contexts when the intelligent data mining and information retrieval comes to the play.

8. References

[1] R. Baeza-Yates, B.Ribeiro-Neto,*"Modern information retrieval", ACM press*, 1999.

[2] Y. Liu, J. Miao, M. Zhang, S. Ma, L. Ru, "How do users describe their information need: Query recommendation based on snippet click model", *Expert Systems with Applications*, 2011.

[3] M. Sanderson, W. B. Croft, "The history of information retrieval research", *Proceedings of the IEEE*, 2012.

[4] S. Brin, L. Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer networks and ISDN systems*, 1998.

[5] M. d. Kunder, Retrieved from WorldWideWebSize.com,Daily Estimated Size of the World Wide Web: http://www.worldwidewebsize.com/, 2013.

[6] Z. G.Gonzalez,M. Kelly, T. E. Murphy Jr, M. Nisenson,"Search Engine Indexing", *Patent Application*, 2012.

[7] W. B. Croft, D. Metzler,T. Strohman , "Search Engines: Information Retrieval in Practice", *Information Processing & Management*, 2010.

[8] A. Emtage, P. Deutsch, "Archie: An electronic directory service for the internet". In *Proceedings of the USENIX Conference*, 1992.

[9] R. Raucci, "Gophers, Web Encyclopedias, and Search Engines". In*Netscape for Macintosh*, Springer, 1996.

[10] S. Brin, L. Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer networks and ISDN systems*, 1998.

[11] U. Manber, A. Patel, J. Robison, "Experience with Personalization ON Yahoo", *Communications of the ACM*, 2000.

[12] B. Horling, M. Kulick, *Personalized Search for everyone*, Retrieved from Google official blog:http://googleblog.blogspot.co.uk/2009/12/personalized-search-for-everyone.html, 2009.

[13] E. Darrudi, H. BaradaranHashemi, A. AleAhmad, A. M. ZareBidoki, A. H. Habibian, F. Mahdikhani, A. Shakery, M. Rahgozar, "dorIR collection for Persian web retrieval", *University of Tehran*, 2010.

[14] M.Vidal,G. V.Menezes, K.Berlt, E. S.de Moura,K.Okada,N.Ziviani, M.Cristo, "Selecting keywords to represent web pages using Wikipedia information". In *Proceedings of the 18th Brazilian symposium on Multimedia and the web, ACM*, 2012.