International Journal of Engineering & Technology, 7 (2.7) (2018) 1025-1029

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Optimizing webpage relevancy using page ranking and content based ranking

J. Satish Babu [1] *, T. Ravi Kumar [1], Dr. Shahana Bano [1]

[1] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, V addeswaram, AP, 522502, India
**Corresponding author E-mail: jampanisatishbabu@kluniversity.in*

Abstract

Systems for web information mining can be isolated into a few classifications as indicated by a sort of mined data and objectives that specif-ic classifications set: Web structure mining, Web utilization mining, and Web Content Mining. This paper proposes another Web Content Mining system for page significance positioning taking into account the page content investigation. The strategy, we call it Page Content Rank (PCR) in the paper, consolidates various heuristics that appear to be critical for breaking down the substance of Web pages. The page significance is resolved on the base of the significance of terms which the page contains. The significance of a term is determined concern-ing a given inquiry q and it depends on its measurable and linguistic elements. As a source set of pages for mining we utilize an arrangement of pages reacted by a web search tool to the question q. PCR utilizes a neural system as its inward order structure. We depict a usage of the proposed strategy and an examination of its outcomes with the other existing characterization framework –page rank algorithm.

Keywords: *Web Content Mining; Web Content Ranking; Page Ranking; Search Engine Optimization; Information Retrieval.*

1. Introduction

Data mining is basically to extract information or knowledge from a huge collection of data called as knowledge discovery of data in databases (KDD). Web comprises of extensive vault of heterogeneous information sources which mean extricating of information turns into an unpredictable errand for the clients. A great deal of troublesome is confronted which incorporates finding of significant information, how and what to gather from the accessible data. Information Retrieval is a technique used in Data Mining to retrieve related documents and information by searching huge data bases. Information Retrieval provides with wide range of diverse application for information extraction, clustering, searching in the web for related information, document classification etc. With the increase of data the number of web pages is continuously increasing, with this the number of queries given to the search engine for retrieval is growing rapidly. The search engine is said to be efficient as it uses a ranking mechanism which proves in better retrieval mechanism.

A discussion on page ranking presupposes knowledge of web mining, The major tasks of web mining are resource finding, information selection, preprocessing, generalization and analysis. First, data are extracted from online or offline text data available on the web. The next step is automatic selection and preprocessing from the retrieved web resources. The third step is an automatic discovery of a general pattern at individual or multiple sites. Finally, the results are validated and the analysis arrived at plays a major role in pattern mining. Types of web mining include web content mining, web structure mining and web usage mining. Various ranking algorithms are used depending upon the different web mining techniques like web structure mining which deals with modeling the link structure of the document exhibiting the popularity of the pages, web content mining which is retrieving of information based on its content i.e.; how relevant the content of the page is with respect o the query searched and lastly web usage mining depending upon the behavior of user how they interact with the websites.

1.1. Web content mining

The discovery of useful information forms web content/ data/ documents, and the process is also known as text mining, which is scanning and mining the text, pictures and graphs of a web page to determine the relevance of the content to a search query, and is related to data mining because lots of techniques used in data mining are also used in web content mining. It is the process of retrieving information from the WWW into a more structured form, and provides results lists to search engines in order of the highest degree of their relevance to the keywords in a query. Also, it does not provide information about the structure of the content that users are searching for or the various categories of documents found.

1.2. Web structure mining

This is the process of discovering a model of the link structure of web pages. For the purpose of generating data, similarities and relationships are established using hyperlinks. Both page rank and hyperlink analysis fall into this category, the idea being to generate a structured summary of a website and a web page. Accordingly, web structure mining can be divided into two kinds to minimize two chief problems the WWW comes up against as a result of the vast amount of information at its disposal. The first problem has to do with irrelevant search results, and the next is the inability to index the vast volume of information provided on the web.

1.3. Web usage mining

This refers to the automatic discovery and analysis of patterns in a click stream and associated data collected or generated as a result of user interactions with web resources on one or more websites. Table below represents web mining categories with views of data, main data, representations and methods of web content mining, web structured mining and web usage mining [11], [12].

Table 1: Web Mining Categories.

Search Engine	2012	2013	2014	2015	Change in % for last two years
GOOGLE	82	77	83	78	-5
BING	81	76	73	72	-1
YAHOO	78	76	71	75	4
MSN	78	74	73	74	1
AOL	74	71	70	74	4

1.4. Architecture

Data mining is used to extract information from a huge collection of data called as knowledge discovery of data in databases (KDD). Web comprises of extensive vault of heterogeneous information sources which mean extricating of information turns into an unpredictable errand for the clients. A great deal of troublesome is confronted which incorporates finding of significant information, how and what to gather from the accessible data.
Various ranking algorithms are used depending upon the different web mining techniques like web structure mining which deals with modeling the link structure of the document exhibiting the popularity of the pages, web content mining which is retrieving of information based on its content i.e.; how relevant the content of the page is with respect to the query searched and lastly web usage mining depending upon the behavior of user how they interact with the websites.

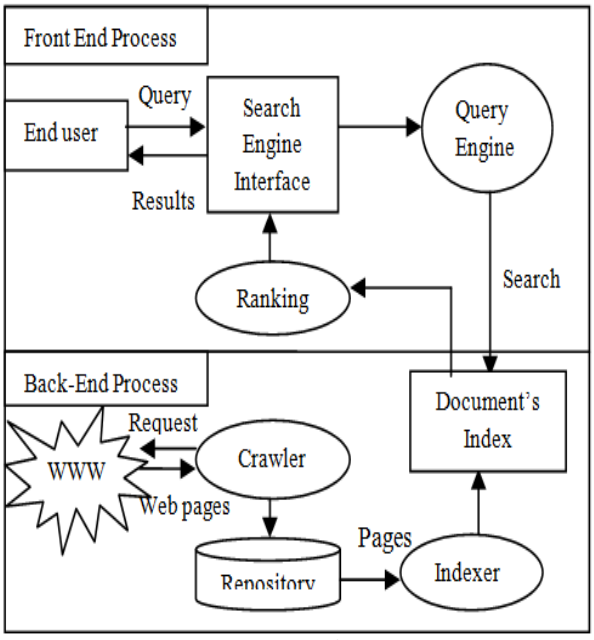

Fig 1: Block Structure of Search Engine.

Information Retrieval is a technique used in Data Mining to retrieve related documents and information by searching huge databases. IR provides with wide range of diverse application for information extraction, clustering, searching in the web for related information, document classification etc. With the increase of data the number of web pages is continuously increasing, with this the number of queries given to the search engine for retrieval is growing rapidly. The search engine is said to be efficient is it uses a ranking mechanism which proves in better retrieval mechanism.
The fundamental segments of web search tool which makes the query item noteworthy are indexer, the crawler and the instrument utilized for positioning [1]. The crawler is a robot which skims through the web, peruses the pages and the data present in numerous sites and aides in web creating so as to index sections for the pursuit list. At the point when client enters an inquiry, the catchphrases indicated in the question are exchanged on the internet searcher interface, where the question area coordinates the watchwords against the alphabetic list kept up with the indexer and present the client with the individual URL's of the concerned website page. By this technique a number of web pages will be retrieved so in order to make the retrieval efficient a ranking mechanism is used to present the users with the most relevant page giving it a higher ranking using some page ranking algorithm. This makes the search outcome relevant for the user.

1.5. The basics of search engine

The web page how it is visible to us is entirely different from how it is interpreted by the search engine [2]. In the below paragraph basically how a web page should be so that it is structured and well implemented both by the search engine and the users.
The first factor involved in its implementation is:

1.5.1. Index able content

For better search of pages by the engine it is important that a web page should be in HTML text format. Other elements such as images, plug-in files, the search engine crawlers should eliminate flash files. One way to ensure that the words or phrases which are given by the user to the search engine should be included in html text of the page.
Other more advance methods to perform the above said thing may be by:

1) Assigning images in any format in html so that we can give a text description of the visual content.
2) If we provide a search box which provide a link to go back and forth for easy retrieval.
3) Along with the text if we supply with flash files or plug-in.

1.5.2. Usage of the keyword

Keywords are the essential element of the search process. The entire process of information retrieval is based on the usage of keywords [7-8]. As the web crawler perform its crawling function and records the contents of the pages and they keep a track of those pages in keyword based files instead of putting away those million pages all together in one database. Different little databases all focusing on a specific keyword term or phrase permit the web indexes to recover the information they require in few moments. The keywords play an important role how we interact with the search engine. The order of words, spellings, punctuation provides additional information to the search engine which helps in better ranking of the pages and as well as better retrieval also.

1.6. Threats

Duplicate version of content also poses a threat but this problem is avoided to some extent by the search engine in last few years by detecting those pages and assigning those pages with lower ranking.

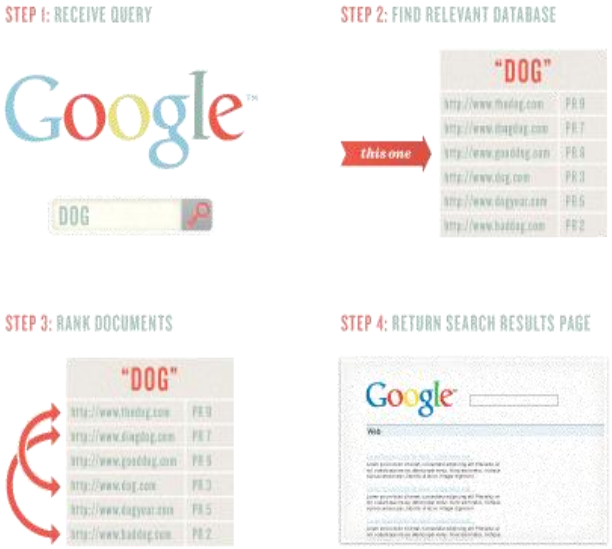

Fig. 2: Process of Information Retrieval.

Canonicalization is another problem faced where two or more different URL'S contain the two or more duplicate version of the web page. Duplicate version of the content may appear on different pages. If multiple version of the content exist as displayed in the above figure where diamond keyword is present in many websites will result in error where more ranked page will not be displayed.

1.6.1. Functions of search engine

The search engine basically aims to improve their performance by providing best possible results of the user query. It aims to provide the kind of pages that satisfy the searchers.
Generally, these sites have various functions in common:

- It facilitates ease of use, navigation, and understanding.
- It provides direct, actionable information which are relevant to the user query.
- It is designed by professionals and is thus easily accessible to modern browsers.
- Deliver high quality content which are legitimate and valuable.

1.6.2. Effects of search engine

- Usability and user experience greatly influence search engine ranking success. These factors provide indirect but valuable experience to the sites popularity and thus influence the ranking of the page.
- Engagement Metric: At the point when a web crawler conveys a page of results to you, it can gauge the achievement of the rankings by watching how you connect with those outcomes. It mainly affects content based search engine ranking [8].
- Machine learning and linking patterns also affect the content based search engine rankings [9].

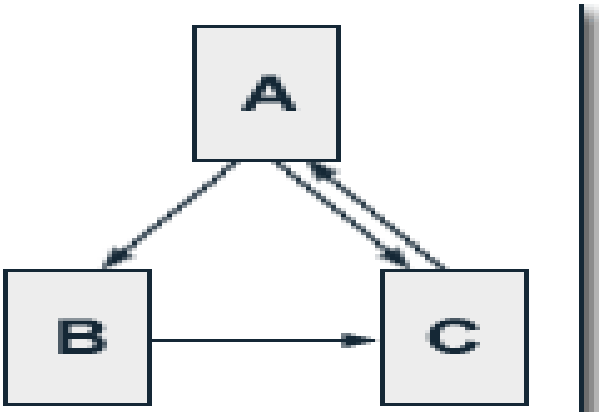

Fig. 3: Linking of Web Pages.

2. Existing algorithms

Generally to rank a web page different criteria's are used by ranking algorithms like page rank, hits, link analysis which is used by Google to rank the pages. But this manual approach using the above said algorithm is a hectic task and although this approach may be efficient but to make web page ranking more accurate we go for content based ranking (PCR) and page ranking where more preference is given based on the importance of terms, multimedia files, images that a web page contains.

2.1. Page ranking algorithm

The ranking algorithm is important as the result of the searched query must display the efficient retrieval page then only the search engine is said to be efficient. This algorithm uses the link structure of the document to decide the page score [3-5]. For example if a web page is having more out links the page score will be higher than compared to other pages and will be displayed with higher rank and on the top of the resulting web pages displayed. This can be better understood with help of an example depicting how a rank for a page is calculated.

2.2. Content based page ranking

Content based page ranking purely depends on the content that is present in the web page. Client request is handled for web search tool to acquire the outcomes. Seek results are removed and sent for pre-handling. Pre-Processing is an essential stride in content based mining [6]. Real world information have a tendency to be noisy, inadequate and conflicting. Information pre-handling strategies can enhance the nature of the information, in this manner enhancing the exactness and effectiveness of the consequent mining process [10]. Information pre-processing is a critical stride in the learning revelation process, since quality choices must be founded on quality information. All client enquiry, watchwords and substance words are pre-processed to evacuate boisterous words. After the synonyms were built using the dictionary, comparison of keywords is done.

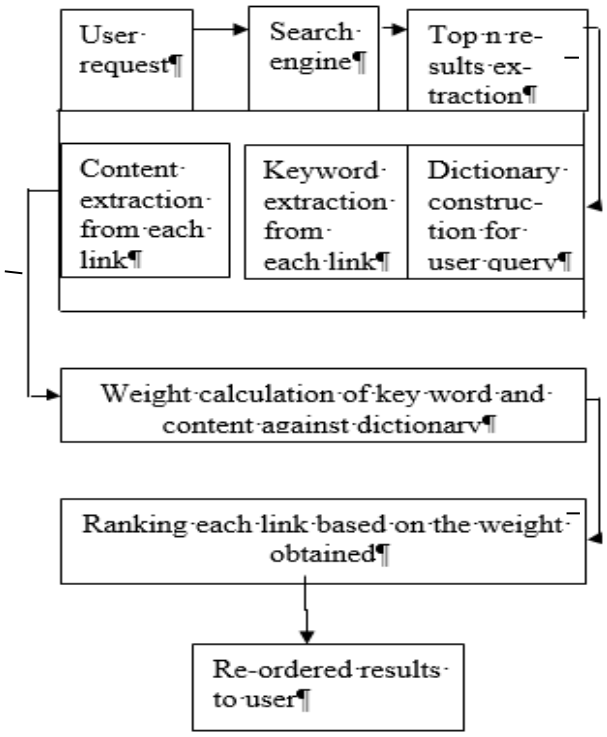

Fig. 4: Content Based Ranking Architecture Design.

Then if any redundancy is found points are given according to their position at last all coordinated keywords are condensed and standardized so that the total aggregate must be not exactly or equivalent to one.

3. Proposed system

The Page Content Ranking identifies the pages which are more significant in response to a search based on the content of the document and it better explains the web content mining. It helps in obtaining a ranked cluster with respect to a given query and this must not only be specifically from cluster of pages but also from the cluster of document. The linking of real time examples with the theory discussed helps us understand the process to further optimize a web site and also tells us how effectively and in less time the site can be analyzed well. Also various tools can be used to produce reports and these are well analyzed accordingly to give a clear view for the further enhancements of the website. So we know that we have content based analysis by doing that we get some database by combining both page ranking and content ranking we showed the better results.

4. Module description

4.1. Admin module

In this module we have done

- Insertion of keywords.
- Deletion of link from database.
- View all links that are presented in database.

4.2. User module

In this module we have done

- Query interface.
- View of links as per keyword entered in search bar.

4.3. Ranking module

In this module we have done

- Collecting the web pages from the static keywords.
- Giving ranks to those pages.

5. Performance analysis

In this project we propose a method that can simplify this problem and come out with a better performance of search engine. In this paper we use both page ranking database and content ranking database to get good relevancy of pages so that the user can get the information in top links that are displayed. The final output is with better results when compared to both page ranking and page content ranking. It will get the input that is needed, through the database which is initially taken with keywords and links. Page rank is given statically using the SEO tools. Then when both content and page ranking are checked and a change in ranking is done. Then again according to that ranking the positions of the links change i.e, the first link may move to last and the last one to first. All this process will be done using java and SQL. Java platform will be perfect to show the things and SQL to store the database. Entire project database is stored in SQL and the things are displayed to the users using java platform.

6. Conclusion

The Page Content Ranking identifies the pages which are more significant in response to a search based on the content of the document and it better explains the web content mining. It helps in obtaining a ranked cluster with respect to a given query and this must not only be specifically from cluster of pages but also from the cluster of document. It is efficient method then the page rank which ignores whether or not the page is relevant to the query at hand. This gives a better view of things that happen in a search engine. Reusability, interoperability and extensibility are the major aspects of content development techniques. As we know that in page ranking mechanism every page is retrieved from database based on the no of links in the page. Pages that are displaying on search engine may or may not have the correct information that is needed to the user. So, the user need to surf the all the pages which are displayed to obtain needed information. It becomes hectic to all the users. To solve this problem we implement the project using both page ranking and content based page ranking. In which ranking is given based on the weights of the keyword.

7. Future Scope

We have done work related to web page relevancy in this we have taken content ranking by considering the weight of keyword, in the mere future web page relevancy can be done more precisely.

References

[1] J. Singh Chouhan and A. Gadwal, “Improving web search user query relevance using content based page rank”, International Conference on Computer, Communication and Control (IC4), Indore, 2015, pp. 1-5. (2015)

[2] Sudhakar, P., G. Poonkuzhali, and R. Kishore Kumar, “Content Based Ranking for Search Engines”, Proceedings of International Multi Conference of Engineers and Computer Scientists (IMECS 12).March 14-16, 2012, Hongkong. (2012)

[3] Shalya, Nidhi, Shashwat Shukla, and Deepak Arora, “An Effective Content Based Web Page Ranking Approach”, International Journal of Engineering Science and Technology, (IJEST), Vol. 4, No. 08 (2012).

[4] Harish Kumar B T, Vibha Lakshmikantha, Venugopal K R, “Content Based Web Page Re-Ranking Using Relevancy Algorithm” Journal of Electronics and Communication Engineering Research, Vol.2, No.7, pp.1-8. (2014)

[5] Pokorny, Jaroslav, and Jozef Smizansky, “Page content rank: an approach to the web content mining”, Proceedings of the IADIS International Conference on Applied Computing, Vol. 2, pp. 22-25. (2005)

[6] Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto, “ACM press”, Modern Information Retrieval, Addison-Wesley (1999).

[7] Kosala, Raymond, and Hendrik Blockeel, "Web mining research: A survey", ACM Sigkdd Explorations Newsletter 2, No. 1, pp. 1-15. (2000)
[8] Frikh, Bouchra, Brahim Ouhbi, and Amine Ameur, "A comparative study of link analysis algorithms for information retrieval", Next Generation Networks and Services (NGNS), 2012, pp. 54-58. IEEE, (2012)
[9] Van Meteren, Robin, and Maarten Van Someren, "Using content-based filtering for recommendation", Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop, pp. 47-56. (2000)
[10] Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava, "Web mining: Information and pattern discovery on the world wide web", Proceedings of Ninth IEEE International Conference on Tools with Artificial Intelligence, 1997, pp. 558-567. IEEE, (1997)
[11] Selvan, Mercy Paul, A. Chandra Sekar, and A. Priya Dharshini, "Survey on web page ranking algorithms", International Journal of Computer Applications, Vol. 41, No. 19 (2012).
[12] Rani, Seema, and Upasana Garg, "A Review Paper on Web Page Ranking Algorithms", International Journal of Engineering and Computer Science, Vol. 3, No. 8, pp. 7946-7949. (2014).