An Integrated Set of Web Mining Tools for Research

D.Aravind
M.Phil. Research Scholar, Department of Computer Science
Morappur Kongu College of Arts and Sciences

OPEN ACCESS

Volume : 6

Special Issue : 1

Month : September

Year: 2018

ISSN: 2321-788X

Impact Factor: 3.025

Citation:
Aravind, D. (2018). An Integrated Set of Web Mining Tools for Research. *Shanlax International Journal of Arts, Science and Humanities*, 6(S1), pp.87–97.

DOI:
https://doi.org/10.5281/zenodo.1410995

Abstract

The evolution of the World Wide Web has brought us enormous and ever-growing amounts of data and information. A large number of data provided by the web, it has become a special area for research. In a sense, the web data mining with its design and implementation provides well bound utilizing information from the web for research. This paper describes th e design and implementation of web data mining research. It gives a method of identifying, extracting, filtering and analyzing data for web resources. It sequentially starts with Information Retrieval (IR), Information Extraction (IE), Generalization, analysis, and Validation. IR will identify web sources by predefined categories with classification. IE will use a hybrid extraction way to select portions from a web page and give data into a database. Generalization will clean data and use database techniques to analyze collected data. Validation uses model-based data extraction to validate the data correctness. This work offers an integrated set of web mining tools that will help advance sophisticated in supporting researcher doing online research.

Keywords; Web data mining, Information Retrieval, Information Extraction.

Introduction

The evolution of theWorld Wide Web has brought us enormous and ever-growing amounts of data and information. It influences almost all aspects of people's lives. Also, with the abundant data provided by the web, it has become the main resource for research. Furthermore, the low cost of web data makes it more attractive to researchers. Researchers can retrieve web data by browsing and keyword searching [12]. However, there are several limitations to these techniques. It is hard for researchers to retrieve data by browsing because there are many following links contained in a web page. Keyword searching will return a big amount of irrelevant data. On the other hand, traditional data extraction and mining techniques cannot be applied directly to the web due to its semi-structured or even unstructured nature. Web pages are Hypertext documents, which contain both text and hyperlinks to other documents. Furthermore, other data sources also exist, such as mailing lists, newsgroups, forums, etc. Thus, design and implementation of a web mining research support system have become a challenge for people with interest in utilizing information from the web for their research.

Usually, web mining is classified into web content mining, web structure mining and web usage mining. Web content mining studies the search and retrieval of information on the web. Web structure mining focuses on the structure of the hyperlinks (inter-document structure) within a web. Web usage mining discovers and analyzes user access patterns [3].

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. Web usage mining provides the support for the website design, providing personalization server and other business making a decision, etc. To better serve for the users, web mining applies the data mining, the artificial intelligence and the charting technology and so on to the web data and traces users' visiting characteristics, and then extracts the users' using pattern[l]. It has quickly become one of the most important areas in Computer and Information Sciences because of its direct applications in e-commerce, CRM, Web analytics, information retrieval and filtering, and Web information systems.

According to the differences between the mining objects, there are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent; based technology may also fall in this category. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs.

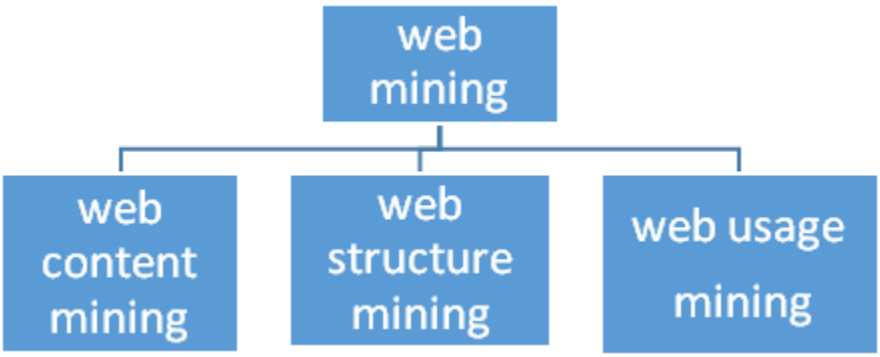


Figure 1: Taxonomy of Web Mining

Web Usage Mining

The Concept of web usage mining:

Discovery of meaningful patterns from data generated by client-server transactions on one or more Web servers.

Typical Sources of data:

1. Automatically generated data stored in server access logs, referrer logs, agent logs, and client-side cookies
2. E-commerce and product-oriented user events (e.g., shopping cart changes, ad or product click-troughs, etc.)
3. User profiles and user ratings
4. Meta-data, page attributes page content, site structure

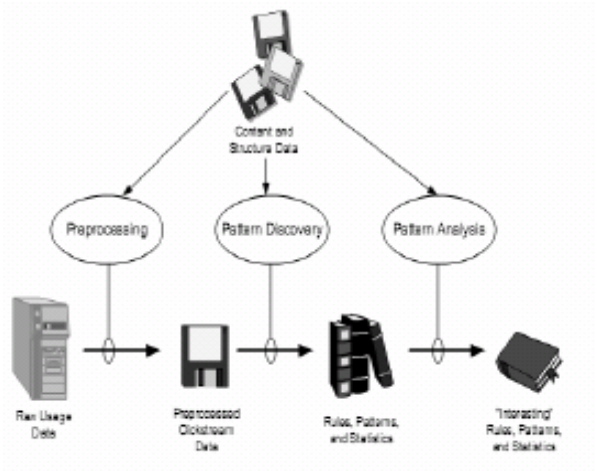


Figure 2: Web Usage Mining Process

Web Log Format

A web server log file contains requests made to the web server, recorded in chronological order. The most popular log file formats are the Common Log Format (CLF) and the extended CLF. A common Log Format file is created by the web server to keep track of the requests that occur on a website.

Debug

Access #1: Got some data! HTTP/1.1 200 OK Server: Zeus/4.2 Date: Mon, 12 Jul 2008 15:46:34 GMT Last-Modified: Thu, 10 July 2008 01:36:00 GMT Content-Type: text/plain Expires: Thu, 18 July 2008 01:36:00 GMT Content-Length: 705 Accept-Ranges: bytes Cache-Control: max-age=0: HTTP/1.1 304 Not Modified Server: Zeus/4.2 Date: Mon, 12 Jul 2008 15:46:34 GMT Expires: Thu, 18 July 2008 01:36:00 GMT Accept-Ranges: bytes Cache-Control: max-age=0

Example of typical server log

Web Information Retrieval

The web can be treated as a large data source, which contains many different data sources. The phase of web mining research is to identify web resources for a specific research topic. Providing an efficient and effective web information retrieval tool is important in such a system.

Data Sources

Because the World Wide Web is a valuable resource for research, it is important to know what kinds of data sources exist and what type of contents they contain. One of the most important tasks for researchers is to find the available on-line data sources. Commonly used data sources can be classified as online databases, documents, and archives.

With the increasing popularity of the web, many online databases have appeared to provide quick and easy access to research data. GenBank [13] created by the National Center for Biotechnology Information (NCBI) [14] is an online database to provide genetic sequence information. The Department of Anthropology at the California Academy of Sciences built an on-line database with 17,000 objects and over 8,000 images.

However, a large amount of on-line information is not in the form of on-line databases. In most cases, information is contained in web pages. Many websites maintain statistical documentation

on their websites. For example, SourceForge [15] maintains a statistical web page for each project hosted on this site, which has information such as the number of page views, downloads, bug reports, etc. Many stock trading websites provide price, volume and trend charts for stocks. Such statistical information is an essential data resource in research analysis.

Finally, some public archives are valuable resources for researchers because they include information about users' behaviors. Forums are an admired way of communication on the web. A website can maintain many different forums for special purposes, which are helpful in studying different activities of web users. Newsgroups are a communication method through which users can download and read the news. They provide information on user-to-user assistance. Other useful archives include mailing lists, software releases, etc.

Motivation

There are two most important categories of searching tools on the Web: directories (Yahoo, Netscape, etc.) and search engines (Lycos, Google, etc.). Both apparatus require an indexing system. The index may contain URLs, titles, headings, etc. Directories are also called indices or catalogs. They are subject lists created by specific indexing criteria. Directories consist of a list of topics which contain a list of web sites. Users can click on those websites to look up contents. With the increase in size, many sites employ search engines to assist query of directories. Search engines are commonly used to query and retrieve information from the web. Users perform a query through keywords when searching web content. Automated programs such as crawlers and robots are used to search the web. Such programs traverse the web to recursively retrieve all relevant documents. A search engine consists of three components: a crawler which visits web sites, indexing which is updated when a crawler finds a site and a ranking algorithm which records those relevant websites.

Current IR systems have several difficulties during web searching. Low precision

(i.e., too many irrelevant documents) And low recall (i.e., too little of the web is covered by well-categorized directories) are two main problems [16]. Furthermore, current directories and search engines are designed for general uses, not for research needs. For example, when we query with Open Source Software, both Yahoo and Google will return many irrelevant results. And those results are not well classified as what a research project needs, such as organizations, papers, websites, etc.

Users must manually browse those results to classify their interesting webpages. CiteSeer [17] is an automatic indexing research support system, but it only provides related paper information. There should be more information included for a research project.

Here, we propose to develop a web research search tool which combines directories and search engines to provide a classified interface to users. By using this search tool, users submit an uncertainty based on their research needs, and the search tool will return categorized directories related to their query. Unlike Yahoo, our returned directories will be built by automatic classification.

Web Information Extraction

Because web data are semi-structured or even unstructured, which cannot be manipulated by traditional database techniques, it is imperative to extract web data to port them into databases for further handling. The purpose of Web Information Extraction (IE) in our web mining research support system is to extract a specific portion of web documents useful for a research project. A specific web data set can be scattered among different web hosts and have different formats. IE takes web documents as input, identifies a core fragment, and transforms that fragment into a structured and unambiguous format [18]. This chapter describes the design of a web IE system. We first introduce the concept of wrappers.

Wrappers

Information extraction on the web brings us new challenges. The volume of web documents is enormous, documents change often, and the contents of documents are dynamic. Designing a general web IE system is a hard task. Most IE systems use a wrapper to extract in formation from a particular website. A wrapper consists of a set of extraction rules and specific code to apply rules to a particular site. A wrapper can be generated manually, semi-automatically or automatically.

The manual generation of a wrapper requires writing ad hoc codes based on the creator's understanding of the web documents. Programmers need to understand the structure of a web page and write codes to translate it. Techniques have been developed to use expressive grammars which describe the structure of a web page, and to generate extraction codes based on the speci_ed grammar. TSIMMIS [20] is an example of the manual wrapper. TSIMMIS focuses on developing tools to facilitate the integration of heterogeneous information sources. A simple self-describing (tagged) An object model is adapted to convert data objects to a common information model. Languages and tools are developed to support the wrapping process. The manual way of wrapper generation cannot adapt to the dynamic changes of websites. If new pages appear or the format of existing sources is changed, a wrapper must be modi_ed to adapt the new change.

Semi-automatic wrapper generation uses heuristic tools to support wrapper generation. In such a system, sample pages are provided by users to hypothesize the underlying structure of the whole website. Based on the hypothesized structure, wrappers are generated to extract the information from the site. This approach does not require programmer knowledge about web pages, but the demonstration of sample pages are required for each new site. Wrappers can be generated automatically by using machine learning and data mining techniques to learn extraction rules or patterns. These systems can train themselves to learn the structure of web pages. Learning algorithms must be developed to guide the training process.

Wrapper Generation Tools

Many apparatus has been created to generate wrappers. Such tools include Languages for Wrapper Development, NLP-based Tools, Modeling-based Tools, and Ontology-based Tools.

Some languages are specifically designed to assist users to address the problem of wrapper generation. Minerva combines the benefits of a declarative and grammar-based approach with the edibility of procedural programming by enriching regular grammars with an explicit exception-handling mechanism. Minerva dense the grammar in Extended Backus-Naur Form (EBNF) style which adds the regular expression syntax of regular languages to the BNF notation, to allow very compact specifications. It also provides an explicit procedural mechanism for handling exceptions inside the grammar parser. Web-QOL (Object Query Language) [3] is a declarative query language which aimed at performing SQL-like queries over the web. A generic HTML wrapper parses a web page and produces an abstract HTML syntax tree. Using the syntax of the language, users can write queries to locate data in the syntax tree and output data in some formats, i.e., tables. Such tools require users to examine web documents and write a program to Separate extraction data.

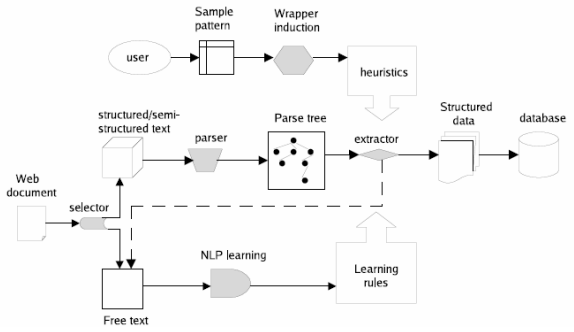
Natural language processing (NLP) techniques are used to learn extraction rules existing in natural language documents. These rules identify the relevant information in a document by using syntactic and semantic constraints. The wrapper induction tools also generate extraction rules from a given set of training samples. The difference between wrapper induction tools and those with NLP is that they format features to implicitly delineate the structure of data, while NLP relies on linguistic constraints. Such difference makes these tools more suitable for HTML documents. WIEN takes a set of sample web pages where interesting data are labeled and uses specific induction heuristics to generate a specific wrapper.

Before extracting data, the wrapper partitions an HTML string into tokens then uses a learning algorithm to induce extraction rules based on the context formed by the tokens. The resulting FST takes a sequence of tokens to match the context separators to determine state transitions. This approach can wrap a wide range of semistructured Web pages because FSTs can encode each

Different attribute permutation as a path. It can deal with hierarchical data extraction. There are two inputs: a set of training examples in the form of a sequence of tokens representing the environment of data to be extracted; and an Embedded Catalog Tree to describe the page structure. The rules generated try to cover as many as possible of the given examples. If there are uncovered examples, it generates a new disjunctive rule. When all examples are covered, STALKER returns a set of disjunctive extraction rules. Modeling-based tools locate portions of data in a web page according to a given target structure. The structure is provided based on a set of modeling primitives which conform to an underling data model. NoDoSE [1] is an interactive tool for semi-automatically determining the structure of such documents and extracting their data. Using a GUI, the user hierarchically decomposes the file, outlining its regions and then describing their semantics. In the decomposition process, users build a complex-structured object and crumble it into a simpler object. Such decomposition is a training process through which NoDoSE can learn to construct objects and identify other objects in the documents. This task is expedited by a mining component that attempts to infer the grammar of the file from the information the user has input so far. All tools presented above generate extraction rules based on the structure of the data in a document. Ontology-based tools rely directly on the data. Given a specific domain application, an ontology tool can locate constants in the page and construct objects with them. In, they present an approach to extract information from unstructured documents based on an application ontology that describes a domain of interest. By using the ontology, rules are formulated to extract constants and context keywords from unstructured documents on the web. For each unstructured document of interest, a recognizer is applied to organize extracted constants as attribute values of tuples in a generated database schema. This tool requires the construction of an ontology by experts and the ontology construction needs to be validated.

Hybrid Information Extraction

The manual wrapper is easy to develop and implement. However, users must create different wrappers for use with each change in web documents. Maintenance cost will be too high for this manual wrapper approach. Moreover, many text-like documents exist on the web, e.g. discussion board, news group. We can classify web contents as two types: structured/semi-structured text and free text. The first type has data items (e.g., names, SSN, etc.). Example web documents of this type include on-line statistics, tables, etc. The second type consists of free languages, e.g., advertisements, messages, etc. Techniques such as wrapper induction and modeling-based tools are suitable for pages of the first type because



The hybrid information extraction architecture

Such tools rely on delimiters of data to create extraction rules. NLP techniques are based on syntactic and semantic constraints can work with both types. However, grammar analysis and learning rules generations are complex. These techniques are costly for structured text. A web mining research support system must deal with both types because both contain useful information for research. The proposed solution is a hybrid information extraction system, shown in Figure a web document will be checked for its type by a selector. Wrapper induction techniques will be developed to extract information from a structured/semi-structured text, while NLP techniques are used for free text. For a free text type web document, grammar and syntax are analyzed by NLP learning. Then, extraction rules are generated based on the analysis. The extractor extracts data according to those extraction rules and, stores extracted data into the database. The extraction of a structured/semi-structured web document is as follows.

Firstly, the parser creates a parse tree for the document. Secondly, users input sample patterns which they want to extract. Then, extraction heuristics are generated to match the sample patterns. Wrappers are created based on extraction heuristics to extract data. If the extracted data contains free text which needs further extraction, the process will be changed to use NLP techniques. Otherwise, data are stored in the database.

Generalization

The purpose of generalization is to discover information patterns in the extracted web content. Analysis of web data can help organizations understand their users' information and improve their web applications. For example, many companies gathered web data to study their customers' interests, predict their behaviors and determine marketing strategies.

Generalization can be divided into two steps. The first step is preprocessing the data. Preprocessing is necessary because the extracted web content may include missing data, erroneous data, wrong formats and unnecessary characters. The second step is to find patterns by using some advanced techniques such as association rules, clustering, classification and sequential patterns.

Preprocessing

Preprocessing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of data cleansing, user identification, and session identification.

Data cleansing eliminates irrelevant or unnecessary items in the analyzed data.

A web site can be accessed by millions of users. Different users may use different formats when creating data. Furthermore, overlapping data and incomplete data also exist. Data cleansing, errors and inconsistencies will be detected and removed to improve the quality of data.

Another task of Preprocessing is user identification. A single user may use multiple IP addresses, while an IP address can be used by multiple users. To study users' behaviors, we must identify individual users. Techniques and algorithms for identifying users can be performed by analyzing user actions recorded in server logs. Session identification divides the page accesses of a single user, who has multiple visits to a website, into individual sessions. Like user identification, this task can be performed based on server logs. A server can set up sessions. Session IDs can be embedded into each URI and recorded in server logs.

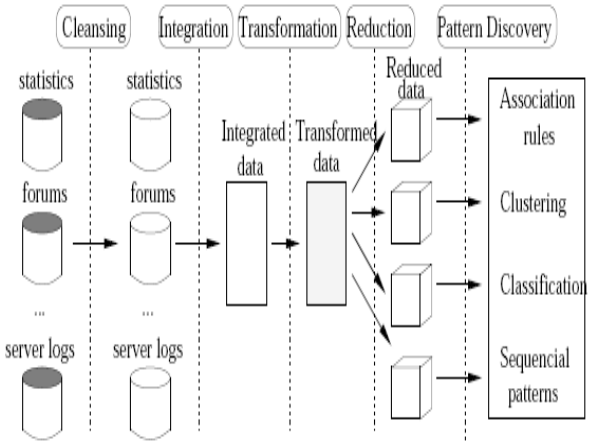
Discovery Techniques

Several data mining techniques can be applied in discovering patterns after web data have been preprocessed, examples of which are provided below.

- Association Rules find interesting associations or correlation relationships among a large set of data items. If X and Y are sets of items, association rule mining discovers all associations and correlations among data items where the presence of X n a transaction implies the presence of Y with a certain degree of condense. The rule condenses denied as the percentage of transactions containing X also containing Y.
- Association rule discovery techniques can be generally applied to the web mining research support system. This technique can be performed to analyze the behavior of a given user. Each transaction is comprised of a set of URLs accessed by a user in one visit to the server. For example, using association rule discovery techniques, we can and correlations in OSS study such as the following:
    1. 40% of users who accessed the web page with URL/project1, also accessed /project2; or
    2. 30% of users who accessed /project1, downloaded software in /product1.

With massive amounts of data continuously being collected from the web, companies can use association rules to help to make effective marketing strategies. Also, association rules discovered from WWW access logs can help organizations design their web page.

- Clustering is a technique to group a set of items having similar characteristics.
- Clustering is applied in the web usage mining to find two kinds of interesting clusters: usage clusters and page clusters [84]. Usage clusters group users who exhibit similar browsing patterns. Clustering of client information or data items can facilitate the development and execution of future marketing strategies. Page clusters discover groups of pages having related content. This information is useful for Internet search engines and Web assistance providers. By using clustering, a website can dynamically create HTML pages according to the user's query and user's information such as past needs.
- Classification is another extensively studied topic in data mining. Classification maps a data item into one of several predefined classes. One task of classification is to extract and select features that best describe the properties of a given class or category. In web mining, classification rules allow one to develop a profile of items belonging to a particular group according to their common attributes. For example, classification on SourceForge access logs may lead to the discovery of relationships such as the following:
    1. users from universities who visit the site tend to be interested in the page/project1; or
    2. 50% of users who downloaded software in /product2, were developers of Open Source Software and worked in IT companies.
- Sequential Patterns find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes [28]. In other words, sequential patterns refer to the frequently occurring patterns related to time or other sequences and have been widely applied to prediction. For example, this technique can be applied to web access server transaction logs on OSS web sites to discover sequential patterns to indicate users' visit patterns over a certain period and predict their future visit patterns:
    1. 40% of users who visited /project1, had done a searched SourceForge, within the past week on keyword x; or
    2. 70% if clients who downloaded software in /project1, also downloaded software in / project2 within 15 days.



The Generalization Infrastructure

Analysis and Validation

The analysis involves the validation and interpretation of the mined patterns. With data extracted from the web and information discovered from generalization, models can be built to simulate the studied phenomenon. The simulation models can be used to validate their correctness.

One objective of the Open Source Software study is to model OSS developers to understand their behaviors and be able to predict the developers' network development. This task can be divided into two-steps: simulate the developers' network and validate the simulation models. We use agent-based tools to simulate and validate the OSS developers' network. This proposal will be focused on how to perform validation.

Conclusion

This paper combines web retrieval and data mining techniques to provide an efficient infrastructure to support web data mining for research. This system is composed of several stages. Features of each stage are explored,and implementation techniques are presented. IR will identify web sources by predefined categories with automatic classification. IE will use a hybrid extraction way to select portions from a web page and put data into databases. Generalization will clean data and use database techniques to analyze collected data. Analysis and Validation will build models based on those data and validate their correctness. Even there the huge number of tools for research in web mining. This work consolidated the general idea on it.

References

Qingtian Han, Xiaoyang Gao, Wenguo Wu, (2008). "Study on Web Mining Algorithm Based on Usage Mining," Computer- Aided Industrial Design and Conceptual Design, 2008. CAID/ CD 2008. 9 th International Conference on 22-25.

Qingtian Han, Xiaoyan Gao, (2009) "Research of Distributed algorithm based on Usage Mining," Knowledge Discovery and Data mining, 2009, WKDD 2009, Second International Workshop on 23-25.

Ranieri Baraglia., & Fabrizio Silvestri, (2004). "An Online recommender System for LargeWeb Sites," Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on 20-24.

Suresh Babu, D. et al. / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5), 2011, 2390-2393 "Web Usage Mining: A Research Concept of Web Mining"

Madey, G, Freeh, V, Tynan, R, Gao, Y., &Ho_man, C (2003). Agent-based modeling and simulation of collaborative social networks. In Americas Conference on Information Systems (AMCIS2003), Tampa, FL.

Madey, G, Freeh, V, Tynan, R., & Ho_man, C. (2003). An analysis of open source software development using social network theory and agent-based modeling. In The 2nd Lake Arrowhead Conference on Human Complex Systems, LakeArrowhead, CA, USA.

Xu, J, Huang, Y., & Madey. G. (2003). A research support system framework for web data mining. In Workshop on Applications, Products and Services of Web-based Support Systems at the Joint International Conference on Web Intelligence (2003 IEEE/WIC) and Intelligent Agent Technology, pages 37{41, Halifax, Canada.

Zaiane, O. R, Xin, M., & Han. J. (1998). Discovering web access patterns and trends by applying OLAP and data mining technology on weblogs. In Advances in Digital Libraries, pages 19{29}.

International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-9, December 2015 ISSN: 2395-3470 www.ijseas.com ANALYSIS OF WEBSITE USAGE WITH USER DETAILS USING DATA MINING PATTERN RECOGNITION

OlfaNasraoui & Christopher Peter s: Combining Web usage Mining and Fuzzy Inference for Website personalization.

Cooley, R, Mobasher, B., &Srivastava, J. (1997). Web mining: Information and pattern discovery on the World Wide Web. In International Conference on Tools with Arti_cial Intelligence, pages 558{567, Newport Beach.

Laender, A, Ribeiro-Neto B, Silva, A., & Teixeira, J. (2002). A brief survey of web data extraction tools. In SIGMOD Record, volume 31,.

Gengbank homepage. http://www.ncbi.nlm.nih.gov/ Genbank/index.html.

_NCBI homepage. http://www.ncbi.nlm.nih.gov/.

Source forge homepage. http://sourceforge.net.

Chekuri, C, Goldwasser, M, Raghavan, P., & Upfal, E. (1996). A web search using automatic classification. In Proceedings of WWW-96, 6th International Conference on the World Wide Web, San Jose, US.

Chakrabarti, S, Berg, M., & Dom, B. (1999). Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks, 31(11{16):16231640.

Eikvil L. (1999). Information extraction from the world wide web - a survey. Technical Report 945, Norweigan Computing Center.

Axtell, R, Axelrod, R, Epstein, J., & Cohen, M. Aligning simulation models: A case study and results. Computational and Mathematical Organization Theory,1(2):123.

Chawathe, S, Garcia-Molina, H, Hammer, J, Ireland, K, Apakonstantinou, Y, Ullman, J. D., & Widom. J. (1994). The IMMIS project: Integration of heterogeneous information sources. In 16th Meeting of the Information Processing Society of Japan, pages 7{18, Tokyo, Japan.

Web Sources

https://www.slideshare.net/Tommy96/jinproposalslidesppt
https://www3.nd.edu/~oss/Papers/Jin_diss_proposal.pdf
https://www.ukessays.com/essays/computer-science/web-mining-research-support-system-computer-science-essay.php
https://www3.nd.edu/~oss/Papers/Jin_diss_proposal.pdf
https://www.slideshare.net/Tommy96/jinproposalslidesppt
https://www.slideshare.net/IOSR/h0314450-26684752
https://pdfs.semanticscholar.org/efe2/c1592d6b03a1ca88ccf63ed23b9366d8e87f.pdf
https://www3.nd.edu/~oss/Papers/Jin_diss_proposal.pdf