FlexiRank: An Algorithm Offering Flexibility and Accuracy for Ranking the Web Pages

Debajyoti Mukhopadhyay[1] and Pradipta Biswas[2]

[1] Cellular Automata Research Lab, Techno India,
(affiliated to W.B. University of Technology),
EM 4/1 Salt Lake Sector V, Calcutta 700091, India
debm@vsnl.com
[2] Indian Institute of Technology, School of Information Technology,
Kharagpur 721302, India
pbiswas@sit.iitkgp.ernet.in

Abstract. The existing search engines sometimes give unsatisfactory search result for lack of any categorization. If there is some means to know the preference of user about the search result and rank pages accordingly, the result will be more useful and accurate to the user. In the present paper a web page ranking algorithm is proposed based on syntactic classification of web pages. The proposed approach mainly consists of three steps: select some properties of web pages based on user's demand, measure them, and give different weightage to each property during ranking for different types of pages. The existence of syntactic classification is supported by running fuzzy c-means algorithm and neural network classifier on a set of web pages. It has been demonstrated that, for different types of pages, the same query string has produced different page ranking.

1 Introduction

Web page ranking algorithms are used to order web pages according to their relevance. Exactly what information the user wants is unpredictable. So the web page ranking algorithms are designed to anticipate the user requirements from various static (e.g., number of hyperlinks, textual content) and dynamic (e.g., popularity) features. The goal of the present paper is to introduce an algorithm called FlexiRank to offer some flexibility to the user while searching the web pages. A search engine interface is incorporated with some option buttons to fine-tune the options while sending the query to the search engine. The option buttons are easy to use for naïve users and not as complicated as some of the existing advanced search engine interfaces.

2 Related Work

Among the existing page ranking algorithms the most important algorithms are Kleinberg's HITS algorithm, Brin & Page's PageRank algorithm, SALSA algorithm,

G. Chakraborty (Ed.): ICDCIT 2005, LNCS 3816, pp. 308 – 313, 2005.

CLEVER Project etc. The AltaVista Search Engine implements HITS algorithm. But the HITS (Hyperlink Induced Topic Search) is a purely link structure-based computation, ignoring the textual content [1]. According to PageRank algorithm used in Google [2], a page has a high rank if the sum of the ranks of its back-links is high. CLEVER project [3] mainly emphasizes on enhancements to HITS algorithm, hypertext classification, focused crawling, mining communities, modeling the web as a graph. The weight assignment to hyperlinks is more exploited in [4] where each link gets a weight based on its position at the page, length of anchor text and on the tag where the link is inserted. In [5] the links of a web page are weighted based on the number of in-links and out-links of their reference pages. In [6] a new approach of dissecting queries into crisp and fuzzy part has been introduced. In [7], a parameter viz. query sensitiveness is measured which signifies the relevance of a document with respect to a term or topic. In [8], the damping factor of PageRank algorithm is changed to a parameter viz. confidence of a page with respect to a particular topic. The confidence is defined as the probability of accessing a particular page for a particular topic.

3 Our Approach

Approach taken in this paper is to make a classification of web pages based on only syntax of the page. This type of classification is independent to the semantics of the content of a page. The search engine interface is incorporated with some option buttons to take the proper class of a page along with the query topic. The web page classification will be like Index page, Home Page, Article, Definition, Advertisement Pages etc. As for example, if a search topic is given like "Antivirus Software" and given category of page is "Homepage" then the homepages of different Antivirus companies will get higher ranking. If for the same query, the category given is "Article", then the pages giving general description of Antivirus Software will get higher ranking. Again if the given category is "Index" then a page having large number of links to different antivirus software vendors will get higher ranking. Thus in the proposed page-ranking algorithm for a single query term, a particular page can get different ranking based on users' demand.

4 Parameters Used for Ranking

In this section different parameters, selected for web page ranking, are discussed. The page ranking will be done by taking a weighted average of all or some of the parameters. The weight given to a particular parameter will depend upon the category of the page. In the proposed algorithm a single query may give different ranking to a page depending on the category of the page-which is not possible in any existing search engines. The algorithm is flexible in the sense that just by changing the weights, the same algorithm satisfies user demands for different types of pages.

4.1 Relevance Weight

Relevance weight measures the relevance of a page with respect to a query topic by counting the number of occurrences of the query topic or part of the query topic within

the text of the document. In the present paper, the page relevance algorithm used has taken an approach of the Three Level Scoring method. In the proposed algorithm, firstly the words in "Stop List" are removed from the search string. After proper stemming, the relevant keywords or terms are extracted from the search string. Next, the occurrence of each term is found out, and a weightage is given to it as the ratio of its length to the length of the given query topic. As for example, for a query string "data mining," the term "data mining" will get a weightage of "1" whereas the term "mining" will get a weightage of "6/11" i.e., 0.545. Finally the algorithm is as follows:

```
function Calc_Relev_Wt(File F: A Text File, String S: The
Search String)
return Relev_weight
/* relevance of textual content of file F w.r.t. Search
string S */
var KEYWORD_SET[1…N]
/* To store the subset of relevant strings within the search
string */
var CNT /*Number of relevant substrings */
var OCCURRENCE[1…N]
/* OCCURRENCE[I]= Occurrences of substring KEYWORD_SET[I]
within file F */
KEYWORD_SET=Set of relevant substrings within S
CNT=|KEYWORD_SET|
For (I=1 to CNT)
OCCURRENCE[I]= Number of Occurrences of substring
KEYWORD_SET[I] within file F
For (I=1 to CNT)
Relev_Weight=Relev_Weight+(Length(KEYWORD_SET[I])/Length(S))*
OCCURRENCE[I]
```

4.2 Hub and Authority Weight

Authorities are pages that are recognized as providing significant, trustworthy, and useful information on a topic. Hubs are index pages that provide lots of useful links to relevant content pages. The authority value of page p is the sum of hub scores of all the pages that points to p and the hub value of page p is the sum of authority scores of all the pages that p points to. It has been observed that the small number of pages with the largest authority converged value should be the pages that had the best authorities for the topic.

4.3 Link Analysis of a Page

The HITS algorithm analyzes the link structure information of a web graph. The hyperlink information of a single page (e.g., number of hyperlinks, anchor text and positions of the pages in the domain tree with respect to a particular page) are also found to give useful information during syntactic categorization of a web page. The number of hyperlinks of a page is calculated by getting the total number of **a *href*** tags. For getting the exact number of hyperlinks the number of ***frame src*** tags should be added to the number of ***a href*** tags and links to the same page should be excluded. By analyzing anchor text the glossary pages can very easily be identified. It has been found the portals have large number of hyperlinks pointing to same level nodes in the

domain tree rooted at the next higher level node of the source of the page; e.g., if source is **a.b.com** nature of hyperlinks is **x.b.com** or **y.b.com**. The site maps and home pages have large number of hyperlinks pointing to lower level nodes in the domain tree rooted at the source of the page; e.g., if source is **a.b.com** nature of hyperlinks is **a.b.com/x**, **a.b.com/y**.

4.4 Types of Content

The syntactic analysis of the content also gives useful properties about the type of a page. Examples of this type of properties are: number of images in a page; proportion of text length to number of images; relevance weight of the query string within special tags like header tag, title tag, etc.

5 The FlexiRank Algorithm

The FlexiRank algorithm operates on a set of web pages returned by a web crawler and gives a ranking of the pages as output. It operates according to the following steps:

- **Select attributes based on user demand:** Based on the users' demand the algorithm chooses a set of properties of a web page. Some properties are chosen irrespective of the users' demand. Examples of such mandatory properties are Relevance weight, Hub weight and Authority weight. The other attributes are chosen based on user demand to provide an accurate ranking. Examples of such optional attributes are number of hyperlinks, number of images, properties of anchor text, etc.
- **Measure the attributes:** The selected attributes are measured for each web page.
- **Calculate rank:** The rank is calculated by taking a weighted average of the measured values. The weight assigned to each attribute is based on users' demand.

The algorithm provides flexibility in two grounds:

- **In selection of properties:** As for example when the users' demand is index type pages, number of hyperlinks of a page will be measured whereas number of images or text to image proportion will not be measured.
- **In determining weightages of properties:** The selected attributes get different weightage for difference in user demand. As for example, for article type of pages, relevance weight and authority weight will get highest weightage whereas for advertisement type of pages, number of thumbnails (i.e., number of images) and hub weight will get higher weightage.

Due to these varying selections of properties and their corresponding weightages, the algorithm provides more flexibility to the user and also gives more accurate result.

6 Experimental Results

The experiment has been done in two parts. In the first part, several web pages are downloaded and classified according to the proposed properties. In the second part, some web pages are downloaded again from an existing search engine and ranked according to the FlexiRank algorithm. Each of these parts is discussed below.

6.1 Clustering the Web Pages

In this part about 50 web pages are downloaded from Google search engine. The pages are clustered according to different properties like Relevance weight, Number of Images, Number of Links, Document Length etc. For clustering purpose, Fuzzy c-means algorithm is used. Cluster validation is done by Classification Entropy. With c = 4, we got a hint of the existence of syntactic classification. To confirm the existence of syntactic classification, we use a neural network software viz. NeuNet Pro downloaded from http://www.cormactech.com. Using this software we define a feed forward neural network with 5 hidden nodes and use back-propagation learning algorithm for classifying 30 web pages downloaded from Google. After completing 1000 cycles with learning rate=60 and verify rate=10 (these rates are defined by the software internally) we get the following scatter graph in Fig. 1 and time series graph in Fig. 2. Since the classification is carried on using only 7 properties, we do not get a very accurate classification. Still the result of the fuzzy clustering algorithm and the less than 20% R.M.S. error in classification confirm the existence of syntactic classification of web pages.

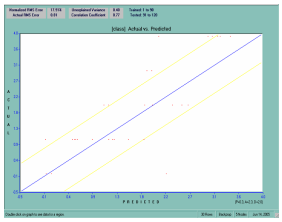

Fig. 1. Scatter graph for Syntactic Classification

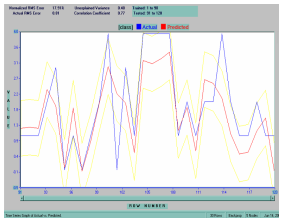

Fig. 2. Time Series Graph for Syntactic Classification

6.2 Ranking the Web Pages

For testing the actual change in ranking for different types of pages, the proposed ranking algorithm is run on top 30 pages downloaded using Google search engine with the search topic “Human Computer Interaction”. The screenshot of the proposed interface of a search engine is shown in Fig. 3. When the type of page is given as index, the following three pages get first three ranks:

1. http://is.twi.tudelft.nl/hci/
2. http://dmoz.org/Computers/Human-Computer_Interaction/
3. http://www-hcid.soi.city.ac.uk/

The first two pages are literally index pages while the third one is the home page of Centre of HCI Design, City University London. The page contains a lot of hyperlinks. Again when the type of page is given as article the following three pages get first three ranks:

1. http://sigchi.org/cdg/cdg2.html
2. http:// www.cs.cmu.edu/~amulet/papers/uihistory.tr.html
3. http://www.id-book.com/

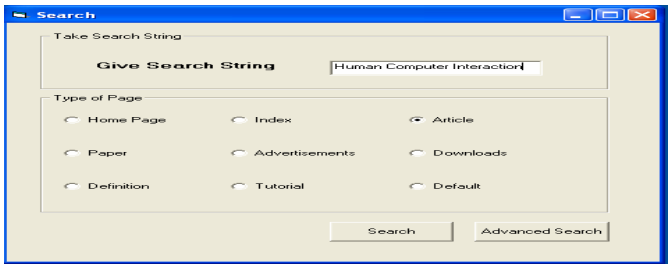

Fig. 3. Screenshot of the Proposed Interface of a Search Engine

Here also, the first two sites are text intensive articles. As can be seen in the interface a default option is also being kept for ranking all types of pages.

7 Conclusion

The present paper discusses a web page ranking algorithm, which consolidates web page classification with web page ranking to offer flexibility to the user as well as to produce more accurate search result. The classification is done based on several properties of a web page which are not dependent on the meaning of its content. The existence of this type of classification is supported by applying fuzzy c-means algorithm and neural network classification on a set of web pages. The typical interface of a web search engine is proposed to change to a more flexible interface which can take the type of the web page along with the search string.

References

1. Kleinberg, Jon; "Authoritative Sources in a Hyperlinked Environment;" Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998; pp. 668-677
2. Brin, Sergey; Page, Lawrence; "The Anatomy of a Large-Scale Hypertextual Web Search Engine;" 7th Int. WWW Conf. Proceedings, Brisbane, Australia; April 1998
3. Chakrabarti, S. et. al.,; "Mining the link structure of the World Wide Web;" IEEE Computer, 32(8), August 1999
4. Baeza-Yates,Ricardo; Davis, Emilio; "Web page ranking using link attributes," Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters**,** May 2004
5. Xing, W.; Ghorbani, A.; "Weighted PageRank algorithm;**"** Proceedings of the Second Annual Conference on Communication Networks and Services Research, 19-21 May 2004; pp. 305 – 314
6. Dae-Young Choi ; "Enhancing the power of Web search engines by means of fuzzy query" Decision Support Systems, Volume 35, Issue 1, April 2003, pp. 31-44
7. Wen-Xue Tao; Wan-Li Zuo;" Query-sensitive self-adaptable web page ranking algorithm**"** International Conference on Machine Learning and Cybernetics, Vol. 1, 2-5 Nov. 2003; pp. 413 - 418
8. Mukhopadhyay, Debajyoti; Giri, Debasis; Singh, Sanasam Ranbir; "An Approach to Confidence Based Page Ranking for User Oriented Web Search;" SIGMOD Record, Vol.32, No.2, June 2003; pp. 28-33