Word Disambiguation in Web Search

Rekha Jain
Computer Science, Banasthali University,
Rajasthan, India
Email: rekha_leo2003@rediffmail.com

G.N. Purohit
Computer Science, Banasthali University,
Rajasthan, India
Email: gn_purohitjaipur@yahoo.co.in

***Abstract*— Internet is huge like a sea as the amount of information is growing rapidly on WEB. Whenever user searches something on Internet the Search Engine provides an incredible amount of information that increases the complexity of dealing with information. Various algorithms have been developed that help the user to retrieve the web contents. Sometimes these algorithms do not give fruitful results especially in case of Homographs. In this paper authors discuss a disambiguation algorithm that is used for information retrieval in web search. The experimental study is done to find out whether the approach provides better results.**

Keywords-Homograph, Word Disambiguation, Web Search.

I. INTRODUCTION

The web is an informative environment that is highly distributed as well as heterogeneous. Web has no authorship standards, no editorial board and no hierarchy, so the documents on the web produce various challenges for Search Engine in terms of storage space, crawling speed and retrieval of relevant documents [1]. Because of heterogeneity engineering of a Search Engine is becoming a challenging task. The problem in designing a Search Engine is that most of the users are not expert in Information Retrieval. They do not have enough experience to format their query. Their only requirement is that relevant pages should occur at the top of search result. Various algorithms like PageRank, Weighted PageRank, HITS (Hypertext Induced Topic Search) have been developed to search the content and show the pages in ranking order. They catalog the links and generate the information like similarity and relations in between the links by using hyperlink topology. These algorithms are based on Web Structure Mining that is a specific branch of Web Mining. Web Mining is a technique that is used to crawl through various web resources to collect required information [2]. It is a process of scraping and data extraction using software and various tools.

But sometimes above mentioned algorithms do not produce fruitful results especially in case of Homographs. Homographs are the words that have same spellings but different meanings and origins. As homographs spelt identically, so some technique is required while searching those keywords because there is no meaning to extract additional contextual data.

In this paper an algorithm is introduced for disambiguation of words that have same spellings but different meanings in different contexts in web search. The goal is to disambiguate those words which look alike but semantically different. In this scenario the preferences of users play an important role in calculating PageRank values. The idea is implemented to explore the improvements in search results. In the following sections the idea and details of implementation are discussed.

II. RELATED WORK

Many algorithms are in developing process for the improvement of efficient retrieval from WEB. The term Web Mining was initially discovered by Etzioni. He assumed that the information on the web is sufficiently structured [3]. Web Mining is an interactive process to discover the knowledge for understanding consumer and business activities on the web [4]. According to R. Kosala and H. Blockeel there are three Web Mining categories – 1)Web Content Mining, 2)Web Structure Mining and 3)Web Usage Mining [5]. S. Brin and L. Page developed the PageRank algorithm in their Ph. D. work at Stanford University [6]. This algorithm is based on citation analysis. Weighted PageRank algorithm that is an extension of PageRank algorithm is proposed by Wenpu Xing and Ali Ghorbani [7]. Here larger rank values are assigned to more important pages rather than

dividing it evenly among all its outgoing linked pages. Kleinberg introduced HITS algorithm that is based on link analysis. He described two different classes of pages- Hub and Authority. These are distinct but interrelated types of pages [8]. According to L. John Old Natural Language Processing must deal with disambiguation of Homographs [9]. The disambiguation was based on Formal Concept Analysis to extract and visualize the disambiguate words. David Yarowsky described a technique for Word Sense Disambiguation that is based on statistical model of word class [10]. Each word is given a weight that helped in selection of categories.

III. STATEMENT OF PROBLEM

People generally use Google search Engine for searching their information on web. At the time of searching when ambiguous word is used in a sentence people never bother about multiple meanings of words to be searched. The result contains the pages from all the possible contexts. This research provides better solutions for searching the word on Internet that have multiple meanings. These words are known as Homographs. IR (Information Retrieval) is a field that corresponds with searching the information for documents, within documents and for metadata about documents as well as in relational database and World Wide Web.

In the present scenario when the user performs the search on search engine, the results of search are poorly organized. Suppose a user is an ornithologist and he/she is interested in searching the details about word Crane. (This crane may be a machine or a bird or anything else). The search result contains the pages that have details about **bird crane** as well as pages that have details about **machine crane**. It is the responsibility of user to discard the irrelevant documents that are provided by Search Engine as search result. Each and every user experiences this problem but most of us taken it for granted that it is end user's responsibility to discard irrelevant pages of different contexts. This research is aimed to customize web search results so that pages related to same context may appear altogether.

IV. OUR APPROACH

The context disambiguator is designed to solve above mentioned problems, which have to work on a corpus or a collection of web pages. This corpus would provide more insight into the nature of the solution. The approach was to implement it as a meta-service that passed the query to the search engine after resolving ambiguation. This is known as preprocessing. This solution would be fast as user would receive only the relevant content and the user has to analyze only handful of pages.

V. DESIGN AND IMPLEMENTATION

A. User Interfaces: The disambiguation Search Engine was designed as user friendly and it was easy to use application. The main aim behind the development of algorithm was to minimize complications and provide simple and efficient user interface. User Interface of search engine is modular in structure and flexible for modifications.

B. Database: For development of this algorithm the SQL 2000 is taken as back end the information stored in it is from various domains.

VI. METHODOLOGY

Two users are considered in this experiment. Each user was asked to specify his/her domain of interest. It had been reported that generally the users were interested to explore only 8-10 pages of search result, so the query result should be relevant according to users' interest. First user was an Ornithologist whose domain was to study the birds, and second user was an Engineer. This user was interested in searching the information about various parts and types of machines.

VII. EXPERIMENTAL EVALUATION

To show the experimental results two registered users are considered- an Ornithologist and an Engineer. The disambiguation algorithm remembers the primary domain of interest and retrieves more meaningful contents to the users.

An ornithologist searched the word **crane** via Google Search Engine and entered the word **crane** on search engine interface as shown in Figure 1.

Figure 1

The results received were not up to the mark because he/she was expected the details about the bird crane not about machine or anything else.

The algorithm resolved the ambiguation between Noun Homographs. At the time of searching users never bothered about the multiple meanings of the word; their only requirement is that their relevant content must appear at top of result.

But when the same user performed the same search through our developed module the result varies. Those results were more relevant as compared to earlier results as shown in Figure1, because the pages appear at the top of result provided the details regarding the **Crane bird.**

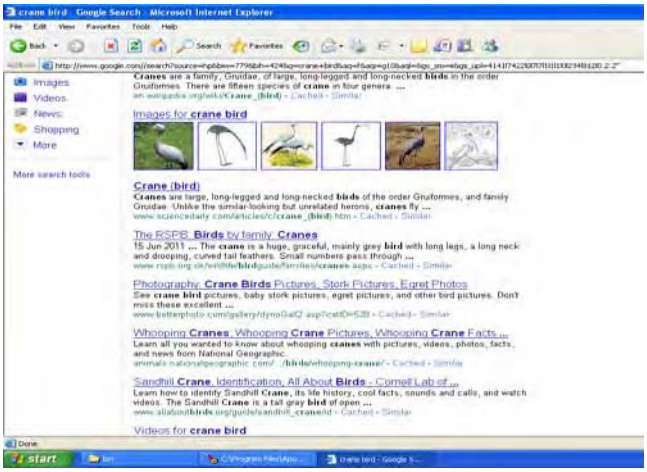

Figure 2

As the user entered the keyword to be searched and clicked on search button the results shown in Figure 2 were displayed. The results were more relevant as compare to earlier results shown in Figure 1.

If the user is an engineer then it is obvious that he/she is interested in searching the details for crane as machine. Figure3 shows the results in following manner such that if an engineer searched the details for word Crane. Here the top of result provided the details for the **Crane machine.**

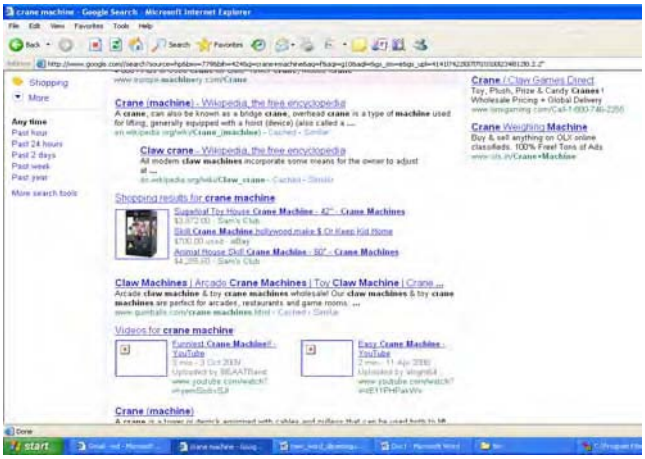

Figure 3

VIII. ANALYSIS OF RESULT

Figure1 shows the result when the user is directly entered crane keyword on to Google interface. Here Google searches all the possible pages having word crane in them and then arranges them in the descending order of their page ranks. It includes pages from all the possible domains. In new developed algorithm user never enters search keywords on to Google interface instead he/she performs the search via our algorithm's search interface. The algorithm provides the result in different manner as it can be seen in Figure 2 and Figure 3 that both the users (Ornithologist and Engineer) enter the same word to search and disambiguation algorithm performs some preprocessing and then passes the resultant query to Google and as a result the ornithologist and engineer receive respective web pages.

CONCLUSIONS

In this paper a solution has been proposed that can be easily improved the quality of search results. The proposed approach is based on an algorithm that can resolved the meaning of Homographs. Various algorithms are used by various search engines, but majority of them are based on link analysis. If we run our disambiguation algorithm before sending the query to search engine we can get better results. By simulating this algorithm we have achieved the better quality of search engine results.

REFERENCES

[1] Biswas Pradipta, "Recent Researches on Web Page Ranking", Indian Institute ofTechnology, Kharagpur, India

[2] Web Data Mining" available at www.web-datamining.net/

[3] O. Etizioni, 1996, "The World Wide Web: Quagmire or Gold Mine", Communications of the ACM, 39(11): 65-68.

[4] "Web page" available at http://en.wikipedia.org /wiki/Web_page

[5] R. Kosala, H. Bloskeel, 2000, "Web Mining Research: A Survey", in SIGKDD Explorations, Newsletter of ACM Special interest group on Knowledge Discovery and Data Mining Vol2 No. 1 Pages 1-15.

[6] S. Brin, and L. Page, April 1998, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", in Seventh International World-Wide Web Conference (WWW 1998), Brisbane, Australia.

[7] Wenpu Xing and Ali Ghorbani, 2004, "Weighted Page Rank Algorithm", Preceedings of the Second Annual Conference on Communication Networks and Services Reasearch (CNSR'04).

[8] Joel C. Miller, Gregory Rae and Fred Schafer, Sep 2001 "Modification of Kleinberg's HITS Algorithm Using Matrix Exponentiation and Web Log Records", SIGIR'01, ACM 1-58113-331-6/01/0009

[9] L. John Old, "Homograph Disambiguation Using Formal Concept Analysis", 4th International Conference, ICFCA 2006, Dresden, Germany, Feburary 13-17, 2006, Proceedings. Lecture Notes in Computer Science Vol. 3874. Berlin: Springer, 2006. ISBN 9783540322030.

[10] Yarowsky David, "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora" Proceedings. Of COLING-92, NANTES, AUG. 23-28, 1992

AUTHORS PROFILE

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.

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 around 40 research papers in various journals.