An Efficient Approach for Enhancing Web Search Results Delivery

Prof. N.V. Pardakhe, Prof. S. S. Dandge, Prof. N. M. Yawale

Abstract— In today's e-world search engines play an essential role in retrieving and organizing relevant data for various purposes. However, in the real ground relevance of results produced by search engines are still arguable because it returns much amount of irrelevant and redundant results. Providing relevant information to user is the primary goal of the website owner. Web mining is ample and powerful research area in which retrieval of relevant information from the web resources in a faster and better manner. Web content mining improves the searching process and provides relevant information by eliminating the redundant and irrelevant contents. However for a broad-topic and ambiguous query, different users may have different search goals when they submit it to a search engine. Web usage mining plays an important role in inferring user search goals as they can be very useful in improving search engine relevance and user experience. The paper focuses on two important issues: improving search-engine performance through dynamic caching of search results, and helping users to find interesting web pages.

1 INTRODUCTION

It is not exaggerated to say the Web World Web is the most excited impacts to the human society in the last 10 years. It changes the ways of doing business, providing and receiving education, managing the organization etc. The most direct effect is the completed change of information collection, conveying, and exchange. Today, Web has turned to be the largest information source available in this planet. The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of view – users, Web service providers, business analysts. With the exponential growth of WWW, it has become difficult to access desired information that matches with user needs and interest. The users want to have the effective search tools to find relevant information easily and precisely. Therefore, Web mining becomes a popular research field.

Web mining is the process of discovering knowledge, such as patterns and relations, from Web data. Web mining generally has been divided into three main areas: content mining, structure mining and usage mining. Each one of these areas are associated mostly, but not exclusively, to these three predominant types of data found in the Web:

Content: The real data that the document was designed to give to its users. In general this data consists mainly of text and multimedia.

Structure: This data describes the organization of the content within the Web. This includes the organization inside a Web page, internal and external links and the website hierarchy.

Usage: This data describes the use of a website or search engine, reflected in the Web server's access logs, as well as in logs for specific applications. There is not a clear-cut distinction among these categories, and all three mining tasks can be combined.

In order to retrieve user requested information, search engine plays a major role for crawling web content on different node and organizing them into result pages so that user can easily select the required information by navigating through the result pages link. This strategy worked well in earlier because, number of resources available for user request is limited. Also, it is feasible to identify the relevant information directly by the user from the search engine results. When the Internet era increases, sharing of resource also increases and this leads to develop an automated technique to rank each web content resource. Different search engine uses different techniques to rank search results for the user query.

2 WEB SEARCH ENGINE

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as Search Engine Result Pages (SERPs). The information may be a specialist in web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain Real Time Computing information by running an algorithm on a web crawler.

2.1 Components of Web Search Engine:

- *Nilima V.Pardakhe is currently working as an Asst. Professor in Comp.Science Department at PRMIT&R, Badnera.*
- *Sangram S. Dandge is currently working as an Asst. Professor in Comp.Science Department at PRMIT&R, Badnera.*
- *Nupoor M. Yewale is currently working as an Asst. Professor in Comp.Science Department at PRMIT&R, Badnera.*

User Interface: The user interface, in the industrial design field of human machine interaction is the space where interaction between humans and machines occurs. The goal of this interaction is effective operation and control of the machine on the user's end, and feedback from the machine, which aids the operator in making operational decisions.

Parser: It is the component providing term (keyword) extraction for both sides. The parsers determine the keywords of the user query and all the terms of the Web documents which have been scanning by the crawler.

Web Crawler: A web crawler is a relatively simple automated program, or script that methodically scans or "crawls" through Internet pages to create an index of the data it is looking for. Alternative names for a web crawler include web spider, web robot, crawler, and automatic indexer. When a web crawler visits a web page, it reads the visible text, the hyperlinks, and the content of the various tags used in the site, such as keyword rich Meta tags. Using the information gathered from the crawler, a search engine will then determine what the site is about and index the information. Lastly, the website is included in the search engine's database and its page ranking process.

Database: It is the component that all the text and metadata specifying the web documents scanned by the crawler.

Ranking Engine: The component is mainly the ranking algorithm operating on the current data, which is indexed by the crawler, to be able to provide some order of relevance, for the web documents, with respect to the user query.

2.2 Search and Matching Function

How systems carry out their search and matching functions differs according to which theoretical model of information retrieval underlies the system's design philosophy. Searching the inverted file for documents meeting the query requirements, referred to simply as "matching," is typically a standard binary search, no matter whether the search ends after the first two, five, or all seven steps of query processing. While the computational processing required for simple, unweighted, non-Boolean query matching is far simpler than when the model is an NLP-based query within a weighted, Boolean model, it also follows that the simpler the document representation, the query representation, and the matching algorithm, the less relevant the results, except for very simple queries, such as one-word, non-ambiguous queries seeking the most generally known information.

Having determined which subset of documents or pages matches the query requirements to some degree, a similarity score is computed between the query and each document/page based on the scoring algorithm used by the system. Scoring algorithms rankings are based on the presence/absence of query term(s), term frequency, tf/idf, Boolean logic fulfillment, or query term weights. Some search engines use scoring algorithms not based on document contents, but rather, on relations among documents or past retrieval history of documents/pages. After computing the similarity of each document in the subset of documents, the system presents an ordered list to the user. The sophistication of the ordering of the documents again depends on the model the system uses, as well as the richness of the document and query weighting mechanisms. For example, search engines that only require the presence of any alpha-numeric string from the query occurring anywhere, in any order, in a document would produce a very different ranking than one by a search engine that performed linguistically correct phrasing for both document and query representation and that utilized the proven tf/idf weighting scheme. However the search engine determines rank, the ranked results list goes to the users, who can then simply click and follow the system's internal pointers to the selected document/page.

2.3 Limitations of Web Search Engine

With the dramatically quick and explosive growth of information available over the Internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. Due to the properties of the huge, diverse, dynamic and unstructured nature of Web data, Web data research has encountered a lot of challenges, such as scalability, multimedia and temporal issues etc. As a result, Web users are always drowning in an “ocean” of information and facing the problem of information overload when interacting with the web.

Typically, the following problems are often mentioned in

A. Finding relevant information: To find specific information on the web, users often either browse Web documents directly or use a search engine as a search assistant. When a user utilizes a search engine to locate information, he or she often enters one or several keywords as a query, then the search engine returns a list of ranked pages based on the relevance to the query. However, there are usually two major concerns associated with the query-based Web search. The first problem is low precision, which is caused by a lot of irrelevant pages returned by the search engine. The second problem is low recall, which is due to the lack of capability of indexing all Web pages available on the Internet. This causes the difficulty in locating the unindexed information that is actually relevant.

B. Finding needed information: Most search engines perform in a query-triggered way that is mainly on a basis of one keyword or several keywords entered. Sometimes the results returned by the search engine don’t exactly match what a user really needs due to the fact of the existence of the homology. For example, when one user with an information technology background wishes to search information with respect to "Python" programming language, he/she might be presented with information on the creatural python, one kind of snake rather than the programming language, given entering only one "python" word as query. In other words, the semantics of Web data is rarely taken into account in the context of Web search.

C. Learning useful knowledge: With traditional Web search service, query results relevant to query input are returned to Web users in a ranked list of pages. In some cases, we are interested in not only browsing the returned collection of Web pages, but also extracting potentially useful knowledge out of them.

3 RELATED WORKS

Due to the heterogeneity of network resources and the lack of structure of web data, automated discovery of targeted knowledge retrieval mechanism is still facing many research challenges. Moreover, the semi-structured and unstructured nature of web data creates the need for web content mining. In Paper [9], the author differentiates web content mining from two different points of view. Information Retrieval view and Database view. Characteristics of web and various issues on web content mining presented in [1]. In paper [8] research areas of web mining and different categories of web mining are discussed briefly. They also summarized the research works done for unstructured data and semi-structured data from information retrieval (IR) view. In IR view, the unstructured text is represented by bag of words and semi-structured words are represented by HTML structure and hyperlink structure [8]. In Database (DB) view, the mining always tries to infer the structure of the web site to transform a web site into a database. A new method for relevance ranking of web pages with respect to given query was determined in paper[5]. Various problem of identifying content such as a sequence labeling problem, a common problem structure in machine learning and natural language processing is identified in [3]. A survey of web content mining plays as an efficient tool in extracting structured and semi structured data and mining them into useful knowledge is presented in [6]. A framework is proposed to provide facilities to the user during search [7]. In this framework user does not need to visit the homepages of companies to get the information about any product, instead the user write the name of the product in the Query Interface (QI) and the framework searches all the available web pages related to the text, and the user gets the information with little efforts. In [10]-[12] Statistical approach using proportions and chi-square for retrieving relevant information from both structured and unstructured documents are presented. The authors applied correlation method to detect and remove redundant web documents Nowadays, most of the people rely on web search engines to find and retrieve information. When a user uses a search engine such as Yahoo or Google or bing to seek specific information, an enormous quantity of results are returned containing both the relevant document as well as outlier document which is mostly irrelevant to the user. Therefore discovering essential information from the web data sources becomes very important for web mining research community. Chakrabarti et al (1999) describes a new hypertext resource discovery system called focused crawler which analyze its crawl boundary to find the links that are likely to be most relevant for the crawler and avoids irrelevant regions of the web. Mei Kobayashi and Koichi Takeda (2000) discussed the development of new techniques targeted to resolve some of the problems such as slow retrieval speed, noise and broken links associated with web based information retrieval and speculates on future needs. Mayfield et al (1998) explores the indexing using both Ngrams and words by using HAIRCUT (Hopkinks Automated Information Retrieving for Combing Unstructured Text) System. Junghoo Cho et al (2000) present the efficient method for identifying replicated document collections to improve web crawlers, archivers and ranking functions used in search engines. Sungrim Kim and Joonhee Kwon (2009) propose an information retrieval method using the context information on the web 2.0 environment by adopting page rank and context tags algorithms. Brin et al (1998) gives an in-depth description of large scale web search engines and described the page rank algorithm. The algorithm states that the relevance of a page increases with the number of hyperlinks to it from other relevant pages. Bin et al (2003) explained web mining process and the Taxonomy of web mining. Georgioes (2007) provide an overview of web mining and the latest developments on web mining application in beneficial to society.

4 PROPOSED SYSTEM

The proposed methodology aims to provide the results to the users which are more relevant to the user query. It tries to overcome the problem of page ranking, in which an approach of relevant search which ranks the web pages based on the frequency or count of keywords (searched by user) is proposed. The web page containing maximum frequency or counts of keyword searched by the user is more relevant and displayed first in the list of web page links on the user screen. Every result is individually analyzed based on frequency of keywords and thus based on the user query, search results are obtained.

Proposed methodology works as follows:

It involves user request to search for the particular query to obtain the search results according to the user query. In Proposed methodology user has to first enter the query. Then preprocessing is performed on that entered query which involves three steps such as text filtering, stemming, stop words removal. After preprocessing of the query keywords are obtained. Then web snippets related to that keyword are fetched from the dataset and frequency of that particular keyword is calculated and finally on the basis of that frequency of keyword, search results are ranked that is the search results are displayed in descending order of frequency of keyword to the user.

A. Dataset used:

Dataset is created by collecting web snippets for some particular keywords. So, here for implementation of ranking rather than fetching the snippets from any search engine AMBIENT [13] data set is used, in which numbers of snippets already has stored. This means that this work considers that, the work had already done for extraction of top 200-500 snippets from top search engine and can be stored in text file. The implementation is done with the AMBIENT dataset.

B. AMBIENT

It is a dataset designed for evaluating the subtopic information retrieval. It consists of 44 topics which are selected from Wikipedia disambiguation page. Each topic has a set of subtopics. Each subtopic has a set of documents that comprises of URL, title and snippet, retrieved from a Web search engine as of January 2012. They are annotated with subtopic relevance judgments. The AMBIENT dataset has 44 topics with an average of 17 subtopics under each topic. The topics and its subtopics which do not have any appropriate terms within the search result are considered to be noise and are removed from the dataset.

4.1 Algorithm for Mining Web Content

Algorithm: Relevancy and keyword frequency based approach

Input: User query

Output: Reordered search results

Step 1: Enter the user query.

Step 2: Perform preprocessing of user query.

Step 3: Obtain keywords from processed query.

Step 4: Extract the web snippets from the dataset related to the specified keyword.

Step 5: Find frequency of the keyword.

Step 6: Display the search results in descending order of keyword frequency.

5 PERFORMANCE EVALUATION

Performance evaluation of the proposed approach is done based on classification context scenario. Precision, Recall, Accuracy and F1 - Score plays a major role in classification based performance. Precision measure is calculated based on the formula

$$\text{Precision} = \frac{tp}{tp+fp}$$

Recall is calculated based on the formula

$$\text{Recall} = \frac{tp}{tp+fn}$$

Accuracy is calculated based on the formula

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

Where

tp – True Positive (Correct result)

tn – True Negative (Correct absence of result)

fp – False Positive (Unexpected Result)

fn – False Negative (Missing result)

F-Measure is calculated based on the formula

$$F= 2. \frac{\text{Precision .Recall}}{\text{Precision+ Recall}}$$

In this proposed work sample Dataset is consider for evaluation purpose and top 10 documents that are more relevant to the user based on user decision is classified manually with different users . Now the same relevant dataset is evaluated against retrieved dataset. Comparison results of the proposed approach are given in the TABLE I.

TABLE I.
RANKING COMPARISION

TABLE III represents the matching of manual ranking against proposed approach

Search Engine ranking (Dataset)	url	Manual Ranking	Proposed approach ranking
1	http://www.camelproductions.com/	7	7
2	http://en.wikipedia.org/wiki/Camel	9	6
3	http://en.wikipedia.org/wiki/Camel_(band)	2	2
4	http://inertron.com/camel/	3	3
5	http://www.math.ca/	4	4
6	http://www.progarchives.com/artist.asp?id=50	6	8
7	http://www.britannica.com/eb/article 9018795/camel	8	9
8	http://lexicorient.com/e.o/camel.htm	10	10
9	http://www.youtube.com/watch?v=ZTVnCyDoQlQ	1	1
10	http://www.sandiegozoo.org/animalbytes/t-camel.html	5	5

TABLE I contains result for evaluating the proposed approach against various performance measures like Precision, Recall, Accuracy and F-Measure. There is mismatching of manual ranking against proposed approach. From the table, it is understood that precision of the proposed system is 0.7 out of 1 where as search-engine precision is 0.1 out of 1The results of the performance measure are plotted in following Figure.

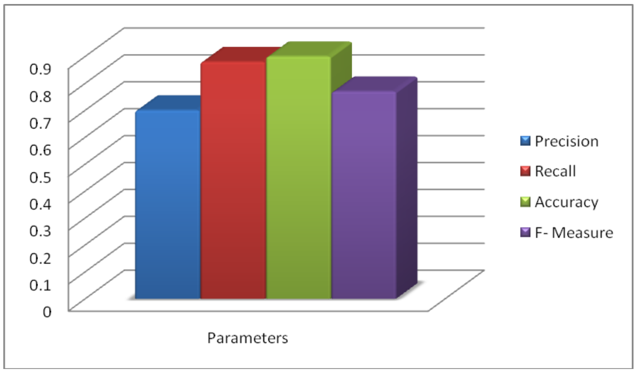

Fig. 2 Performance of Proposed system

6 CONCLUSION

The Proposed approach gives far better results compared with search-engine ranking. However, more fine tuning process to be needed to bring the best result. Proposed methodology focus only on text based mining to rank the relevancy of the web pages where nowadays relevant information may be available in any format like images, audio and video files. Forth coming research work will focus on all types of data sets.

REFERENCES

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

[2] 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.

[3] I. Mele, " Web Usage Mining for Enhancing Search –Result Delivery and Helping Users to Find Interesting Web Content," ACM SIGIR Conf. Research and Development in Information Retrival (SIGIR '13), pp. 765-769, 2013.

[4] P. Sudhakar, G. Poonkuzhali, R. Kishor Kumar, "Content Based Ranking for Search Engines," Proc. International Multi Conference of Engineers and Computer Scientists (IMECS 12), 2012.

[5] Z. Lu, H. Zha, X. Yang, W. Lin, Z. Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions," Proc. IEEE Transactions on Knowledge and Data Engineering, pp. 502-513, 2013

[6] http://searchenginewatch.com/.

[7] Guandong Xu, "Web Mining Techniques for Recommendation and Personalization", A Dissertation submitted to The School of Computer Science & Mathematics Faculty of Health, Engineering & Science Victoria University, Australia For the degree of Doctor of Philosophy March 2008

[8] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining - Concepts, Applications & Research Directions", University of Minnesota, USA, Chapter 3. Pg 52-71.

[9] Oren Etzioni. The World Wide Web: Quagmire or gold mine. Communications of the
ACM, 39(11):65-68, 1996

[10] Sule Gundus, "Rcommendation Model for Web Users: User Interest Model and Click Stream Tree", Ph.D. Thesis, Istanbul Technical University, Institute of Science and Technology, October 2003.

[11] A. Jain, R. Sharma, Gireesh Dixit, V. Tomar, "Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages", 2013 International Conference on Communication Systems and Network Technologies, IEEE 2013.

[12] K.Wang and H. Liu. Discovering Typical Structures of Documents: A Road Map Approach. In 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 146–154, 1998.

[13] http://www.web-datamining.net/structure/

[14] J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan. Web Usage Mining: Discovery
and Applications of Usage Patterns from Web Data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[15] Y. Wang, "Web Mining and Knowledge Discovery of Usage Patterns", CS 748T Project, February 2000.

[16] J. Gou, "Web Content Mining & Structured Data Extraction & Integration", University of Illinois at Urbana- Champign.

[17] http://en.wikipedia.org/wiki/Web_search_engine

[18] A. Gupta, A. Dixit, A. K. Sharma, "Relevant Document Crawling with Usage Pattern and Domain Profile Based Page Ranking", IEEE, 2013.

[19]http://www.infotoday.com/searcher/may01/liddy.htm

[20] S. Brin and L. Page. "The anatomy of a large-scale hypertextual Web search engine". *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[21] Ron Weiss, Bienvenido Velez, Mark A. Sheldon, Chanathip Manprempre, Peter Szi- lagyi, Andrzej Duda, and David K. Gi ord. HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In Proceedings of the 7th ACM Conference on Hypertext, pages 180{193, New York, 16{20 March 1996. ACM Press.

[22] Ellen Spertus. Parasite: Mining structural information on the web. In Proceedings of the Sixth International WWW Conference, Santa Claram USA, April,1997,

[23] Sougata Mukherjea, James D. Foley, and Scott Hudson. "Visualizing complex hyper- media networks through multiple hierarchical views". In Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, volume 1 of Papers: Creating Visualizations, pages 331{337, 1995.

[24] J. M. Kleinberg. "Authoritative sources in a hyperlinked environment". *Journal of the ACM*, 46(5):604–632, September 1999.

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

[26] Taher H. Haveliwala, "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search", IEEE Transaction on Knowledge and Data Engineering, Vol. 15, No. 4, July/August 2004

[27] Chekuri, C., Goldwasser, M., Raghavan, P. and Upfal, E. "Web search using automated classification". In Sixth International World Wide Web Conference, Santa Clara, California, Apr. 1997.

[28] Hao Chen and Susan Dumais, "Bringing Order To the Web : Automatically Categorizing Search results", Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.