# Devising a Methodology for Link Analysis by Reducing Noise

Mamta M. Hegde
MIT, Pune
Department of Computer Engineering
Pune, India

Prof. M. V. Phatak
MIT, Pune
Department of Computer Engineering
Pune, India

***Abstract-* With the huge volume of web pages that exist in today's world, search engines play a vital role in the current Internet. But even if they allow finding relevant pages for any search topic, nowadays the number of results returned is often too big to be carefully explored. The need of the users varies, so that what may be interesting for one may be completely irrelevant to another. Hyperlink Analysis is the name given to a collection of techniques that have emerged to analyze the hyperlink structure that exists in the Web. This paper proposes a method for ordering the Web pages returned from search engines so as to allow relevant URL's to be ranked higher. The factors include keyword popularity, keyword to Web page popularity, and Web page popularity; hit score, active time spent by the user on a web page. These factors capture the preferences of the users. Using these factors we are able to rank more popular pages higher, which will help most users find the more popular and the more relevant pages. The objective of the approach is to provide right solutions to acquire information on user topic of interest and identify noise while reducing it, thus efficient mining can made be possible.**

***Keywords— Web mining, Web Structure mining, Hyperlink analysis, Noise reduction***

## I. Introduction

Finding information relevant to what we are seeking is becoming more important as the Web is growing in explosive speed. Nowadays, most people try to find information on the Web by using search engines. Given a few search keywords, most search engines today will retrieve more than a few thousand Web pages. The problem now is that we need to scan pages after pages, manually and time consumedly, to find what we need or often give up without getting the needed information. We need to address the problem of helping Web users to find the information that they need. There are several approaches to address the problem. The currently most popular method to address the problem is by ordering the search results and presenting to the users the most relevant pages first. This method is called page ranking, which is one of the important factors that makes Ranking seem to help Web users find the needed information quicker [6].

In this paper, focus is on the ranking relevant URL's based on user query and reducing noise. It attempts to improve existing page ranking methods, to allow relevant Web pages to be ranked higher. The system attempts to capture the search history and the preferences of search engine users to better serve their needs.

This paper is organized as follows: In section II an overview of hyperlink analysis is provided. In section III literature review is done. In section IV our approach and basics are discussed. In section V Proposed algorithm is described followed by section VI which describes experimental analysis and finally, in section VII the paper is concluded.

## II. HYPERLINK ANALYSIS

Hyperlink Analysis by itself is a part of bigger research. Hyperlink analysis can be used for a variety of purposes. Some of the main uses are:

- Measuring the extent of support that the ideas and statements on a page provide for a particular topic. This information also helps to rank Web pages according to their relative importance.
- Serving as an effective tool in classifying Web pages according to various topics and functionalities.
- Improving the efficiency of crawling by identifying the relative importance of pages that need to be crawled first.
- When combined with usage statistics, hyperlink analysis can be used for predicting user-browsing behavior and help the user to surf the Web better.

The methodology for using hyperlink analysis for an application can be described as the following sequence of steps:

- Analyse the needs of the application to determine the type of information it needs from hyperlink analysis. For example, the web search application requires that pages that are relevant to a user query be ranked in some order of importance. The information model here is a ranked list of URLs.
- Next, determine the metric(s) that need to be calculated to quantify various aspects of the information model. For example, for Google the metric is PageRank, while in the HITS approach it is Hub Score and Authority Score. As newer applications of hyperlink analysis are being discovered, new metrics will have to be developed to suit their needs. Algorithms to compute the selected metrics need to be selected/designed next.
- Next, the analysis scope relevant to the application must be decided. The choices are single page level, groups of pages and links, or an entire graph.
- Finally, it must be decided if hyperlink analysis is to be done just by itself, or in conjunction with web content and

web usage analysis. If so, then the results must be integrated with those of the other analysis [11].

## III. LITERATURE REVIEW

### A. *Traditional Data analysis approach*

One approach is to use search engines, which can locate interesting information based on user’s requirement or interests. Another approach is to apply traditional data mining techniques to get hidden knowledge from web data. Comparing with traditional data, web data are different in the following ways:

- The amount of data on the web is massive.
- Web data is semi structured. Also the format of web data differs a lot; it can be an HTML document, an XML document, an audio file, a video file, an image or TXT file etc.
- Web data is highly dynamic. Data on the web change in different ways at any time.

Only a small portion of web data is useful. A particular individual is interested in only a tiny portion of web data.

### B. *Modern Data analysis approach*

One approach for web mining is to transform web data into certain formats such that traditional mining techniques can be applied. Another approach is to modify traditional mining techniques so that they can handle web data. Based on the types of data being mined, web mining can be divided into web content mining, web structure mining and web usage mining.

- Web content mining focuses on extracting knowledge from text of web document such as HTML documents, XML documents etc.
- Web structure mining is to extract useful structure of the overall web or sets of web documents. It can be used to construct web communities, identify authoritative web pages, and rank the search result according to the connectivity and assigning weights to web pages.
- Web usage mining is to discover information from web log files. By using different techniques we can get the patterns of how web surfers navigate the web, consequently we can provide personalize web services and pre fetch web documents.

### C. *Challenges And Issues of Hyperlink analysis approach*

One issue so-called dead links; links to a Wiki page that doesn't yet exist or which are pointed to non-index able Webpages. A Webpage is considered dangling when it doesn't have any followable links on it. Although this situation could exist on the real internet, it is mostly an artifact of not having downloaded all pages that need to be evaluated. This issue arises because it is virtually impossible to download all pages on the internet.

Another issue is that link analysis works best on queries that will have a lot of results for more specific queries, merging the ranks for the web pages as calculated by PageRank with ranks calculated by traditional information retrieval scoring methods. It is also a challenge to rank web pages in the order of their significance, both overall as well as pertaining to a particular query. There are many aspects of a web page that make it relevant such as:

- Web page changes;
- The frequency of Web page changes;
- Keyword changes and keyword count changes;
- The number of new backlinks; and
- Data availability and stability.

The above features are quite replicable. Competing profit-seeking ventures may manipulate their web pages' properties in order to project a false importance within the web, thus skewing search results enormously. Any evaluation strategy that counts replicable features of web pages is prone to manipulation [7].

Scalability is becoming a challenge to the system. With the growth of numbers of users and web documents, the system needs more resources for processing information and forming recommendations. Majority of resources are consumed with the purpose of determining users with similar interests.

## IV. OUR APPROACH AND BASICS

Hyperlink analysis with noise reduction system is used to recommend and rank relevant URL’s and web pages to the user based on his search and calculation of popularity score and importance of the page. There is lots of information available on internet and the user should get what he needs else he remains unsatisfied. The system is helpful to recommend desired pages and URL’s to the users. The application can be used for personalized web search based on users browsing history. Proposed approach does the following:

1. Search information quickly and correctly with respect to the query proximity.
2. Automatic concept generation and clustering the retrieved information.
3. Link Analysis and identify dead links, reduce primary noises.

## V. PROPOSED ALGORITHM

The proposed Hyperlink analysis with noise reduction system is used to display most relevant URLS’ to the user. According to the user query web pages are fetched and displayed to the user. The steps in the algorithm could be briefly summarised as follows:

1. Preprocessing
2. Clustering is done to group web search results.
3. Calculate popularity scores.
4. Calculate active time spent by user on page
5. Identify noise and reduce them.
6. Sort results and display most relevant links to the user.

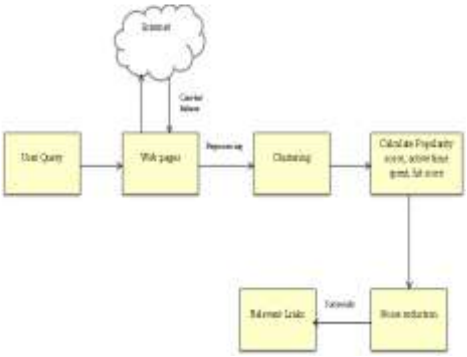

Figure 1. Architecture of Proposed approach

- Data Preprocessing

In web search results clustering, it is the web snippets that serve as the input data for the grouping algorithm. Due to the rather small size of the snippets and the fact that they are automatically generated summaries of the original documents, proper data pre-processing is of enormous importance. Data preprocessing has following stages:

- Text filtering
- Language identification
- Stemming
- Mark Stop word

In the text filtering step, all terms that are useless or would introduce noise in cluster labels are removed from the input documents. Among such terms are:

• HTML tags (e.g. <table>) and entities (e.g. &amp;)

• Non-letter characters such as "$", "%" or "#" (except white spaces and sentence markers such as '.', '?' or '!')

In the Language identification step, before proceeding with stemming and stop words marking, for each input document separately, LINGO tries to recognise its language. In this way, for each snippet, appropriate stemming algorithm and stoplist can be selected.

In Stemming, if the required stemmer is available, inflection suffixes and prefixes are removed from each term appearing in the input collection. This guarantees that all inflected forms of a term are treated as one single term, which increases their descriptive power.

While Stop words marking, although they alone do not present any descriptive value, stop words may help to understand or disambiguate the meaning of a phrase.

- Clustering

Clustering is a process of forming groups (clusters) of similar objects from a given set of inputs. Good clusters have the characteristic that objects belonging to the same cluster are "similar" to each other, while objects from two different clusters are "dissimilar". Clustering has following stages:

- Feature extraction
- Cluster label induction
- Cluster content discovery
- Final cluster formation [8]

The aim of the feature extraction phase is to discover phrases and single terms that will potentially be capable of explaining the verbal meaning behind the LSI-found abstract concepts. To be considered as a candidate for a cluster label, a phrase or term must: Appear in the input documents at least a specified number of times. Cluster Label Induction consists of following steps:

- Term-document matrix building - build the term-document matrix A for the input snippet collection, as index terms use the non-stop words that exceed the predefined term frequency threshold. Use the tf-idf (term frequency inverse document frequency) weighting scheme; which aims at balancing the local and the global term occurrences in the documents. In this scheme

$$aij=tfij\cdot \log(N/dfi)$$

Where *tfij* is the term frequency, *dfi* denotes the number of documents in which term *i* appears, and *N* represents the total number of documents in the collection. The log(*N*/*dfi*), which is very often referred to as the *idf* (inverse document frequency) factor, accounts for the global weighting of term *i*.

Abstract concept discovery - To achieve this, the original term-document matrix is approximated by a limited number of orthogonal factors – the column vectors of the SVD's *U* matrix .Intuitively, the factors can be perceived as a set of abstract concepts each of which conveys some common meaning present in a subset of the input collection. Now perform the Singular Value Decomposition of the term-document matrix to obtain U, S and V matrices; based on the value of the q parameter and using the S matrix - calculate the desired number k of abstract concepts; use the first k columns of the U matrix to form the Uk matrix;

$$\frac{\| A_k \|_F}{\| A \|_F} = \frac{\sqrt{\sum_{i=1}^{k} (\sigma_i^2)}}{\sqrt{\sum_{i=1}^{r_A} (\sigma_i^2)}} \geq q$$

- Phrase matching - using the tf-idf term weighting create the phrase matrix P; for each column of the Uk matrix {multiplicate the column by the P matrix; find the largest value in the resulting vector to determine the best matching phrase;}
- Candidate label pruning - calculate similarities between all pairs of candidate labels; form groups of labels that exceed a predefined similarity threshold; For each group of similar labels {Select one label with the highest score}

In the cluster content discovery phase, the classic Vector Space Model is used to assign the input snippets to the cluster labels induced in the previous phase.

$$\cos\theta_j = \frac{a_j^T q}{\| a_j \| \| q \|} = \frac{\sum_{i=1}^{t} a_{ij} q_i}{\sqrt{\sum_{i=1}^{t} a_{ij}^2}\sqrt{\sum_{i=1}^{t} q_i^2}}$$

- In the final phase of the algorithm, cluster scores are calculated according to the following formula: *cluster-score = label-score * member-count*
- In the presentation interface the resulting clusters are sorted according to their score, the user will be presented with the well-described and relatively large groups in the first place [8].
- Calculate Popularity Score and importance of page – Web users interact with search engines by providing several search keywords and selecting Web pages from the search results. .The system attempts to capture as much usage information as possible and to make use of captured information.

- Keyword popularity - When a user entered keywords and clicks search, the system will store the keywords in the database. Each of the term will be associated with a weight that records the frequency that the term has been used.
- Keyword to Web page popularity- After the search engine returns the search results to the user, the user will select Web pages for viewing. The relationships between the search keywords and the selected Web pages will be recorded. The relationships capture the preferences of the users.
- Web page popularity- By highlighting the number of popular keywords contained in the page. The idea is that if a Web page contains a large number of popular keywords, then it should be considered as more popular [6].

$$Pop(q, d) = kPnorm \sum_{t \in q} keywordPop(t, d) + keywordwebpagePop(q, d).kWpnorm + WebpagePop(d).WPnorm$$

- Calculate active time spent by user on a page – Web page popularity should be accompanied by measuring the amount of time a user spent interacting on the Web page. Time on page or visit duration can be an indication of the level of interest or involvement that a visitor has with the web page.
- Noise reduction

Hyperlink structure usually reflects the implicit logical relationship among web pages. Core information, redundant information and hidden information are the three types of Web document data. The content that a user needs to view from a Web page is known as Core information. To improve Web content convenience or business attractiveness redundant information is used. Web documents also comprise 'hidden information' like HTML tags, script language and programming comments, since it is not visible for end users. Several challenges must be surmounted to extract information from these pages. Recognition and elimination of noise are the vital problem for extraction of information from the web. Web pages more often than not contain different contents, which are relevant or irrelevant with the key topic. In addition to primary content, web pages commonly have image-maps, logos; advertisements search boxes, footers and headers, navigational links, related links and copyright information along with the primary content. The approach implements following steps for noise reduction:

- Identify irrelevant links- the aim is to order the search results based on query proximity such that where search terms are in close proximity those URL's appear at top of the search results and others are considered noise in relation to the search term.
- Identify and remove primary noises- the approach emphasizes to remove primary noises - Navigation bars, Page Headers and Footers, Copyright and Privacy Notices, Advertisements, Make a donation and Logos.
- Identify and display dead links to the user from whom the response code from server is other than 200 (like 404 – Page does not exist) to minimise user access time.
- Hit score– A hit is a request to a web server for a file, like a web page, image. The click through score is the measure of the number of times an URL is clicked by the user for a specified keyword.

## VI. EXPERIMENTAL ANALYSIS

To evaluate the effectiveness of the system, performance is measured using two factors like time and accuracy. Accuracy is the measure of relevant links to the total links fetched in percentage. Time is active time spent by the user while interacting with the page. So based on above proposed system we have worked on practical evaluation using JAVA. We have done implementation through the web application as shown in the following Figure 2:

Figure 2. Search results for Hyperlink analysis and noise reduction system

Following graphs show the performance evaluation of the proposed algorithm based on our popularity index and other factors. The following table readings and graphs show the improved performance as compared to existing system. Comparative study of accuracy between proposed and existing method based on number of pages ranked. Table 1 and 2 shows the readings we got during our practical analysis and Figure 3 and 4 shows the graph for those readings: The table and the following graph shows that the proposed approach shows improved results as compared to the previous approach.

Table 1: Accuracy comparative analysis

| No. of Search results | % Accuracy (relevant links) | |
|---|---|---|
| | Hyperlink analysis and noise reduction system | Existing system |
| 1 | 70 | 40 |
| 2 | 32 | 19 |
| 3 | 82 | 53 |
| 4 | 60 | 47 |
| 5 | 61 | 55 |
| 6 | 59 | 62 |
| 7 | 78 | 70 |
| 8 | 39 | 21 |
| 9 | 50 | 20 |

Table 2: Time comparative analysis

| No. of Search results | Time in sec | |
|---|---|---|
| | Hyperlink analysis and noise reduction system | Existing system |
| 1 | 12 | 29 |
| 2 | 17 | 23 |
| 3 | 62 | 68 |
| 4 | 48 | 76 |
| 5 | 33 | 86 |
| 6 | 48 | 76 |
| 7 | 12 | 37 |
| 8 | 28 | 68 |
| 9 | 23 | 59 |

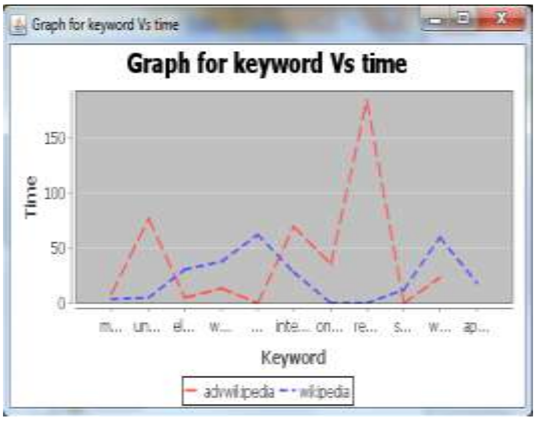

Figure 3. Time spent comparative analysis

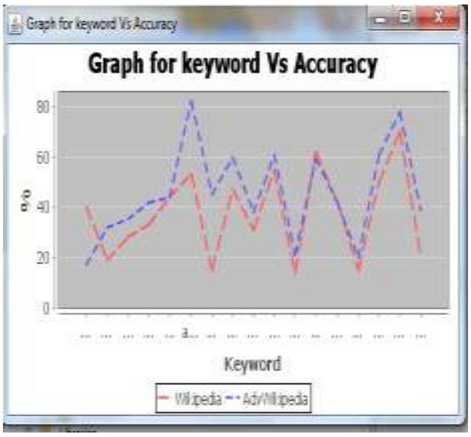

Figure 4. Accuracy comparative analysis

## VII. CONCLUSION AND FUTURE SCOPE

Web mining technologies are right solutions for knowledge discovery / extraction on the web. Web structure mining is a new area of research. The important decision can be done regarding structure of the website i.e. the excellent webpage‘s will be moved very near to the home page. The pages with more hit count can be given the preference to be brought closer to the home page. The system focuses on the link analysis and calculates the importance of webpage by displaying most relevant links.

The Hyperlink analysis and noise reduction system is used to present users the most relevant links on the topic of their interest. The application can be used for personalized recommendation to give personalized relevance based on user's profile. In the system some results are relevant to some users under certain conditions but may not be relevant to other users because idea of relevance is subjective and hence cannot be measured, more future research should be done in this direction. Another future research direction is to create new type of search engine that allows the users to have more interaction and control.

## REFERENCES

1. Sekhar Babu Boddu, V.P Krishna Anne, Rajesekhara Rao Kurra, Durgesh Kumar Mishra, ‟ Knowledge Discovery and Retrieval on World Wide Web Using Web Structure Mining”, Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation 2010.
2. Soumen Chakrabarti, Byron E. Dom, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, David Gibson, Jon Kleinberg,‟Mining the Web's Link Structure”,IEEE 1999.
3. P. Sivakumar, R. M. S Parvathi, "An Efficient Approach of Noise Removal from Web Page for Effectual Web Content Mining", European Journal of Scientific Research ISSN 1450-216X Vol.50 No.3 (2011), pp.340-351 © EuroJournals Publishing, Inc. 2011.
4. Feng Li, "Extracting Structure of Web Site Based on Hyperlink Analysis‟, IEEE 2008.
5. Lili Yan, Yingbin Wei, Zhanji Gui , Yizhuo Chen,‟ Research on PageRank and Hyperlink-Induced Topic Search in Web Structure Mining", IEEE 2011.
6. Ben Choi,Sumit Tyagi,"Ranking Web Pages relevant to search keywords", ISBN:978-972-8924-93-5 © 2009 IADIS.
7. Monika R Henzinger ,"Hyperlink Analysis for the Web", IEEE internet computing, 2001.
8. Stanislaw Osinski, "Algorithm for clustering of web search results" , Carrot 2 Search, 2003.
9. P Ravi Kumar, A. K. Singh, "Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval", American Journal of Applied Sciences 7 (6): 840-845, 2010 ISSN 1546-9239 © 2010Science Publications.
10. Monika R Henzinger,"Link Analysis in Web Information Retrieval", Computer Society Technical Committee on Data Engineering, IEEE 2000.
11. Pang Ning Tan, Jaideep S,Prasanna D,Vipin K,"Hyperlink Analysis :Techniques and Applications", High Performance Computing Research Center.