An Integrated Page Ranking Algorithm for Personalized Web Search

J.Jayanthi
Assistant Professor
CSE Department
Sona College of Technology, Salem

Dr.K.S.Jayakumar
Associate Professor
MECH Department
SSN College of Engineering, Chennai

ABSTRACT
Today search engines constitute the most powerful tools for organizing and extracting information from the Web. However, it is not uncommon that even the most renowned search engines return result sets including many pages that are definitely useless for the user. This is mainly due to the fact that the very basic relevance criterions underlying their information retrieval strategies rely on the presence of query keywords within the returned pages. Web Search Personalization is a process of customizing the Web search experience of individual users. The goal of such personalization may range from simply providing the user with a more satisfied results by relevant information. Such a system must be able to deduce the information needs of the user. It is worth observing that statistical algorithms are applied to "tune" the result and, more importantly, approaches based on the concept of relevance feedback are used in order to maximize the satisfaction of user's needs. Nevertheless, in some cases, this is not sufficient. In this paper search results are ranked based on user preferences in content and link. The preference of content and link is integrated in order to rank the results.

Keywords
Personalization, Page ranking, Information retrieval, semantics, links, HITs.

1. INTRODUCTION
Experiments conducted by Yahoo! show that quite a large number of users would prefer personalized search over the generic search systems currently available. With the exponential growth of the internet, we also have a market for personalized web search systems. For example, a keyword search for "pen" is ambiguous. The user might be looking for the a writing implement with a point from which ink flows or a pen — female swan. Which of the category would be of most interest for a particular user? If a search system could answer such questions, information retrieval on the web would be so much more comfortable. Many users are willing to ignore privacy and security issues when choosing personalized search. Such facts show that there is worthwhile market for designing a large scale commercial personalized search engine.

Polysemy is the existence of multiple meanings for a single word. Synonymy is the existence of multiple words with the same meaning. They cause confusion in the keyword search. This is known as the vocabulary problem. It results in mismatches between the query space and the document space. Synonymy causes relevant information to be missed if the query does not contain the exact keywords occurring in the documents, inducing a recall reduction. Polysemy causes irrelevant documents to appear in the result lists, affecting negatively the system precision. Also, looking at the searching behavior of users, we know users are prone to start a session with queries that are formed easily rather than spend time in forming a better query. Users often learn the right keywords for a domain of information while actually browsing the web.

2. RELATED WORK
The impulsive growth and popularity of the World Wide Web has resulted in a substantial amount of information sources on the Internet, creating a scenario where the answers to information needs of the users are available online somewhere in some format; but in order to find the appropriate information users need to scan through endless list of digital data. Different typologies of users explore the Web in various ways according to their requirements and experiences; some users, for instance, may survey an area of knowledge to get a general understanding on it, while others to look for specific information. In either of the cases, they need to access and analyze all the documents available and this process is time consuming. For these reasons, they normally tend to compromise themselves with the information they have received. This clearly indicates the presence of information overload [3], [4]. Personalized web content [5], [6] is one of the proposed solutions to solve this problem. Moreover another feature of information available on the Web makes difficult identify opportune, automatic and effective methodologies of access and retrieval: the most part of information is present in the form of unstructured free text, written in natural languages. Examples are blogs, forum, corporate memos, research reports, emails, blogs and historical documents [7].According to recent studies more than 80% of queries submitted by users to search engines are estimated informational in nature [8]. This means that most of them could be answered properly by providing structured and normalized form of information, like to key notes of entities, price lists of items for sale, document summaries. The purpose of Information extraction (IE) is to structure the possible unstructured text; in other words, IE is the process of populating a template of structured information starting from unstructured or loosely formatted text, which can be given directly to user or can be stored in a database for further

processing [7], [9].Cuwe et al. in [10] suggested improving retrieval efficiency by tracking the user and exploring his/her logs. The authors reported that their algorithm dramatically improved the result's efficiency. They investigated the user's log files in the search engine and used them in the subsequent queries. This method directs the search engine toward common information in documents for each user.

The focal point for each user is distinct and in agreement with the log file. Haveliwala et al. in [11] investigated the possibility to find a web page relevant to a reference web page. Although the objective of the project is quite similar to this paper, it was implemented using a totally different strategy. The authors used the reference page only to represent the knowledge-based system. Whereas in this work, we demonstrated how this approach is inadequate in comparison with the positive and negative examples method we provided. Poincot et al. in [12] introduced a new approach to compare documents and calculate their similarity using machine learning. The authors showed how documents' similarities could be calculated using neural network (Kohonen maps). Chakrabartwe et al., in [13] presented an algorithm in mining the web using hub and authority's techniques to discover relevant web pages. Ahonen et al. in [14] experimented with the co-occurring text phrase and they concluded that a promising result was found. Liu et al. in [9] proposed a similar method to the one provided in this paper. The authors provided a personalized web search for improving retrieval effectiveness. They have implemented a machine learning algorithm to capture the user interests. Every time the user connects to a URL, the system keeps track of that URL and categorizes it. This process improved the overall system performance as every URL is reflected on the subsequent query.

The authors of [15] presented another approach to address the same problem which was handled in this paper. However, they introduced the learning algorithm as a mandatory and without the user awareness, and provided more evidence as in [12, 16], and [21], which advocated the advent of machine learning in web mining and information retrieval

3. PROBLEM

The problem is to personalize Web search for improving retrieval effectiveness. Our strategy includes three steps. The first step is to map a user query to a set of links. The second step is to utilize both the query and its context to retrieve Web pages using ontology.

Table 1. Document And Term Matrix

Category / Term	Oracle corporation	Relational dbms	Enterprise software	Oracle university	Oracle certification	Oracle careers
Oracle Corporation	1	0.6	0.5	0	0.4	0.5
Oracle University	0	0	0	1	0.5	0

Table 2. Document And Category Matrix

Doc/Term	Oracle corporation	Relational dbms	Enterprise software	Oracle university	Oracle certification	Oracle careers
D1	1	0	0	0	0.5	0.5
D2	0.5	0.5	0	0	0.5	0
D3	0	0	1	0	0	0.5
D4	0	0	0	1	0.5	0

In order to accomplish the first step, ODP is used as a resource. A tree model approach is used to represent a user's search history and describe how a user's search history can be collected without his/her direct involvement. The user submits a query to the search engine. The search engine produces set of results composed with the relevant and irrelevant page collections.

Table 3 . Category And Term Matrix

Doc/Category	Oracle Corporation	Oracle University
D1	1	0
D2	1	0
D3	1	0
D4	0	1

The relevant or irrelevant page identification .is a complex task to the user. Anyway, the presence of unwanted pages in the result set would force him or her to perform a post processing on retrieved information to discard unneeded ones.

Even though several automatic techniques have been recently proposed, result refinement remains a time-waste and click-expensive process, which is even more critical when the result set has to be processed by automatic software agents. The Semantic Web will offer the way for solving this problem at the architecture level. In fact, in the Semantic Web, each page possesses semantic metadata that record additional details concerning the Web page itself. Annotations are based on classes of concepts and relations among them. The "vocabulary" for the annotation is usually expressed by means of an ontology that provides a common understanding of terms within a given domain.

Relations among concepts embedded into semantic annotations can be effectively exploited to define a ranking strategy for Semantic Web search engines. This sort of ranking behaves at an inner level that is, it exploits more precise information that can be made available within a Web page and can be used in conjunction with other established ranking strategies to further improve the accuracy of query results. With respect to other ranking strategies for the Semantic Web, our approach only relies on the knowledge of the user query, the Web pages to be ranked, and the underlying ontology. Thus, it allows us to effectively manage the search space and to reduce the complexity associated with the ranking task.

4. PROPOSED ALGORITHM

Search result ranking operations are done under the search engine environment. Page links and contents are used individually for the ranking process. The semantic relations are used to analyze the contents of the web pages. The ontology is used to analyze the content relationship. Concept relationships are maintained in the ontology. The proposed system improves the ranking mechanism using the semantic relations and hyperlink relations. The semantic relations indicate the content relevancy.

The hyperlink network is used to represent content referenced by the other pages. The hyperlink also shows the related sources. In links and out links details are used to reflect the page relationship. The Hyperlink Induced Topic Search (HITS) algorithm is used for the page ranking process. The authority and hub values are estimated under the link based ranking process. The link based ranking scheme does not considers the content relationship. The content based ranking scheme does not consider the page content values. The content and link relationship is used in the proposed system for the ranking process. The search query values are prepared using the semantic information. Domain selection is used to support the search query optimization. The result page analysis operations are performed under the client environment. The cleaning process is used to remove noisy data under the web pages. The tag elements and script sources are considered as noise data. The results are ranked and irrelevant pages are removed from the result.

Table 4 .Category And Link Table

Category/Link	Oracle corporation
Oracle Corporation	www.**oracle**.com www.linkedin.com/companies/**oracle** *www.crunchbase.com* www.hidglobal.com/documents/casestudy_**oracle**.pdf www.silobreaker.com/**oracle-corporation**-11_3660330 www.orafaq.com www.oriole**corp**.com/ www.evri.com/organization/**oracle-corporation**-0x49aae www.gocertify.com/vendors/**OracleCorporation** www.**corp**orateinformation.com www.mysql.com www.indeed.co.in/**Oracle-Corporation**-jobs www.sun.com/third-party/global/**oracle** www.bnamericas.com/.../en/**Oracle_Corporation-Oracle**

Table 5.Catergory And User Matrix

Category/user1	Oracle corporation	Relational dbms	Enterprise software	Oracle university	Oracle certification	Oracle careers
Oracle Corporation	0	1	0	0	1	1
Oracle University	0	0	0	0	0	0

Links relevant to the oracle corporation and oracle university is prioritized based on term weight. Every user is mapped with category and based on their search history weight is assigned. Based on the user's preference the relevant links will be prioritized and the irrelevant links will be omitted. In the above table user 1 is more interested on Relational DBMS, Oracle certification and Oracle careers so other links will be hidden for the users. Rocchio is originally a relevance feedback method [30].We use a simple version of Rocchio adopted in text categorization:

$$M(i,j) = \text{Max}\left(\sum_{k=1}^{m} DT(k,j) * DC(k,i) \right) \quad (1)$$

where M is the matrix representing the user profile, Ni is the number of documents that are related to the ith category, m is

the number of documents in DT, DT(k,j) is the weight of the jth term in the kth document, DC(k,j) is a binary value denoting whether the kth document is related to the ith category. Clearly, M(I,j) is the max weight of the jth term in all documents that are related to the ith category and documents that are not related to the category are not contributing to M(i,j). We call it as MRocchio method .Based on the category term weight ,Category link will be prioritized. To include personalization for every user ,category term weight will be calculated.

$$MU(i,j) = \sum_{u=1}^{n} \sum_{k=1}^{m} DT(k,j) * DC(k,i) \quad (2)$$

Interested terms links will be mapped with the user and will be displayed.

5. RESULT AND DISCUSSION

5.1 Measure of Web Page Retrieval

The measure of effectiveness is essentially the "Precision at 11 standard recall levels" as used in TREC evaluation [22]. It is briefly described as follows:

- For each query, for each list of retrieved documents up to the top 20 documents, all relevant documents are identified. (In practice, a number higher than 20 may be desirable. However, we have a limited amount of human resources to perform manual judgment of relevant documents. Furthermore, most users in the Web environment examine no more than 20 documents per query.)
- The union of all relevant documents in all these lists is assumed to be the set of relevant documents of the query.
- For each value of recall (the percentage of relevant documents retrieved) among all the recall points{0:0; 0:1; . . . :; 1:0}, the precision (the number of relevant document retrieved divided by the number of retrieved documents) is computed.
- Finally, the precision, averaged over all recall points, is computed. For each data set and for each mode of retrieval, we obtain a single precision value by averaging the precision values for all queries. The measure of efficiency is the average wall clock time for processing a user query
- Next, we examine the efficiency of our technique. Table 6shows that the average times for processing a query in seconds. Each of the times reported in the table consists of:
  a) the time to map the user query to a set of categories,
  b) the time for the search engine, Google Directory, to retrieve the documents,
  c) the time for our system to extract lists of documents from the search engine result pages, and
  d) the time to map user interested links from retrieved results
- Thus, the portion of our algorithm which consists of step a and d is efficient.

6. CONCLUSION

We described a strategy for personalization of Web search:

1. A user's search history can be collected without direct user involvement.
2. The categories that are likely to be of interest to the user are deduced based on his/her query
3. These categories are used as a context of the query to improve retrieval effectiveness of Web search.
4. For Each category relevant links are identified and mapped with it
5. User interested categories are tracked and the corresponding links will be mapped.
6. Thus the integrated technique for personalized web search is adopted to improve the precision.

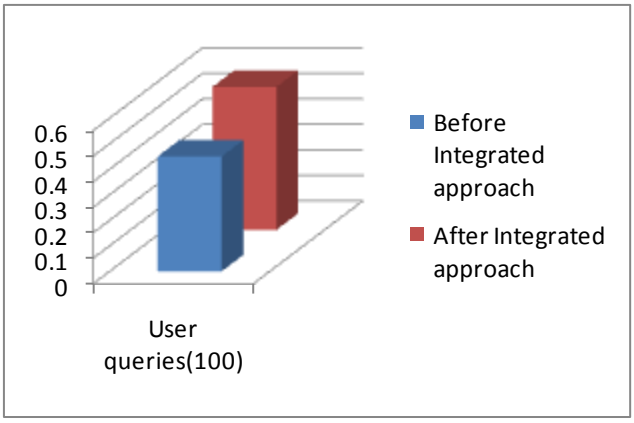

Fig.1: Comparison between precision before and after integrated approach

The ranking scheme is improved with the content and hyperlink in the web pages. The user can easily identify the relevant pages. The ranking scheme produces better results than other ranking.

7. REFERENCE

[1] Boanerges Aleman-Meza, Chris Halaschek, I. Budak Arpinar and Amit Sheth "Context-Aware Semantic Association Ranking" 2003.

[2] Fabrizio Lamberti, Andrea Sanna, and Claudio Demartini "A Relation-Based Page Rank Algorithm for Semantic Web Search Engines" IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 1, January 2009.

[3] W. P. Lee and M. H. Su, "Personalizing information services on wired and wireless networks," in EEE. IEEE Computer Society, 2004, pp.263–266.

[4] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, "User profiles for personalized information access," in The Adaptive Web

[5] Lecture Notes in Computer Science, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds., vol. 4321. Springer, 2007, pp. 54–89.

[6] G. S. B. Nelson, "Avoiding overload:personalizing web content through security, eintelligence and data mining," in Proc. of the SouthEast SAS

Users Group Conference, New Orleans, Louisiana, August 19-22 2001.

[7] P. Brusilovsky and C. Tasso, "Preface to special issue on user modeling for web information retrieval," User Model. User-Adapt. Interact.,vol. 14, no. 2-3, pp. 147–157, 2004.

[8] A. McCallum, "Information extraction: distilling structured data fro unstructured text," ACM Queue, vol. 3, no. 9, pp. 48–57, 2005.

[9] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the informational, navigational, and transactional intent of web queries," Inf. Process. Manage. vol. 44, no. 3, pp. 1251–1266, 2008.

[10] Hang Cui; Ji-Rong Wen; Jian-Yun Nie; Wei-YingMa, "Query expansion by mining user logs," IEEETransactions on Knowledge and Data Engineering, page(s): 829- 839, July-Aug. 2003.

[11] Taher Haveliwala, Aristides Gionis , Dan Klein, and Piotr Indyk. "Evaluating strategies for similarity search on the web," In Proceedings of the Eleventh International World Wide Web Conference, May 2002.

[12] Philippe Poinçot, Soizick Lesteven, and Fionn Murtagh. "Comparison of two document similarity search engines," Library and Information Services in Astronomy III, ASP Conference Series, Vol. 153, 1998

[13] Tomkins."Mining the link structure of the world wide web," IEEE Computer, 32(8): 60-67, 1999.

[14] H. Ahonen, O. Heinonen, M. Klemettinen, and A.Verkamo. "Finding co-occurring text phrases by combining sequence and frequent set discovery." In R.Feldman, editor, Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99Workshop on Text Mining: Fiundations, Techniques and Applications, page 1-9, 1999

[15] Fang Liu; Yu, C.; Weiywe Meng. "Personalized web search for improving retrieval effectiveness," IEEE Transaction on knowledge and data engineering,Volume: 16, Issue: 1, page 28-40, Jan. 2004.

[16] W. Fan, M.D. Gordon, P. Pathak, Personalization of search engine services for effective retrieval and knowledgemanagement, in the Proceedings of the 2000 International Conference on Information Systems(ICIS), 2000, Brisbane, Australia.

[17] Haixuan Yang, Irwin King, and Michael R. Lyu "DiffusionRank: A Possible Penicillin for Web Spamming" Proceedings SIGIR ACM 2007.

[18] Kemafor Anyanwu, Angela Maduko and Amit Sheth "SemRank: Ranking Complex Relationship Search Results on the Semantic Web" ACM, 2005.

[19] Li Ding, Rong Pan, Tim Finin, Anupam Joshi, Yun Peng, and Pranam Kolari "Finding and Ranking Knowledge on the Semantic Web" preprint from the Proceedings of the 4th International Semantic Web Conference, Galway IE, Springer-Verlag, November 2005.

[20] Oren Kurland, Lillian Lee "PageRank without Hyperlinks: Structural ReRanking using Links Induced by Language Models" ACM 2005.

[21] D. Mladenic. " Text -learning and related intelligent agents," IEEE Intelligent Systems , 14 (4): 44-54, 1999.

[22] E.M. Voorhees and D. Harman, eds., "Common Evaluation Measures," Proc. Text REtrieval Conf.(TREC-10), p.A-14 ,2001.

AUTHOR PROFILE

J. JAYANTHI (M'07–M'06) Author became a Member (M) of IEEE in 2007, a Member (M) of CSI in 2006.She was born on 19-02-1978, in salem, Tamilnadu. She has completed her UG(B.E. CSE) in GCT Coimbatore, PG(ME.CSE) in Sona College of Technology, Pursuing Ph.D in Web Personalization, Anna University of Technology. She has published papers in ten national conferences and three international conferences. She is the member of IEEE and CSI.

Dr.K.S.JAYAKUMAR Author was born 0n 05-06-1976 in salem,Tamilnadu.He has completed UG(B.E. Mech) in Angala Amman College of Engineering, Trichy/ Bharathidasan University, Trichy. PG(M.tech), Indian Institute of Technology, Delhi. PG (M.S) Singapore-MIT Alliance, Nanyang Technological University. Ph.D in Nanyang Technological University.He has published two books Natural Language Understanding by Robots and Cognitive Spoken English. He has published papers in international journals like International Journal of Language in India and International Journal of Humanoid Robotics