A Proficient Method for Identifying Absent Repeated Links on Web Communities Using LLI Algorithm

Revathy Ravichandran, Gunavathi Ramasami
Computer Science, Sree Saraswathi Thyagaraja College
Bharathiyar University
revathyravi86@gmail.com; gunaganesh2001@gmail.com

Abstract: Web mining is broadly defined as the discovery and analysis of useful information from the World Wide Web. This describes the automatic search of information resources available online. Due to growth of web, the amount of hardware and network resources needed is large and expensive. In addition search engines are popular tools, so they have heavy constraints on query answer time. These algorithms could be used for various Web applications, such as enhancing Web search. The ideas and techniques in this work would be helpful to other Web-related researches. The community still does not know how many links are missing, where these links are and finally, whether the missing links will change our conceptual model of the Internet topology. An accurate and complete model of the topology would be important for protocol design, performance evaluation and analyses. The goal of this work is to develop methodologies and tools to identify and validate such missing links. Web based information extraction using mining, which can work its way through a website, downloading pages, link and connectivity that it finds so that they can be viewed locally in our own browser, without navigating the entire contents connected to the internet. The system describes the LLI (Latent Linkage Information) algorithm to identify the missing and repeated links between the websites and SVD (Single Value Decomposition) method to divide the websites into multiple blocks and identifies parent child relationship between websites. The conclusion deals with

- Finding the number of links of a particular web site.
- Discovers the missing link from web site.
- The specific links are found in particular web site.
- The parent and child links of a particular web site are found.

Keywords: LLI (Latent Linkage Information); SVD; PR (Page Rank); URI; URL.

1. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is a combination of powerful methods that helps in reducing the costs and risks as well as increasing revenues, by extracting strategically information from the available data. The overall goal of the data

Mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

One problem associated with retrieval of data from web documents is that they are not structured as in traditional databases. There is no schema or division into attributes. Traditionally, Web pages are defined using hypertext markup language (HTML). Web pages created using HTML are only semi structured, thus making querying a more difficult than with well-formed databases containing schemas and attributes with defined domains.HTML ultimately will be replaced by extensible markup language (XML), which will provide structured documents and facilitate easier mining.

Web Usage Mining

Web usage mining is the third category in web mining. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access.

Objectives of Research

The objective of the research is used for searching the information in the web page. It is used for identifying the deeper relationship among the web pages more effectively. Query processing time is faster.

The major goal of this research is used for finding the number of the links in a website. It is used for identifying the missing and repeated links in a website. It is also used for navigating the contents of the web pages more easily by identifying the parent child relationship with the web pages.

Hyperlink

Primarily, the hyperlink is one of the most obvious features of the web and can be easily extracted by parsing the web page codes. The hyperlink analysis has proven success in many web related areas, such as page ranking in the search engine Google web page community construction web search improvement web clustering and visualization and relevant page *finding.*

Page ranking in the search engine Google

PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element *E* is referred to as the *PageRank of E* and denoted by $PR(E)$.other factors like Author Rank can contribute to the importance of an entity.

A PageRank results from a mathematical algorithm based on the web graph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com or usa.gov. The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support.

Predictive Analytics

Predictive analytics is using business intelligence data for forecasting and modeling. It is a way to use predictive analysis data to predict future patterns. It is used widely in the insurance, medical and credit industries. Assessment of credit, and assignment of a credit score is probably the most widely known use of predictive analytics. Using events of the past, managers are able to estimate the likelihood of future events.

Hypertext and WWW Rankings

There have been several approaches to ranking pages in the context of hypertext and the www. In work predating the emergence of the www, Botafogo, Revlon, and Shneiderman [19] worked with focused, stand-alone hypertext environments. They denied the notions of index nodes and reference nodes an index node is one whose out degree is significantly larger than the average out-degree, and a reference node is one whose in-degree is significant larger than the average needed by the link structure.

2. SYSTEM METHODOLOGY

Web URL Identification

In computing, a Uniform Resource Locator (URL) is a type of Uniform Resource Identifier (URI) that specifies where an identified resource is available and the mechanism for retrieving it. In popular usage and in many technical documents and verbal discussions it is often, imprecisely and confusingly, used as a synonym for uniform resource identifier. The confusion in usage stems from historically different interpretations of the semantics of the terms involved. In popular language a URL is also referred to as a Web address.

HTML Parsing

The HTML Parsing module is a class for accessing HTML as tokens. An HTML Parsing object gives you one token at a time, much as a file handle gives you one line at a time from a file. The HTML can be tokenized from a file or string. The tokenizer decodes entities in attributes, but not entities in text. A program that extracts information by working with a stream of tokens doesn’t have to worry about the peculiarity of entity encoding, whitespace, quotes, and trying to work out where a tag ends.

Link extraction

Regular expressions are powerful, but they’re a painfully low-level way of dealing with HTML. The system processes the spaces and new lines, single and double quotes, HTML comments, and a lot more. The next step up from a regular expression is an HTML tokenizer. In this module, we’ll use HTML Parser to extract information from HTML files. Using these techniques, you can extract information from any HTML file, and never again have to worry about character-level trivia of HTML markup. And automatic passage extraction methods from the body may be worthwhile. Implications of the findings for aids to summarization, and specifically the text.

Missing Link Extraction

Extract Link is a powerful, highly accurate, fast threaded link extractor utility to search and extract link (http, ftp, email, news, images) from any type of file (Html,). If the contents are not present in results in link, base, domain separately and supports link compare, URL extraction depth, false link/base removal, domain check list, filters, helps in identifying the missing links.

Repeated Link Identification

The repeated pages form the home page are analyzed. The user can identify the repeated pages and can find the efficiency of websites. By detecting the repeated pages, the back tracking of each link is identified as like missing page detection. That reduces the information search time for the user while extracting the website specific information.

LLI (Latent Linkage Information algorithm)

The LLI algorithm is efficient than the Extended Cocitation algorithm in semantically finding the relevant page. The topologic relationships among the pages in a page source can be easily expressed as a linkage matrix. This matrix makes it possible, by matrix operations, to reveal the deeper relationships among the pages and effectively find relevant pages.

The LLI reveals deeper relationship among the matrix elements and findsthe relevant links of a particular website

A = (aij) m×n

$$Aij = \begin{cases} 1 & \text{when page I is a child of page} \\ 0 & \text{otherwise.} \end{cases}$$

J, pagei€BS, pagej€Pu

B= (bij) p×q

$$bij = \begin{cases} 1 & \text{when page I is a child of page} \\ 0 & \text{otherwise.} \end{cases}$$

J, pagei€FS, pagej€cu

These two matrices imply more beneath their simple definitions. The linkage matrix between the pages in BS and the pages in P*u*is A= $(a_{ij})_{mxn}$ and the linkage matrix between pages in FS and pages in Cu is B= $(b_{ij})_{pxq}$.The ithrow of matrix A can be viewed as the coordinate vector of page i (page i Є BS) in an n-dimensional space spanned by the n pages in P*u* and the ith row of matrix B can be Viewed as the coordinate vector of page i (page i Є FS) in a q-dimensional space spanned by the q pages in C*u*. Similarly, the jthcolumn of matrix A can be viewed as the coordinate vector of page j (page j Є P*u*) in an m-dimensional space spanned by them pages in BS.

The matrix contains main linkage information among the pages and makes it possible to filter those irrelevant pages, which usually have fewer links to the parents of given u, and effectively find relevant pages. In this algorithm, the relevance of a page to the given page u is measured by the similarity between them.

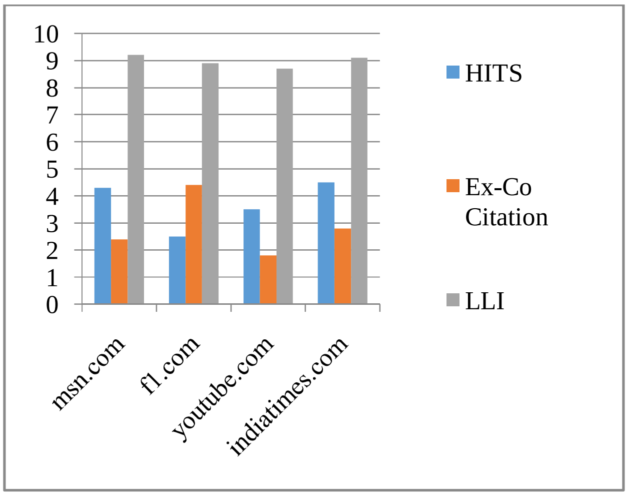
Page similarities in the LLI algorithm are measured by the deeper (mathematical) relationships among the pages that are revealed within the whole of the concerned page source by mathematical operations, especially the SVD of a matrix, not measured by simply counting the number of links.

Since the number of pages in the page source can be controlled by the algorithm and this number is relatively very small compared with the number of pages on the web, the LLI algorithm is feasible for application.

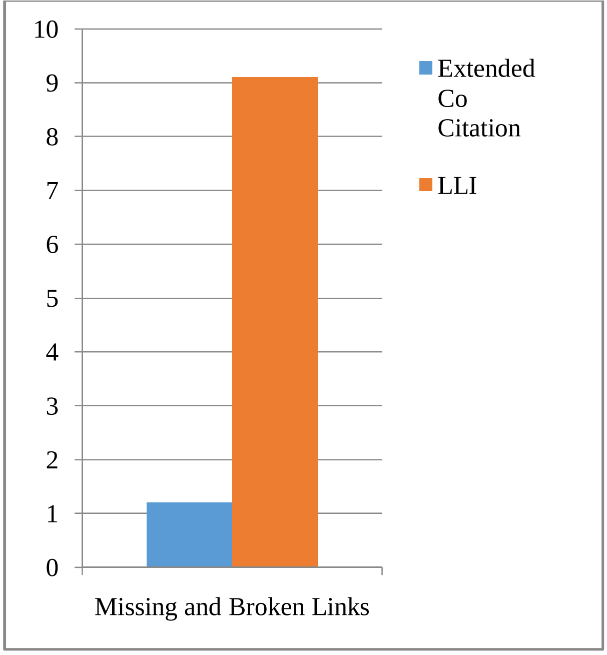
LLI algorithm reduces the influence of the pages in the same website to a reasonable level in the page similarity measurement, avoids some useful information being omitted, and prevents the results from being distorted by malicious hyperlinks. The page similarity in the LLI algorithm could also be adapted for page clustering .Thus this chapter deals with the comparative study of various hyper-link based algorithms and finds out the best algorithm to be used by this project.

3. COMPARISON AND RESULTS

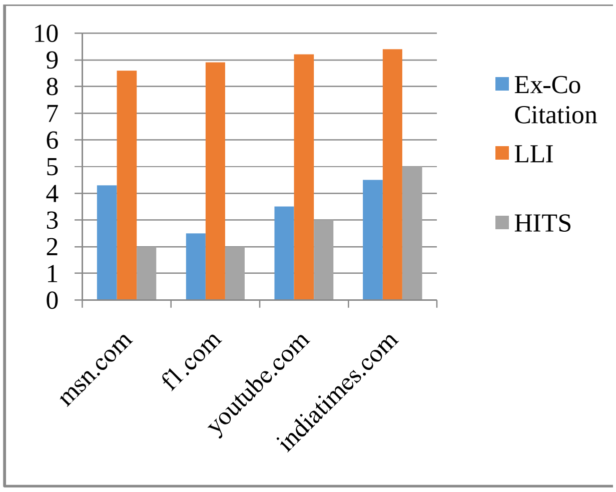
Finding No of Links in a website



Missing link Discovery



Parent Child Links in a Web



4. CONCLUSION

The proposed hyper link based algorithm can perform better than the existing systems. The LLI algorithm reveals deeper (mathematical) relationships among the pages and finds out relevant pages more precisely and effectively. LLI algorithm is based on hyperlink analysis among the pages and takes a new approach to construct the page source. The new page source reduces the influence of pages in the mirror site to a reasonable level in the page similarity measurement, avoids some useful information being omitted, and prevents the results from being distorted by malicious hyperlinks. Algorithms could identify the pages that are relevant to the given page in a broad sense, as well as those pages that semantically relevant to the given page. Furthermore, the LLI algorithm reveals deeper (mathematical) relationships among the pages and finds out relevant pages more precisely and effectively. Experimental results show the advantages of the algorithm.

REFERENCE

[1] A. Y. Levy and D. S. Weld. Intelligent internet systems. Artificial Intelligence, 118(1-2), 2000.

[2] C. Chen and L. Carr, “Trailblazing the Literature of Hypertext: Author Co-Citation Analysis (1989-1998),” Proc. 10th ACM Conf.Hypertext and Hypermedia, pp. 51-60, 1999.

[3] D. Mladenic. Text-learning and related intelligent agents. IEEE Intelligent Systems, 14(4):44–54, 1999.

[4] H. Kautz, B. Selman, and M. Shah. The hidden web.AI magazine, 18(2):27–36, 1997.

[5] J. Borges and M. Levene. Data mining of user navigation patterns. In Proceedings of the WEBKDD’99 Workshop on Web Usage Analysis and User Profiling, pages 31–36, 1999.