Paper from Proceeding of the National Conference "Science in Media 2012" Organized by YMCA University of Science and Technology, Faridabad, Haryana (India) December 3rd -4th 2012 Published by
IJMRS's International Journal of Engineering Sciences, ISSN (Online): 2277-9698
www.ijmrs.com

Document structure similarity methods: a review

Deepika and Ashutosh Dixit, YMCA University of Science and Technology, Faridabad,

***Abstract*—The primary goal of Search Engines is to provide user information relevant to its query. For this purpose a web crawler is used which is a part of search engine and responsible for fetching data. The crawler traverses the web and provides pages to the search engines. Generally crawling is based on content but it is observed that structure of a page plays an important role in getting more relevant data .This paper reviews some methods given by various researchers in which crawling is based on structure of a page rather than content.**

Index Ter*ms*—C Search Engines, Web Crawler, Document Structure

I. INTRODUCTION

In everyday life we take help of search engines (such as google, altavista, yahoo etc.) to search for information on the World Wide Web. Search engines try to maintain real-time information by running an algorithm on a web crawler.

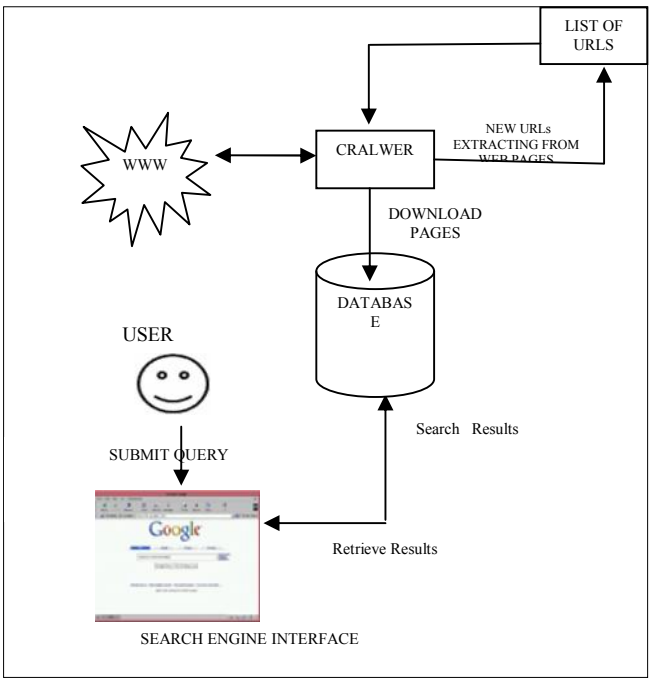

Fig. 1. Architecture of a Search Engine

A Web Crawler is a program that visits Web sites and reads their pages and other information in order to create entries for a search engine index. The architecture of a general search engine is shown above in figure 1.

On everyday basis large number of documents gets added on the Web. This has made document extraction process cumbersome. Focus of most of the research work related to document extraction is on content of the document. But with the increase in number of documents on the web, identifying documents on the basis of their structure seems to be more meaningful. Structural information helps in grouping of large number of pages from various websites. Algorithms based on similarity of structure for searching information compute the minimum cost edit distance between any two document structures. However, as these algorithms are expensive, typically requiring O(n2) or more time to compute the distance, there are opportunities to create algorithms that are faster, but provide slightly less accuracy in computing the distance. In this paper we present an overview of the current approximation algorithms used to detect structural similarity between different web documents.

II. DOCUMENT SIMILARITY METHODS

Overview of the different types of algorithms that have been used to determine document similarity has given below. The different metrics described are tree edit distance similarity, tag similarity, Fourier transforms, and path similarity.

A. Navigation pattern

Vidal et al.[1] proposed that by knowing the structure of a page beforehand which is a sample page of users' query, relevant page(s) can be searched more efficiently. So, like previous crawling methods in which all pages related to search topic has been fetched, now fetch only those pages which will have similar structure as that of sample page in terms of relevancy.

For the desired purpose a tool is developed by Vidal et al.[1] for generating structure-driven crawlers that requires only little efforts from users, since it relies on a sample page of the pages to be fetched. To accomplish this, given a sample page and an entry point to a Web site, the tool greedily traverses the Web site looking for target pages, i.e., pages that are structurally similar to the sample page. Next, it records all paths that lead to target pages and generates a navigation pattern which is composed by sequences of patterns of links a crawler has to follow to reach the target page. Finally, the tool generates a crawler based on these patterns. From this point on, the crawler can be used to fetch pages that are structurally

similar to the sample page, even if new similar pages are added later.

B. X-PATH

It is considered that web pages are in the form of HTML pages. **Wang et al.[2]** represent web Documents as document Object Model (DOM). But before representing HTML pages into Dom tree, first convert them into XHTML pages. Let us assume an example, suppose A & B are two Dom trees corresponding to two web pages. A formula is derived that calculate similarity between documents. The formula follows as:

$$Similarity\ (A, B) = \frac{SimpleTreeMatching\ (A, B)}{(sizes(A) + sizes(B))/2}$$

Where,

SimpleTreeMatching (A, B) → the number of maximum matching nodes of tree A & tree B;
sizes(A) & sizes(B) → the number of nodes on tree A and tree B.

When *Similarity (A,B)* is closer to 1, tree A and tree B are very similar to each other, and the HTML documents they represent are also very similar. For a given specific threshold $\theta(0 \le \theta \le 1)$, if the $Similarity(A,B) \ge \theta$, then the two trees are considered to be matched successfully, and the Web data will be extracted correspondingly; otherwise, the two trees does not match. In this paper, θ is set as 0.6.

C. DOM TREE

Chunying Kang [3] also decomposed these web pages in DOM tree. Then these DOM trees are then traversed in breadth first manner. On the basis of traversal, statistics of its changes, layer by layer DOM node tree comparison and then the sum of all floors of the changes are computed. Some threshold value is fixed on the basis of which it is decided that if their value is less than some threshold then pair of pages are structurally similar otherwise not.

D. BETWEEN XML DOCS

Nierman et al. [4] gives the idea to measure structural similarity between two XML documents. Tree edit distance based measures are used here. The algorithm developed by them is dynamically finds the distance between any pair of documents. A collection of documents are derived from multiple Document Type Descriptors (DTDs) are used here from which pair-wise distances between documents in the collection are computed and cluster the documents using these distances. It is observed that the resulting clusters match the original DTDs and has better results than previously used similarity methods.

E. FOCUSSED CRAWLERS' APPROACH

Ling Yu et al. [5] represents web page in the form of tree. Tree has nodes that denoted as tags such as form, body, table, title etc. Now similarity between two trees is computed as:
Structure × Structure _ [0 . . . 1], which returns the degree of similarity of a structure of the page operating to the structure of the page given. It is considered that if the situation is ideal then this function should have the property that if its value for x1 is greater than for x2 then we can conclude that the similarity of x1 to the page is higher than x2. The main idea behind this function is that usually similar structure in pages belonging to a specific domain.

III. CONCLUSION

With time various methods have been developed for searching relevant web pages. Broadly there are two kinds of methods used for searching pages-Content based and structure based. This paper highlights some of important methods based on similarity of structures. Various methods discussed in the paper shows that structure based methods give better results than content based methods.

REFERENCES

[1] M´arcio L.A. Vidal, Altigran S. da Silva, Edleno S. de Moura, Jo˜ao M. B. Cavalcanti. GOGETIT!: a Tool For Generating Structure-Driven Web Crawler. Published in the proceedings of the 15th international conference on World Wide Web (WWW'06). Pages 1011-1012. ACM New York, USA.

[2] Hua Wang, Y. Z. (2010). Web Data Extraction Based on Simple Tree Matching. Paper presented at the Information Engineering (ICIE), 2010 WASE International Conference

[3] Kang, C. (2009). DOM-based Web Pages to Determine the Structure of the Similarity Algorithm. Paper presented at the Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium.

[4] Nierman and H. V. Jagadish. Evaluating structural similarity in XML documents. In Proceedings of the 5th International Workshop on the Web and Databases (WebDB2002), Madison, Wisconsin, USA, June 2002.

[5] Huo Ling Yu Sch. of Inf., Beijing Wuzi Univ., Beijing, China Liu Bingwu ; Yan Fang (2010). Similarity Computation of Web Pages of Focused Crawler. This paper appears in: Information Technology and Applications (IFITA), 2010 International Forum