AMBIGUITY RESOLUTION IN INFORMATION RETRIEVAL

RekhaJain[1], Rupal Bhargava[2] and G.N Purohit[3]

[1, 2, 3]Deparment of Computer Science, Banasthali Vidyapith, Jaipur

[1]rekha_leo2003@yahoo.com
[2]bhargava.rupal@gmail.com
[3]gn_purohitjaipur@yahoo.co.in

ABSTRACT:

With the advancement of the web it is very difficult to keep up with the amplifying requirements of learning on web, to satisfy user's expectation. Users demand with the updated and accurate results. To solve the queries Search Engines use different techniques. Google the most famous search engine uses Page Ranking Algorithm. Ranking Algorithms arrange the results according to the user's needs. This paper deals with "Page Rank Algorithm". Our proposed algorithm is an extension of page rank algorithm which refines the results so that user gets what he/she expects. We have used a measure Average Precision to compare Page Rank algorithm and the proposed algorithm, and proved that our algorithm provides better results.

KEYWORDS:

Information retrieval, page ranking algorithms, weighted page rank

1. INTRODUCTION

"World Wide Web", a network spread throughout world, an ever expanding network that expands with a very high speed. This network has embedded in our lives slowly and gradually such that all the queries raised can be satisfied with a seconds work. For the requirement of query solving World Wide Web has provided with search engines. These search engines solve different queries with different results. Users always expect to get most relevant results at the top but it is not always possible for Search Engine to understand the actual perspective and meaning of the query. In this paper we have proposed an algorithm that tries to resolve such problems by resolving ambiguities in the query and provide better results.

2. LITERATURE REVIEW

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Today Search Engines are highly used for the purpose of Information Retrieval and with such a high demand it becomes essential to refine the results and present user with the most appropriate top ranked results. Web Mining is the field which deals with such concepts. R. Cooley, B. Mobasher and J. Srivastava[9] proposed the definition of Web Mining, its taxonomy. They also proposed architecture of Web Usage Mining. In July 2000, R. Kosala, H. Blockeel [10] pointed some confusion regarding the usage of the term Web Mining and its categories. Besides this they explored the connection between the Web mining categories and their related agent paradigm. David Hawkin [2] gave the descriptive view of web search engines and its

DOI : 10.5121/ijnlc.2013.2101

components. Eric J. Glover, Steve Lawrence, Michael D. Gordon, William P. Birmingham, C. Lee Giles [3] proposed a search engine that not only produces the relevant results but also allows the users to provide preferences in the form of an information need category. With the help of such information search engines may provide more valuable results.

To provide the results to the usersranking mechanism is used. Till now many Ranking Algorithms have been proposed. S. Brin and L. Page[7][11] proposed the ranking algorithm "Page Rank Algorithm" that is used in most popular search engine Google. This algorithm prioritizes web pages according to its inbound links. It used the concept of citation analysis such that a link coming from an important page was given high weight whereas page that was not so important was given a low weight. Also they gave a formula to calculate the page rank of pages in an iterative manner.

J.Kleinberg[5][6]gave us algorithm "Hyperlink Induced Topic Search" (HITS). He recognized two different forms of web pages called hubs and authorities. Authorities are pages having important contents whereas Hubs are pages that act as resource lists guiding users to authorities. Thus a good hub page for a subject points to many authorities pages on that content, a good authority page is pointed by any good hub pages on the same subject. Kleinberg said that a page may be a good hub and good authority at the same time and this circular relationship lead to the definition of an iterative algorithm HITS. This algorithm was used in search engine "CLEVER" but was not successful because of topic drift and efficiency problems. Also this algorithm worked on both Web Structure Mining as well as Web Content Mining.

Wenpu Xing and Ali Ghorbani [12] proposed "Weighted Page Rank" (WPR) algorithm which is an extension of Page Rank algorithm. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. This algorithm is based on Web Structure Mining.

3. WEB MINING

Web Mining is a process of automating information retrieval. Web Mining can be categorized into three as shown in figure 1:

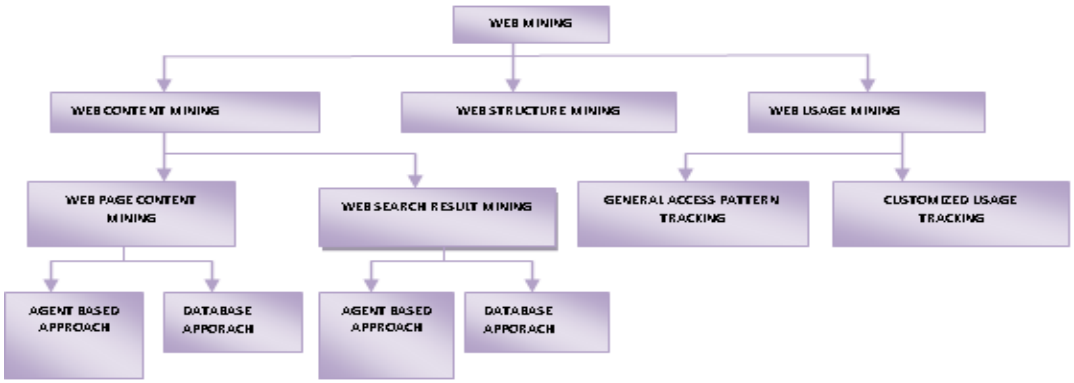

Figure1. Categories of Web Mining

Web Content Mining uses content of page to retrieve the results whereas Web Structure Mining uses the link structure to retrieve results. Web Usage Mining takes into account usage pattern of the user to extract information

4. RANKING MECHANISM

Ranking Algorithms are key part of Search Engines as they reorder the results of Search Engines. They work on different criteria depending upon the need and requirement. Using different web mining categories different ranking algorithms have been developed and used. To provide a better ranking algorithm and understand the key concept of ranking algorithms few ranking algorithms are discussed in next section.

4.1. Page Rank Algorithm

Page Rank Algorithm was developed by Brin and Page [7][11] at Stanford University. This algorithm uses link structure of web pages as it is a graph based algorithm. Also it considers back links for the calculation of ranks of web pages.

To calculate ranks Brin and Page [7][11] proposed a formula given in equation 1. Here T1, T2…Tn are pages pointing to it. Formula is as follows:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \text{ ..... (1)}$$

Where,

d damping factor (whose value is generally 0.85).It is used to stop other pages having too much influence)

C (Ti): number of links going out of Ti
PR (Ti): Page rank of Page Ti

Sum of page rank of all web pages corresponds to 1. It uses an iterative approach to calculate actual page rank of web pages initiating with page rank 1 for all web pages.

4.2. Weighted Page Rank Algorithm

Wenpu Xing and Ali Ghorbani. [12] proposed Weighted Page Rank algorithm. It is an extended version of Page Rank algorithm.This algorithm decides the rank of pages on basis of both in links and out links of the pages. Page Rank of the page is not equally divided among all the outgoing links; instead it divides the rank on the basis of importance of Web pages. Wenpu Xing and Ali Ghorbani[12] gave the modified formula as given in equation 2:

$$PR(u) = (1-d) + d \sum PR(v) W^{in}_{(v,u)} W^{out}_{(v,u)} \text{ ..... (2)}$$

d: damping factor (whose value is generally 0.85).It is used to stop other pages having too much influence)
PR (Ti): Page rank of Page Ti

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p \epsilon R(v)} O_p} \text{ ..... (3)}$$

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \epsilon R(v)} I_p} \text{ ..... (4)}$$

Iu and Ip: represent the number of in links of page u and page p, respectively.

Ou and Op: represent number of out links of page u and page p, respectively

R (v) denotes the reference page list of page v.

4.3. Hypertext Induced Topic Selection (HITS)

Kleinberg [5][6] categorized web pages into hubs and authorities such that authorities were the pages that contain important content or information specific to the query and hubs were the pages that contain links to authorities. Hence hub is a good hub when it points to many authoritive pages on the specific content whereas aauthorityis a good authority it is pointed by many good hubs. Kleinberg [5][6] said that pages may be a good authority and hub at the same time. This iterative relation formed the basis for HITS. In HITS there are two major steps:

- Sampling: A set of relevant pages to the query are collected from the complete set of web pages.
- Iterative Step: Finds hubs and authorities using formulas given below:

$$H_p = \sum_{q \in I(p)} A_q \ldots\ldots\ldots\ldots\ldots (5)$$
$$A_p = \sum_{q \in B(p)} H_q \ldots\ldots\ldots\ldots\ldots (6)$$

Where

Hp :is hub weight of p
Ap:is authority weight of p
Ip :is set of reference pages
Bp : is set of reference pages

5. PROPOSED WORK

This algorithm works as a layer on to the Search Engine that uses Page Rank Algorithm. It helps to resolve ambiguities of the queries and rearranges the results as per the user preferences. For e.g.suppose the user enters the query “bass” it may refer to bass as a guitar or as a fish. Our algorithm disambiguatesthese words and produces results according to the user preference. The flowchart of the proposed algorithm is given in figure 2.

Figure 2.Flowchart for proposed algorithm

6. RESULTS

We have considered two users User A and User B. User A is a fisherman whereas User B is a musician. When user A wants to search bass there is high probability that he will search for bass (fish). Also when User B wants to search same word bass there is a high probability that he meant bass as bass (guitar). When the bass query is given to Google it returns pages for both bass (guitar) and bass (fish) but the bass (guitar) pages are on the top priority (as shown in figure 3) and hence User A won't get satisfactory results.The results produced by our algorithmare not like Google's results. Our algorithm disambiguates the query for the user whether it be User A or User B and provides the results according to the user priority.We have applied Average Precision to compare the efficiency of Page Rank algorithm and our newly proposed algorithm. Average Precision uses Precision and Recall both to compute its value. Precision is the fraction of retrieved documents that are relevant to the search [4]. Recall in information retrieval is the fraction of the documents that are relevant to the query and are successfully retrieved [4]. Average precision computes the average value of $p(r)$over the interval from $r = 0$to$r - 1$[4].

$$\text{AveP} - \int_0^1 p(r)dr.$$

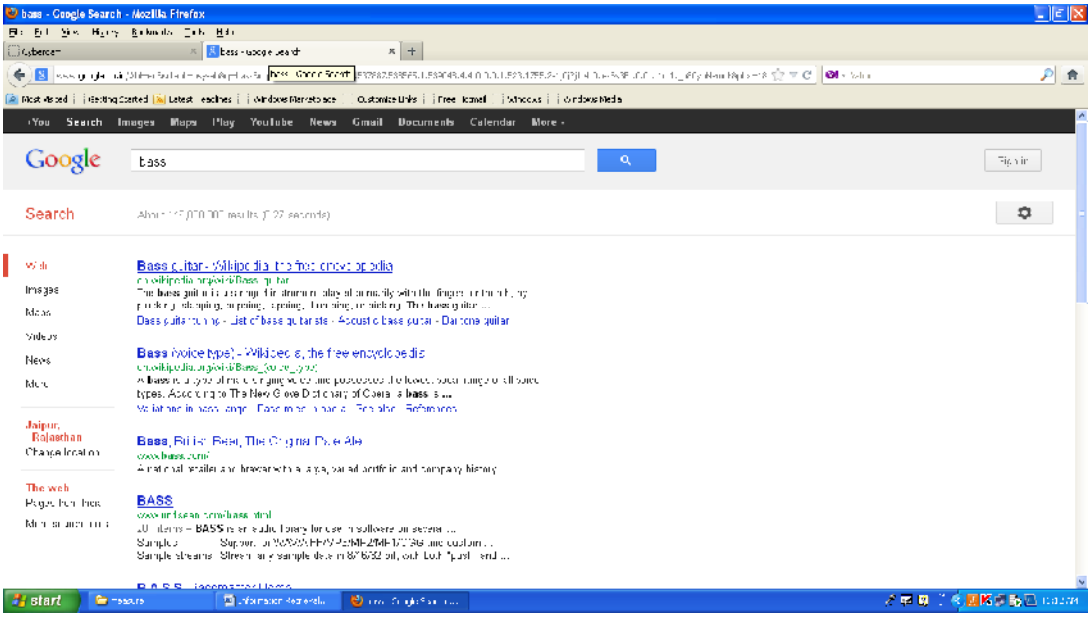

Figure 3.Google results

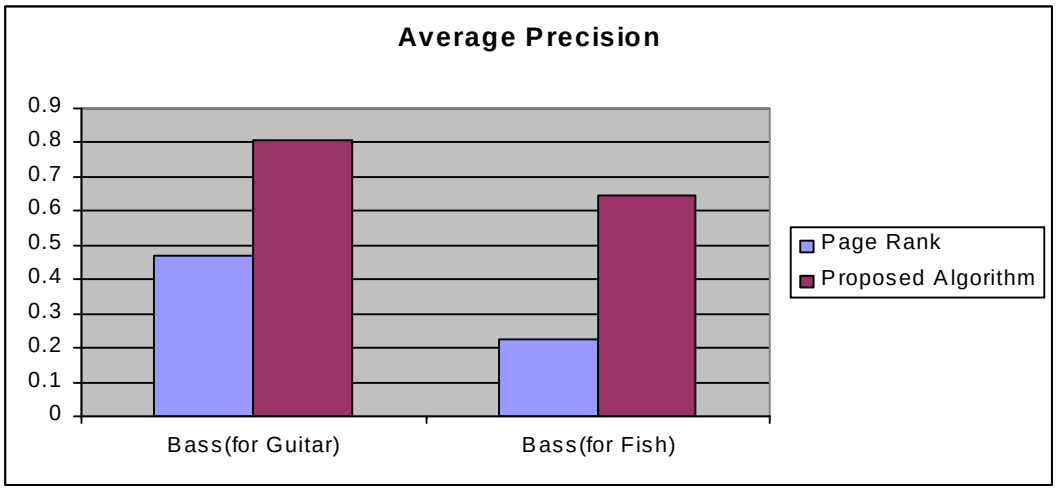

Figure 4. Results for Average Precision

7. CONCLUSION

Ranking algorithms work on different perspectives to sort the results and give user satisfying results. Page Rank algorithm is widely used and accepted algorithm. In this paper we have proposed an extension of Page Rank algorithm which helps to resolve ambiguities of user queries and provide more relevant results to users.We have also compared Page Rank algorithm and new algorithm on the basis of Average Precision and proved that new algorithm provides more relevant results to users.

REFERENCES

[1] Ashutosh Kumar Singh, Ravi Kumar P, A Comparative study of Page Ranking Algorithms for Information Retrieval, International Journal of Electrical and Computer Engineering 4:7 2009

[2] David Hawkin,Web Search Engines, CSIRO ICT Center

[3] Eric J. Glover, Steve Lawrence, Michael D. Gordon, William P. Birmingham, C. Lee Giles, Web Search – Your Way

[4] "Information Retrival" available at http://en.wikipedia.org/wiki/Information_retrieval

[5] J. Kleinberg, "Authoritative Sources in a Hyper-Linked Environment", Journal of the ACM 46(5), pp. 604-632, 1999.

[6] J. Kleinberg, "Hubs, Authorities and Communities", ACM Computing Surveys, 31(4), 1999.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank Citation Ranking: Bringing order to the Web". Technical Report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[8] "Precision and Recall" available at http://en.wikipedia.org/wiki/Precision_and_recall

[9] R. Cooley, B. Mobasher and J. Srivastava, "Web Minig: Information and Pattern Discovery on the World Wide Web". Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, pp. (ICTAI'97), 1997.

[10] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.

[11] S. Brin, and L. Page, The Anatomy of a Large Scale Hypertextual Web Search Engine, Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117,1998.

[12] W. Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proc. Of the Second Annual Conference on Communication Networks and Services Research (CNSR '0), IEEE, 2004.

AUTHORS

Rekha Jain completed her Master Degree in Computer Science from Kurukshetra University in 2004. Now she is working as Assistant Professor in Department of "Apaji Institute of Mathematics & Applied Computer Technology" at Banasthali University, Rajasthan and pursuing Ph.D. under the supervision of Prof. G. N. Purohit. Her current research interest includes Web Mining, Semantic Web and Data Mining. She has various National and International publications and conferences.

RupalBhargava is pursuing her M.Tech in Computer Science from BanasthaliVidyapith, Rajasthan. She is undergoing the training of her M.Tech in supervision of Mrs. Rekha Jain. Her current research interest includes Web Mining, Semantic Web and Data Mining. She has published various papers in the conferences and journals.

Prof. G. N. Purohit is a Professor in Department of Mathematics & Statistics at Banasthali University (Rajasthan). Before joining Banasthali University, he was Professor and Head of the Department of Mathematics, University of Rajasthan, Jaipur. He had been Chief-editor of a research journal and regular reviewer of many journals. His present interest is in O.R., Discrete Mathematics and Communication networks. He has published round 40 research papers in various journals.