INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATION & MANAGEMENT

IJRCM

A Monthly Double-Blind Peer Reviewed (Refereed/Juried) Open Access International e-Journal - Included in the International Serial Directories

Indexed & Listed at:

Ulrich's Periodicals Directory ©, ProQuest, U.S.A., EBSCO Publishing, U.S.A., Cabell's Directories of Publishing Opportunities, U.S.A.,
Open J-Gage, India [link of the same is duly available at Inflibnet of University Grants Commission (U.G.C.)],
Index Copernicus Publishers Panel, Polandwith IC Value of 5.09 &number of libraries all around the world.
Circulated all over the world & Google has verified that scholars of more than 2501 Cities in 159 countries/territories are visiting our journal on regular basis.

Ground Floor, Building No. 1041-C-1, Devi Bhawan Bazar, JAGADHRI – 135 003, Yamunanagar, Haryana, INDIA

http://ijrcm.org.in/

CONTENTS

ASSOCIATE EDITORS

TECHNICAL ADVISOR

FINANCIAL ADVISORS

LEGAL ADVISORS

SUPERINTENDENT

CALL FOR MANUSCRIPTS

We invite unpublished novel, original, empirical and high quality research work pertaining to recent developments & practices in the areas of Computer Science & Applications; Commerce; Business; Finance; Marketing; Human Resource Management; General Management; Banking; Economics; Tourism Administration & Management; Education; Law; Library & Information Science; Defence & Strategic Studies; Electronic Science; Corporate Governance; Industrial Relations; and emerging paradigms in allied subjects like Accounting; Accounting Information Systems; Accounting Theory & Practice; Auditing; Behavioral Accounting; Behavioral Economics; Corporate Finance; Cost Accounting; Econometrics; Economic Development; Economic History; Financial Institutions & Markets; Financial Services; Fiscal Policy; Government & Non Profit Accounting; Industrial Organization; International Economics & Trade; International Finance; Macro Economics; Micro Economics; Rural Economics; Co-operation; Demography: Development Planning; Development Studies; Econometrics; Applied Economics; Development Economics; Business Economics; Monetary Policy; Public Policy Economics; Real Estate; Regional Economics; Political Science; Continuing Education; Labour Welfare; Philosophy; Psychology; Sociology; Tax Accounting; Advertising & Promotion Management; Management Information Systems (MIS); Business Law; Public Responsibility & Ethics; Communication; Direct Marketing; E-Commerce; Global Business; Health Care Administration; Labour Relations & Human Resource Management; Marketing Research; Marketing Theory & Applications; Non-Profit Organizations; Office Administration/Management; Operations Research/Statistics; Organizational Behavior & Theory; Organizational Development; Production/Operations; International Relations; Human Rights & Duties; Public Administration; Population Studies; Purchasing/Materials Management; Retailing; Sales/Selling; Services; Small Business Entrepreneurship; Strategic Management Policy; Technology/Innovation; Tourism & Hospitality; Transportation Distribution; Algorithms; Artificial Intelligence; Compilers & Translation; Computer Aided Design (CAD); Computer Aided Manufacturing; Computer Graphics; Computer Organization & Architecture; Database Structures & Systems; Discrete Structures; Internet; Management Information Systems; Modeling & Simulation; Neural Systems/Neural Networks; Numerical Analysis/Scientific Computing; Object Oriented Programming; Operating Systems; Programming Languages; Robotics; Symbolic & Formal Logic; Web Design and emerging paradigms in allied subjects.

Anybody can submit the **soft copy** of unpublished novel; original; empirical and high quality **research work/manuscript** ***anytime*** in ***M.S. Word format*** after preparing the same as per our **GUIDELINES FOR SUBMISSION**; at our email address i.e. infoijrcm@gmail.com or online by clicking the link **online submission** as given on our website (***FOR ONLINE SUBMISSION, CLICK HERE***).

GUIDELINES FOR SUBMISSION OF MANUSCRIPT

1. **COVERING LETTER FOR SUBMISSION:**

DATED: _____________

THE EDITOR
IJRCM

Subject: **SUBMISSION OF MANUSCRIPT IN THE AREA OF** .

(e.g. Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/Engineering/Mathematics/other, please specify)

DEAR SIR/MADAM

Please find my submission of manuscript entitled '____' for possible publication in your journals.

I hereby affirm that the contents of this manuscript are original. Furthermore, it has neither been published elsewhere in any language fully or partly, nor is it under review for publication elsewhere.

I affirm that all the author (s) have seen and agreed to the submitted version of the manuscript and their inclusion of name (s) as co-author (s).

Also, if my/our manuscript is accepted, I/We agree to comply with the formalities as given on the website of the journal & you are free to publish our contribution in any of your journals.

NAME OF CORRESPONDING AUTHOR:
Designation:
Affiliation with full address, contact numbers & Pin Code:
Residential address with Pin Code:
Mobile Number (s):
Landline Number (s):
E-mail Address:
Alternate E-mail Address:

NOTES:

a) The whole manuscript is required to be in ***ONE MS WORD FILE*** only (pdf. version is liable to be rejected without any consideration), which will start from the covering letter, inside the manuscript.
b) The sender is required to mentionthe following in the **SUBJECT COLUMN** of the mail:
New Manuscript for Review in the area of (Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/Engineering/Mathematics/other, please specify)
c) There is no need to give any text in the body of mail, except the cases where the author wishes to give any specific message w.r.t. to the manuscript.
d) The total size of the file containing the manuscript is required to be below **500 KB**.
e) Abstract alone will not be considered for review, and the author is required to submit the complete manuscript in the first instance.
f) The journal gives acknowledgement w.r.t. the receipt of every email and in case of non-receipt of acknowledgment from the journal, w.r.t. the submission of manuscript, within two days of submission, the corresponding author is required to demand for the same by sending separate mail to the journal.

2. **MANUSCRIPT TITLE**: The title of the paper should be in a 12 point Calibri Font. It should be bold typed, centered and fully capitalised.

3. **AUTHOR NAME (S) & AFFILIATIONS**: The author (s) **full name**, **designation**, **affiliation** (s), **address**, **mobile/landline numbers**, and **email/alternate email address** should be in italic & 11-point Calibri Font. It must be centered underneath the title.

4. **ABSTRACT**: Abstract should be in fully italicized text, not exceeding 250 words. The abstract must be informative and explain the background, aims, methods, results & conclusion in a single para. Abbreviations must be mentioned in full.

5. **KEYWORDS**: Abstract must be followed by a list of keywords, subject to the maximum of five. These should be arranged in alphabetic order separated by commas and full stops at the end.

6. **MANUSCRIPT**: Manuscript must be in ***BRITISH ENGLISH*** prepared on a standard A4 size ***PORTRAIT SETTING PAPER***. It must be prepared on a single space and single column with 1" margin set for top, bottom, left and right. It should be typed in 8 point Calibri Font with page numbers at the bottom and centre of every page. It should be free from grammatical, spelling and punctuation errors and must be thoroughly edited.

7. **HEADINGS**: All the headings should be in a 10 point Calibri Font. These must be bold-faced, aligned left and fully capitalised. Leave a blank line before each heading.

8. **SUB-HEADINGS**: All the sub-headings should be in a 8 point Calibri Font. These must be bold-faced, aligned left and fully capitalised.

9. **MAIN TEXT**: The main text should follow the following sequence:

   INTRODUCTION

   REVIEW OF LITERATURE

   NEED/IMPORTANCE OF THE STUDY

   STATEMENT OF THE PROBLEM

   OBJECTIVES

   HYPOTHESES

   RESEARCH METHODOLOGY

   RESULTS & DISCUSSION

   FINDINGS

   RECOMMENDATIONS/SUGGESTIONS

   CONCLUSIONS

   SCOPE FOR FURTHER RESEARCH

   ACKNOWLEDGMENTS

   REFERENCES

   APPENDIX/ANNEXURE

   It should be in a 8 point Calibri Font, single spaced and justified. The manuscript should preferably not exceed ***5000 WORDS***.

10. **FIGURES &TABLES**: These should be simple, crystal clear, centered, separately numbered &self explained, and **titles must be above the table/figure**. **Sources of data should be mentioned below the table/figure**. It should be ensured that the tables/figures are referred to from the main text.

11. **EQUATIONS**:These should be consecutively numbered in parentheses, horizontally centered with equation number placed at the right.

12. **REFERENCES**: The list of all references should be alphabetically arranged. The author (s) should mention only the actually utilised references in the preparation of manuscript and they are supposed to follow **Harvard Style of Referencing**. The author (s) are supposed to follow the references as per the following:

- All works cited in the text (including sources for tables and figures) should be listed alphabetically.
- Use (**ed.**) for one editor, and (**ed.s)** for multiple editors.
- When listing two or more works by one author, use --- (20xx), such as after Kohl (1997), use --- (2001), etc, in chronologically ascending order.
- Indicate (opening and closing) page numbers for articles in journals and for chapters in books.
- The title of books and journals should be in italics. Double quotation marks are used for titles of journal articles, book chapters, dissertations, reports, working papers, unpublished material, etc.
- For titles in a language other than English, provide an English translation in parentheses.
- The location of endnotes within the text should be indicated by superscript numbers.

PLEASE USE THE FOLLOWING FOR STYLE AND PUNCTUATION IN REFERENCES:

BOOKS

- Bowersox, Donald J., Closs, David J., (1996), "Logistical Management." Tata McGraw, Hill, New Delhi.
- Hunker, H.L. and A.J. Wright (1963), "Factors of Industrial Location in Ohio" Ohio State University, Nigeria.

CONTRIBUTIONS TO BOOKS

- Sharma T., Kwatra, G. (2008) Effectiveness of Social Advertising: A Study of Selected Campaigns, Corporate Social Responsibility, Edited by David Crowther & Nicholas Capaldi, Ashgate Research Companion to Corporate Social Responsibility, Chapter 15, pp 287-303.

JOURNAL AND OTHER ARTICLES

- Schemenner, R.W., Huber, J.C. and Cook, R.L. (1987), "Geographic Differences and the Location of New Manufacturing Facilities," Journal of Urban Economics, Vol. 21, No. 1, pp. 83-104.

CONFERENCE PAPERS

- Garg, Sambhav (2011): "Business Ethics" Paper presented at the Annual International Conference for the All India Management Association, New Delhi, India, 19–22 June.

UNPUBLISHED DISSERTATIONS AND THESES

- Kumar S. (2011): "Customer Value: A Comparative Study of Rural and Urban Customers," Thesis, KurukshetraUniversity, Kurukshetra.

ONLINE RESOURCES

- Always indicate the date that the source was accessed, as online resources are frequently updated or removed.

WEBSITES

- Garg, Bhavet (2011): Towards a New Natural Gas Policy, Political Weekly, Viewed on January 01, 2012 http://epw.in/user/viewabstract.jsp

AN APPROACH TOWARDS RELATIONAL WEB MINING WITH CORRESPONDENCE OF LINK BREAKDOWN STRUCTURE

SM SARAVANAKUMAR
ASST. PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE
PSG COLLEGE OF ARTS & SCIENCE
COIMBATORE

R SHANMUGAVADIVU
ASST. PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE
PSG COLLEGE OF ARTS & SCIENCE
COIMBATORE

ABSTRACT

The need to consolidate the information contained in heterogeneous data sources has been widely documented in recent years. In order to accomplish this goal, an organization must resolve several types of heterogeneity problems, especially the entity heterogeneity problem that arises when the same real-world entity type is represented using different identifiers in different data sources. Statistical record linkage techniques could be used for resolving this problem. However, the use of such techniques for online record linkage could pose a tremendous communication bottleneck in a distributed environment where entity heterogeneity problems are often encountered. In order to resolve this issue, we develop a matching tree, similar to a decision tree, and use it to propose techniques that reduce the communication overhead significantly, while providing matching decisions that are guaranteed to be the same as those obtained using the conventional linkage technique. These techniques have been implemented, and experiments with real-world and synthetic databases show significant reduction in communication overhead. This work introduces a link analysis procedure for discovering relationships in relational web pages, generalizing both simple and multiple correspondence analysis. It is based on a random walk model through the live web pages having as many states as elements in the website.

KEYWORDS

Heterogenous data, Web Mining, Link Analysis.

1. INTRODUCTION

1.1 OVERVIEW OF WEB MINING

Web mining can be defined as the discovery and analysis of useful information from the World Wide Web. Web mining can be defined as the integration of the information gathered by traditional data mining methods and techniques with information related to the web. In a simplified way we could say that its data mining adapted to the particularities of the web.

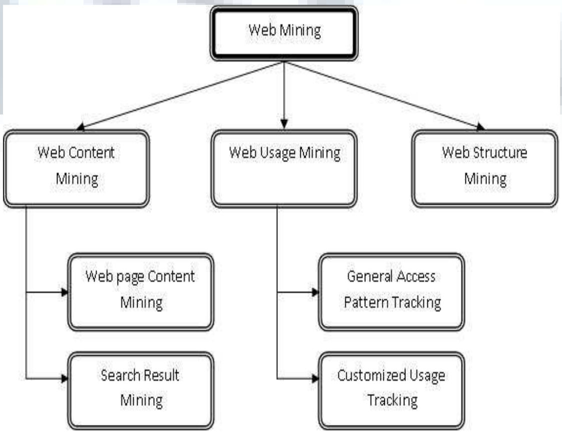
With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and resources. In addition, with the transformation of the Web into the primary tool for electronic commerce, it is imperative for organizations and companies, who have invested millions in Internet and Intranet technologies, to track and analyze user access patterns. These factors give rise to the necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge both across the Internet and in particular Web localities.

There are several important issues, unique to the Web paradigm, that come into play if sophisticated types of analyses are to be done on server side data collections. These include the necessity of integrating various data sources such as server access logs, referrer logs, user registration or profile information; resolving difficulties in the identification of users due to missing unique key attributes in collected data; and the importance of identifying user sessions or transactions from usage data, site topologies, and models of user behavior. We devote the main part of this paper to the discussion of issues and problems that characterize Web usage mining. Furthermore, we survey some of the emerging tools and techniques, and identify several future research directions.

1.2 WEB MINING TECHNIQUES

Web Mining when looked upon in data mining terms, can be said to have three operations namely Clustering, Associations and Sequential Analysis. Web Mining techniques provide a set of techniques that can be used to solve the problems of a user. By providing direct solutions to the problems, WM techniques can be used as a part of the bigger application that addresses many issues. WM techniques are not only the tools to handle the issues but the implementation methods in other related techniques from various research areas like Information Retrieval (IR), Database (DB) and Natural Language Processing (NLP) can also be developed.

FIG. 1: TASKS OF WEB MINING



Data Mining Techniques can be classified into three areas as indicated below based on which part of the web is to be mined. They are classified as follows,

- Web Content Mining

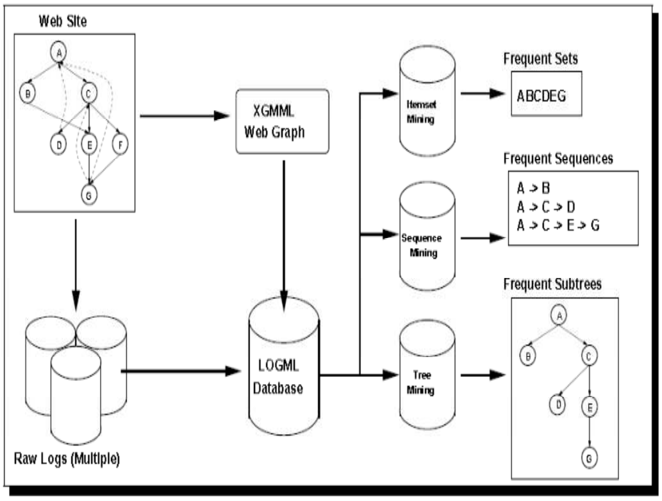
- Web Structure Mining
- Web Usage Mining

1.2.1 WEB USAGE MINING

Web usage Mining is a part of web mining which in turn it is a subset of data mining. Web usage mining involves mining the usage characteristics of the users of web applications. The extracted information can then be used in several ways such as enhancement of the existing application, investigating the fraudulent elements in the web.

As the part of the business intelligence in an organization rather than the technical element, it is used for deciding various strategies in the business through efficient use of web applications. The major problem in web mining in general and web usage mining in particular is the nature of the data they dealt with. In recent days the web data has become huge in nature and a lot of transactions and web usages are taking place by day to day aspects of life.

FIG. 2: WEB USAGE MINING ARCHITECTURE



Web usage mining process

The main processes in web usage mining are:

Preprocessing: Commonly used as a preliminary data mining practice, it transforms the data to a simpler format that will be more effectively processed for the purpose of the user. The different types of preprocessing in web usage mining are listed below:

1) Content Preprocessing
2) Usage Preprocessing
3) Structured Preprocessing

Pattern Discovery: WUM can be used to uncover patterns in server log list but it is often carried out only on samples of data. The mining process will be of ineffective if the sample data are not the good representation of the large data set. Listed below are the methods of pattern discovery:

1) Association rules
2) Statistical methods
3) Classification
4) Clustering methods
5) Sequential patterns
6) Dependency modeling

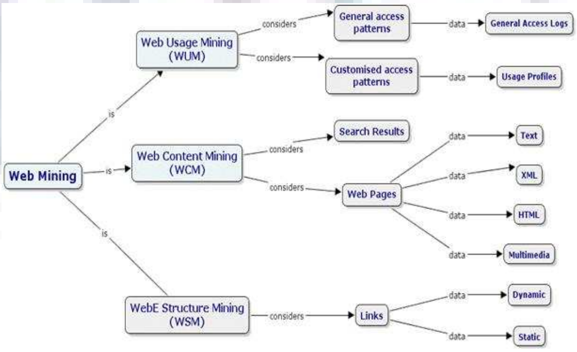
Analysis of the pattern: The obtained usage patterns are analyzed to filter uninterested information and extract the useful information. The OLAP (Online Analytical Processing) and SQL (Structured Query Language) can be used.

Areas of Web Usage Mining

1) System Improvement
2) Personalization
3) Site modification
4) Business Intelligence
5) Usage Characterization

CONCEPTUAL MAP OF WEB MINING

FIG. 3: CONCEPTUAL MAP OF WEB MINING



1.3 LINK ANALYSIS IN WEB INFORMATION RETRIEVAL

The goal of information retrieval is to find all documents relevant for a user query in a collection of documents. Decades of research in information retrieval were successful in developing and refining techniques that are solely word-based. With the advent of the web new sources of information became available, one of them being the hyperlinks between documents and records of user behavior. To be precise, hypertexts (i.e., collections of documents connected by hyperlinks) have existed and have been studied for a long time. What was new was the large number of hyperlinks created by independent individuals.

Hyperlinks provide a valuable source of information for web information retrieval as we will show in this article. This area of information retrieval is commonly called link analysis. A hyperlink is a reference of a web page B that is contained in a web page A. When the hyperlink is clicked on in a web browser, the browser displays page B. This functionality alone is not helpful for web information retrieval. However, the way hyperlinks are typically used by authors of web pages can give them valuable information content. Typically, authors create links because they think they will be useful for the readers of the pages. Thus, links are usually either navigational aids that, for example, bring the reader back to the homepage of the site, or links that point to pages whose content augments the content of the current page.

1.4 WEB PAGE TREE STRUCTURE

The site's blueprint is called a sitemap. The more information the page have, the more difficult architecting the site might be. There is not really a right way or a wrong way about architecting the site. A tree structure for a site could be the following:

TABLE 1

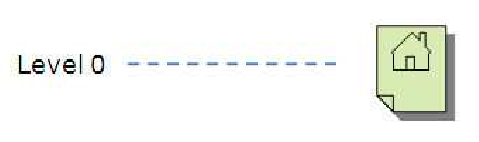
Level 0	Level 1	Level 2
Home Page	Our Company	Company Facts
		Methodology and Techniques
	Services	Construction
		Consulting
		Innovative Solutions
	Experience	Private Sector
		Public Sector
		Customer testimonials
		Case Studies
	Real Estate	Buildings
		Land
	Contact Us	

An important thing the user must also do regarding the structure of the site, which is also related to its usability as the user will see in the next section, is to think of what information will be presented on the home page. Usually, since the home page is the "gateway to the site", the user should place information the user think is important and internal pages (links to them) which the user wants to promote.

SINGLE PAGE

This is the where the entire site content is presented on the home page.

FIG. 4: SINGLE PAGE

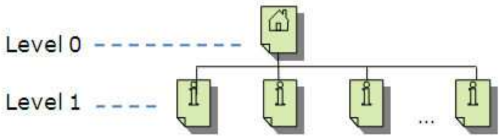


This is applicable only if the site does not have a lot of information to present. An advantage of this model is that it is very simple, requires minimum and the visitor gets the entire information on the spot without having to further navigate in the site. It is especially applicable for a marketing site for a specific product or service.

INDEX LIKE MODEL

In this model, all pages of the website hang below the home page (the tree structure goes one level deep) and all navigation happens through the home page.

FIG. 5: INDEX MODEL

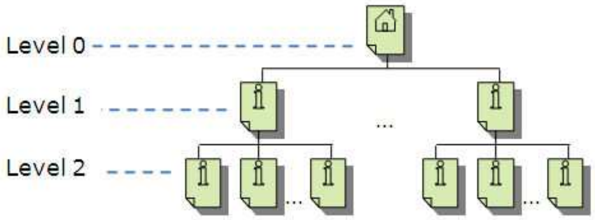


This model is good for a site which has decent amount of information to present and its thematic categories are independent with each other as well as self explanatory so that they don't need to be broken down further.

STRICT HIERARCHY

This model corresponds to an ideal situation where the user can partition the information following a "proper" hierarchical structure. In this structure the user can have as many levels as they want and each page can be accessed only through its parent.

FIG. 6: STRICT HIERARCHY

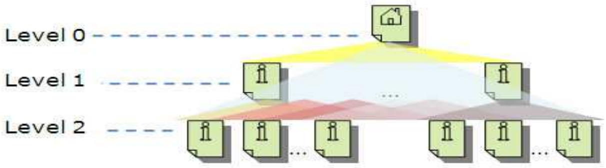


This model can be used for a site where information can be broken down into independent thematic areas, something that in real life is very difficult to find.

MULTI-DIMENSIONAL HIERARCHY

This model is similar to the one of the strict hierarchy, can be as deep as needed, any page can be accessed by any page, thus making it to partition the information easier and more flexible.

FIG. 7: MULTI-DIMENSIONAL HIERARCHY



This is the most commonly used model and the model is recommended for a mid- to large- (in terms of information) sized site.

1.5 THE RELATIONAL DATABASE MODEL

A database can be understood as a collection of related files. How those files are related depends on the model used. Early models included the

- Hierarchical Model - where files are related in a parent/child manner, with each child file having at most one parent file
- Network model – where files are related as owners and members, similar to the network model except that each member file can have more than one owner.

The relational database model was a huge step forward, as it allowed files to be related by means of a common field. In order to relate any two files, they simply need to have a common field, which makes the model extremely flexible.

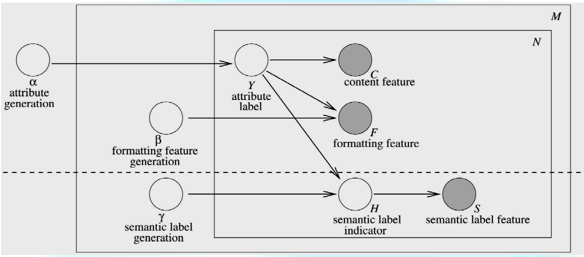
A database based on the relational model developed by E.F. Codd. A relational database allows the definition of data structures, storage and retrieval operations and integrity constraints. In such a database the data and relations between them are organised in tables. A table is a collection of records and each record in a table contains the same fields.

2. METHODOLOGY

2.1 A GENERATIVE MODEL

We develop a Bayesian learning framework to tackle the wrapper adaptation and new attribute discovery problem based on a generative model for the generation of the text fragments related to attributes which depicts the graphical representation of the model. Shaded and unshaded nodes represent observable and unobservable variables, respectively. With respect to the proposed framework using a running example, in each domain, there is an attribute generation variable denoted by α which controls the label Y of each text fragment. Consider the book domain and the Web site, as shown in Figure.

FIG. 8: A GENERATIVE MODELS



This site consists of several Web pages and each of them contains a number of records. The attributes contained in each record are title, author, price, publishing date, etc. For the Web sites, such as the ones shown in above figure, collected from the book catalog domain, records are normally composed of a similar set of book-related attributes. As a result, basically does not change drastically for different Web sites given a domain. Suppose a Web site contains M pages and the mth page contains N$_m$ text fragments. The label of each text fragment is generated according to P (Y| α). Based on the label generated, the content feature C is then generated from the distribution P (C|Y). In essence, C is a feature vector which characterizes the content of the text fragment. For example, an element of C refers to the number of occurrence of a particular token in the text fragment. C is dependent on Y; as a consequence, they are, in turn, dependent on α. Therefore, C remains largely unchanged for different Web sites. For instance, in the book catalog domain, the book titles collected from different Web sites show similar characteristics and orthographic information. Within a particular Web site, there is a formatting feature generation variable denoted by α. The formatting feature F of attributes represents the formatting information such as the font color or location of a text fragment. Similar to C, F is a feature vector characterizing the layout format of the text fragment. Attributes of records from different Web sites are normally presented in different formats or style. Within a Web site, an attribute of a record can be associated with a semantic label. A semantic label basically is a text fragment showing the semantic meaning of a text fragment. For example, in the Web page shown in, the semantic label for the attribute publication date is "Pub. Date:." The semantic labels commonly show certain regularity.

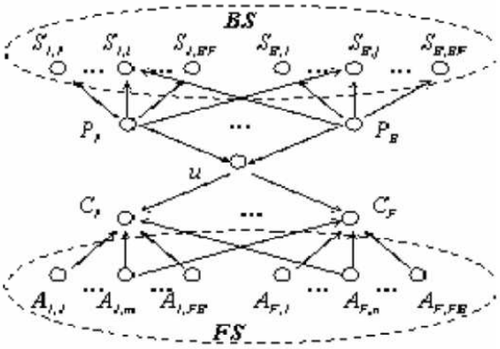
2.2 LATENT LINKAGE INFORMATION ALGORITHM

Yanchun Zhang, Guandang Xu proposed an algorithm named Latent Linkage Information (LLI) for ranking web pages that are closely related to a given page. This algorithm is primarily derived from scientific literature co-citation index algorithm. We give a brief description of this algorithm as follows.

The first step of this algorithm is to construct a web page space (page source) for the given page u from link topology on the web. The page source is constructed as a directed graph with edges indicating hyperlinks and nodes representing the following pages:

1. Page *u*,
2. Up to *B* parent pages of *u*, and up to *BF* child pages of each parent page that are different from *u*,
3. Up to *F* child pages of *u*, and up to *FB* parent pages of each child page that are different from *u*.

FIG. 9: PAGE SOURCE STRUCTURE FOR THE GIVEN PAGE *u*



2.3 WEB URL IDENTIFICATION

Web addresses are recorded in a Uniform Resource Locator (URL), a logical address of a web page that can always used to dynamically retrieve the current physical copy over the internet. The key advantage of the Uniform Resource Locator's (URL) is its universality, since the address is the same no matter where in the world it is used. This is why Tim Berners-Lee proposed in *RFC 1630,* Universal Resource Identifiers in WWW*,* that it be called a Universal Resource Identifier (URI) to suggest his vision of a network where anything could be linked to anything. However, he experienced philosophical resistance to this idea of universality from the IETF team working on the web standards, and so the address became named the now familiar Uniform Resource Locator.

2.4 HTML PARSING AND LINK EXTRACTION

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.

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. Regular expressions are a tool that is insufficiently sophisticated to understand the constructs employed by HTML. HTML is not a regular language and hence cannot be parsed by regular expressions. Regex queries are not equipped to break down HTML into its meaningful parts. so many times but it is not getting to me. Even enhanced irregular regular expressions as used by Perl are not up to the task of parsing HTML. You will never make me crack. HTML is a language of sufficient complexity that it cannot be parsed by regular expressions. It's considered good form to demand that regular expressions be considered verboten, totally off limits for processing HTML, but I think that's just as wrongheaded as demanding every trivial HTML processing task be handled by a full-blown parsing engine.

2.5 DOMAIN AND SUB DOMAIN CLASSIFICATION

The List of domains and sub domains of corresponding web server are identified with respect to the URL address entered. The Numbers of web servers connected to the main server are listed, also the business level integrated web servers of major web site is identified in the proposed module

2.6 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.

2.7 LINKAGE ANALYSIS CORRESPONDENCE

Efficient record linkage techniques based on the matching tree. The overall linkage process is summarizes. The first two stages in this process are performed offline, using the training data. Once the matching tree has been built, the online linkage is done as the final step. We can now characterize the different techniques that can be employed in the last step. Recall that, given a local enquiry record, the ultimate goal of any linkage technique is to identify and fetch all the records from the remote site that has a matching probability or more. Link analysis is a multivariate statistical technique. It is conceptually similar to principal component analysis, but applies to categorical rather than continuous data. In a similar manner to principal component analysis, it provides a means of displaying or summarizing a set of data in two-dimensional graphical form.

All data should be nonnegative and on the same scale for LAC to be applicable, and the method treats rows and columns equivalently. It is traditionally applied to contingency tables — LAC decomposes the chi-square statistic associated with this table into orthogonal factors. Because LAC is a descriptive technique, it can be applied to tables whether or not the chi-square statistic is appropriate. Several variants of LAC are available, including detrended correspondence analysis and canonical correspondence analysis. The extension of correspondence analysis to many categorical variables is called multiple correspondence analysis. An adaptation of correspondence analysis to the problem of discrimination based upon qualitative variables is called discriminant correspondence analysis or barycentric discriminant analysis.

2.8 SIMILARITY BASED WEB PAGE CLUSTERING

The World Wide Web creates many new challenges to information retrieval. The sheer mass and almost anarchic structure of the Web makes effective search difficult. Some good search engine alleviate the problem to some extent by ranking the search results based on the relevancy of the Web pages to user's query. They aim to place the most prominent pages at high ranks. Most of current search engines work by first retrieving a set of Web pages based on traditional text-based search engine and then applying link-based page ranking algorithms to rank this set of Web pages. Current page ranking algorithms have several problems. One of the most important problems is computation complexity since the convergence of those eigenvector-based ranking algorithms requires iteration which is computationally expensive. Full Similarity-based Ranking (FSBR) using densely connected clustering, a novel approach for Web page ranking, is proposed by Prof. Xinhua Zhuang. Under his advising, I did thorough literature overview, proposed a novel Subgraph Chaining Expansion algorithm, built test bed, implemented FSBR algorithm, and conducted simulation and extensive experiments. FSBR is a generic full similarity-based ranking scheme. It allows similarity measures built on link structure and other ranking contributable features. It finds similarity-based densely connected clusters and uses them in page ranking. The experimental results also show that FSBR provides much higher accuracy than the HITS page ranking algorithm.

Link-based clustering in the context of bibliometrics, hypertext and the WWW has focused largely on the problem of decomposing an explicitly represented collection of nodes into a "cohesive" subset. But it has been mainly applied to a moderately sized set of objects such as a focused collection of scientific journals, or the set of pages on a single website.

2.9 PLSA MODEL

Probabilistic Latent Semantic Analysis (PLSA) is derived from a variant of *latent semantic index* (LSI), which is commonly used in text mining and information retrieval. The PLSA model is based on a statistic model called aspect model, which can be utilized to identify the hidden semantic relationships among general co occurrence activities. Similarly, we can conceptually view the user sessions over web pages space as co occurrence activities in the context of web usage mining to discover the latent usage pattern Probabilistic latent semantic analysis (PLSA) is a statistical technique for the analysis of co-occurrence data. In contrast to standard latent semantic analysis, which stems from linear algebra and shrinks the size of occurrence tables (number of words occurring in some documents), PLSA is based on probability and statistics to derive a latent semantic model. Instead of the traditional key-word based data classification, PLSA tries to classify data to its "latent semantic". It's about learning "what was intended" rather than just "what actually has been said or written". After performing the PLSA classification, words which often come together in a same document will be seen as highly connected to each other, and the documents which contain these words therefore will be classified into the same "topic". The whole PLSA process can be divided into two parts, which are corpus classification and the query fold-in. Both parts use the expectation maximization (EM) theory. After running tens of iterations, we can get the final result. PLSA can be used in many areas, such as information retrieval or machine learning, to improve the original results.

3. CONCLUSION

We devised simple, yet powerful, and modular algorithms, to identify primary content blocks from Web pages. Our system outperformed the LH method significantly, in b-precision as well as runtime, without the use of any complex learning technique. The Feature Extractor, provided a feature, can identify the primary content block with respect to that feature. The Content Extractor detects redundant blocks based on the occurrence of the same block across multiple Web pages. The method , thereby, reduce the storage requirements, make indices smaller, and result in faster and more effective searches. Though the savings in file size and the precision and recall values from "Shingling method " is as good as from Content Extractor, Content Extractor outperforms the "Shingling Method " by a high margin in runtime. We intend to deploy our system as a part of a system that crawls Web pages, and extracts primary content blocks from it.

In the next step, we will look at the primary content and identify heuristic algorithms to identify the semantics of the content to generate markup. The storage requirement for indices, the efficiency of the markup methods, and the relevancy measures of documents with respect to keywords in queries should also improve (as we have shown briefly by caching size benefit) since now only the relevant parts of the documents are considered.

REFERENCES

1. Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman, “ Mining of Massive Datasets”, Stanford University, 2012.
2. Bayesian Approach”, ieee transactions on knowledge and data engineering, vol. 22, no. 4, April 2010.
3. Dr.N.P.Gopalan, Mrs. J. Akilandeswari P, Gabriel Sagaya Selvam, “Effectively Finding The Relevant Web Pages From The World Web Wide” .
4. Gary William Flake, Kostas Tsioutsiouliklis, and Leonid Zhukov, “Methods for Mining Web Communities: Bibliometric, Spectral, and Flow”.
5. Jingqian Xu, Prof. Xinhua Zhuang, “Full Similarity-Based Page Ranking”, University of Missouri, May 2008.
6. John R. Punin, Mukkai S. Krishnamoorthy, Mohammed J. Zaki, “Web Usage Mining - Languages and Algorithms”.
7. Keijo Ruohonen (Translation by Janne Tamminen, Kung-Chung Lee and Robert Piché), “GRAPH THEORY”, 2008.
8. MichaelWolverton, Ian Harrison, and David Martin, “Issues in Algorithm Characterization for Link Analysis”, SRI International.
9. Monika Henzinger, “Link Analysis in Web Information Retrieval”, Google Incorporated.
10. Nick Craswell and David Hawking, “Web Information Retrieval”.
11. Prasanna Desikan, Jaideep Srivastava, Vipin Kumar, and Pang-Ning Tan Department of Computer Science,” Hyperlink Analysis: Techniques and Applications” University of Minnesota, Minneapolis, MN, USA.
12. Rekha Jain, Dr. G. N. Purohit, “Page Ranking Algorithms for Web Mining” International Journal of Computer Applications (0975 – 8887),Volume 13– No.5, January 2011.
13. Robert Cooley, Bamshad Mobasher, Jaideep Srivastava , “Web Mining: Information and Pattern Discovery on the World Wide Web”, Department of Computer Science University of Minnesota.
14. Seung Yeol Yoo and Achim Hoffmann, “Clustering-Based Relevance Feedback for Web Pages”, University of New South Wales.
15. Tak-Lam Wong and Wai Lam, Senior Member, IEEE , “Learning to Adapt Web Information Extraction Knowledge and Discovering New Attributes via a
16. Tamanna Bhatia, “Link Analysis Algorithms For Web Mining”, IJCST Vol. 2, Issue 2, June 2011.
17. XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S, Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, “Top 10 algorithms in data mining”.
18. Yanchun Zhang, Guandong Xu, “On Web Communities Mining and Analysis”, School of Computer Science & mathematics,Victoria University, Vic 8001, Australia.
19. Yanchun Zhang, Jingyu Hou, “Effectively finding relevant Web pages from linkage information”, Knowledge and Data Engineering, IEEE Transactions, July-Aug. 2003, Volume: 15 , issue: 4.

REQUEST FOR FEEDBACK

Dear Readers

At the very outset, International Journal of Research in Computer Application and Management (IJRCM) acknowledges & appreciates your efforts in showing interest in our present issue under your kind perusal.

I would like to request you tosupply your critical comments and suggestions about the material published in this issue as well as on the journal as a whole, on our E-mail**infoijrcm@gmail.com** for further improvements in the interest of research.

If youhave any queries please feel free to contact us on our E-mail **infoijrcm@gmail.com**.

I am sure that your feedback and deliberations would make future issues better – a result of our joint effort.

Looking forward an appropriate consideration.

With sincere regards

Thanking you profoundly

Academically yours

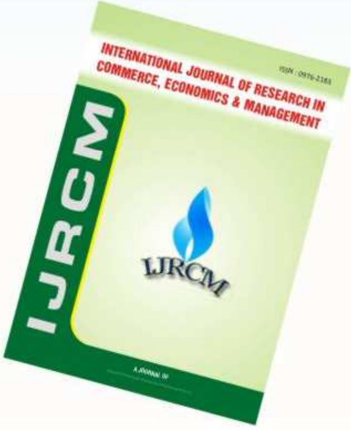
Sd/-

Co-ordinator

ABOUT THE JOURNAL

In this age of Commerce, Economics, Computer, I.T. & Management and cut throat competition, a group of intellectuals felt the need to have some platform, where young and budding managers and academicians could express their views and discuss the problems among their peers. This journal was conceived with this noble intention in view. This journal has been introduced to give an opportunity for expressing refined and innovative ideas in this field. It is our humble endeavour to provide a springboard to the upcoming specialists and give a chance to know about the latest in the sphere of research and knowledge. We have taken a small step and we hope that with the active co-operation of like-minded scholars, we shall be able to serve the society with our humble efforts.

Our Other Journals


ISSN: 0976-2183
INTERNATIONAL JOURNAL OF RESEARCH IN COMMERCE, ECONOMICS & MANAGEMENT
IJRCM
IJRCM

ISSN: 0976-2183
INTERNATIONAL JOURNAL OF RESEARCH IN COMMERCE, IT & MANAGEMENT
IJRCM
IJRCM