Journal of Engineering and Applied Sciences 12 (Special Issue 2): 6135-6142, 2017
ISSN: 1816-949X

A Study on Web Mining Tools and Techniques

Saleh Mowla, Ishita Bedi and Nisha P. Shetty
Department of Information and Communication Technology,
Manipal Institute of Technology, Manipal, 576104 Karnataka, India

Abstract: Web today has become a repository of knowledge in any form such as text, audio, graphics, video and multimedia. With the passage of time world wide web has become clogged up with various information making extraction of vital information arduous and cumbersome. Web mining is a branch of data mining which deals with searching, extracting and filtering useful data stored in web server databases and logs. This study is a detailed study of various techniques involved in mining web data on the basis of its application. The study describes various tools involved in web mining and is concluded with challenges faced, future aspects and applications.

Key words: Web content mining, WEKA, R, web usage mining, web structure mining, web miner

INTRODUCTION

With the development of this huge repository of information called as web, information is not limited to one computer and can be stored, accessed and updated from any computer located in any corner of the world. When user queries a search engine numerous links pop up containing data selection of vital information from these links is crucial. Information in web do not adhere to one single common format because of which mining and preprocessing the humongous data is an absolute necessity. Within the data available, there are many hidden patterns which cannot be detected at one go. In order to find these patterns, data mining is required which automates the task of analyzing the information based on the certain perspective to solve a given problem statement.

Web mining problems (Jayalatchumy and Thambidurai, 2013)

Data extraction: It is important to know if the data being mined is structured or unstructured and accordingly, machine learning and automatic extraction techniques can be used. Also, some data will be incorrect or incomplete and must be examined with great accuracy. Personal data in the web must be suitably protected from unauthorized access.

Information integration and schema matching: Different websites and pages may represent the same information in various manners; identification of similar data and classifying or categorizing them from the vast data warehouse (i.e., the internet) can be a difficult task.

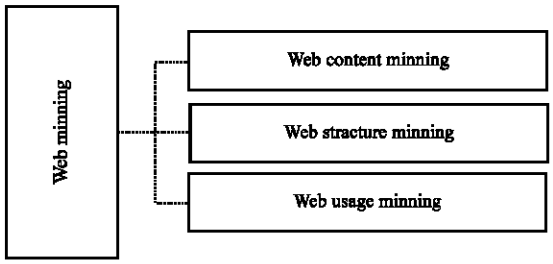

Fig. 1: Types of web mining

Opinion extraction: It is not easy to interpret the tone of opinions collected from various chatrooms, discussion forums and blogs; misinterpretation of data gathered will give a completely different result on analysis.

Knowledge synthesis: Concept hierarchies or ontology can be used in a variety of applications. However, manually generating them is not feasible as it takes a long time to do so. The aim here is to organize bits and pieces of information scattered around the web and get something valuable out of it.

Detecting noises: Very often the main content of any webpage goes unnoticed due to surplus amount of hyperlinks, advertisements, copyright notices, etc., in the web page. Extracting useful information is a tedious but a necessary process.

Types of web mining: Based on the target data web mining can be categorized into three categories (Singh and Singh, 2010) as shown in Fig. 1.

Corresponding Author: Saleh Mowla, Department of Information and Communication Technology,
Manipal Institute of Technology, Manipal, 576104 Karnataka, India

MATERIALS AND METHODS

Web content mining: This technique involves procedures to extract and integrate data from varied web page contents. It aims at evaluating and mapping the information to provide adequate results to user's queries (Malarvizhi and Saraswathi, 2013). Types of web content mining are shown in Fig. 2. When extracting web content information using web mining, there are four typical steps:

- Collect: gather the contents from the web
- Parse: mine usable data from formatted data (HTML, PDF, etc.)
- Analyze, tokenize, rate, classify, cluster, filter, sort, etc
- Produce turn the results of analysis into something useful (report, search index, etc.)

Unstructured data mining: Almost all the data on the web is unstructured. Unstructured data basically refers to data that doesn't fit in any database or structured form. Examples include text, images, videos, etc. Because of the dominance of unstructured data over other types of data, it becomes essential to mine it efficiently (Malarvizhi and Saraswathi, 2013; Saini and Pandey, 2015).

Data/information extraction: Data extraction helps to analyze results and provide services. Since, the data is very huge to extract meaningful information, patterns are matched. Certain keywords and phrases are traced and connections within the text are found. This technique helps extract information from large, unstructured data and the missing information is found too using other rules. Some predictions can be incorrect in which case it is discarded.

Topic tracking: This technique studies the type of documents already viewed by the user and the user's profile. After analysis, it predicts and suggests documents related to the user's interest. Many search engines use this technique. The main task of this technique is to study a stream of resources to find the particular topic which is contained in the positive samples (Gupta and Lehal, 2009). This information is generally scattered and the technique might suggest irrelevant documents as well. It can be applied to fields such as education, medical, finance, business, etc.

Clustering: A cluster is a group of similar objects. For cluster analysis, data is divided into groups based on similarity and labelling is done within groups. Many types of partitioning can be done such as soft, hard, etc. and an object can be allowed or disallowed to be a part of multiple clusters or the objects may be related in a hierarchical manner. Clustering is usually done based on fly because of which useful documents are not omitted (Johnson and Gupta, 2012). This helps user to select a topic of interest.

Information visualization: It is "the use of computer-supported, interactive, visual representation of abstract data to amplify cognition. Useful for finding

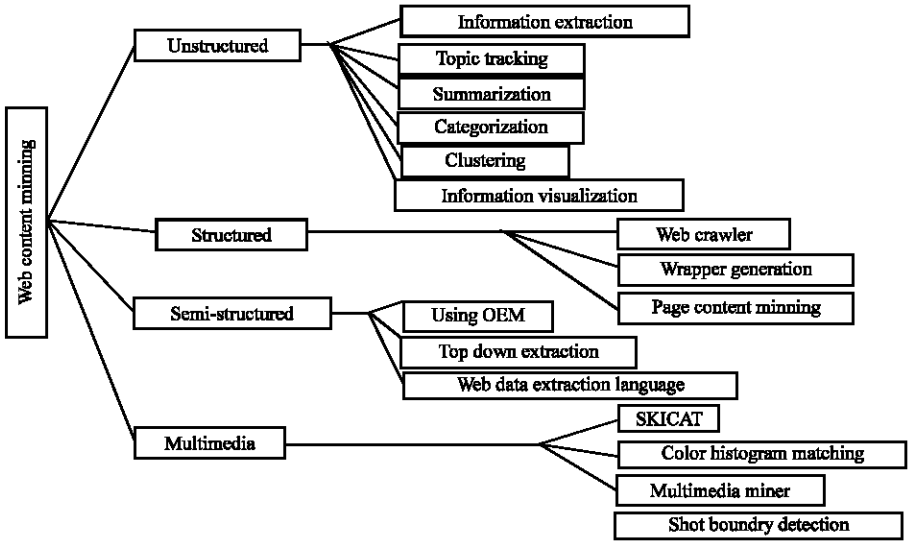

Fig. 2: Web content minning types

similar or related topics from a huge set of data or documents, it is a representation of data in an abstract way in a graphical form using feature extraction and key term indexing. Huge texts are represented hierarchically in graphical form which can be analyzed by zooming in scaling, etc. (Saini and Pandey, 2015).

Summarization: This technique helps to reduce the length of the document thereby keeping only the main points. It is useful to get the gist of the topic. The time taken to summarize a text is comparatively very less. The software should analyze the semantics, scan headings and subheadings, interpret meanings, etc., to summarize a given text. Examples are Microsoft's auto-summarize, online text compactors, etc.

Structured data mining: Highly organized data is classified as structured data. It can refer to data within a file or a record. The degree of organization is such that insertion into databases is seamless and searchable by simple algorithms and search operations as opposed to unstructured data.

Web crawler: A web crawler or web spider is an automated script that browses the web in a planned way. The crawlers regularly scan the content of the pages on the web for the words and the location of the words in the pages. This is converted to an index which is essentially a list of words and web pages they reside in. The external crawler traverses an unknown website and the internal crawler traverses the internal pages of that website. Types include focused, incremental, distributed and parallel web crawlers.

Page content mining: Traditional search engines rank pages based on which pages are retrieved and classified. The results are displayed according to the rank and classification is done as per their importance based on their PCR (Page Content Rank).

Wrapper generation: Traditional search engines rank web pages. In wrapper generation information is provided on the capability of sources. Sources are the query they will answer and the output types. Based on the query, the web pages are retrieved on the basis of their page rank value. Wrappers provide meta-information like statistics, domains, etc. (Saini and Pandey, 2015).

Semi-structured data mining: Semi structured data is not raw data but rather structured data which is not stored in a systematized manner like tables. As the documents in the web unite from diverse sources it is not feasible to store such data in a single format.

Object Exchange Model (OEM): The OEM helps to understand the information structure on the web more accurately and is best suited for an ever-changing and heterogeneous environment. The structures of objects are self-describing in nature.

Web data extraction language: Web data is converted to structured form and stored in a tabular form. End users can hence access it.

Top down extraction: Complex objects from rich web sources are extracted and changed into less complex objects until the most atomic ones are extracted.

Multi-media data mining: Multimedia data mining is the process of examining stimulating patterns from media data like graphics, audio, video and text.

SKICAT: This system is an astronomical data analysis system. It is used to produce a digital catalog of the sky objects. It is a mix of image processing and data classification and classifies objects into human usable classes using machine learning.

Color histogram matching: Two basic constituents of color histogram matching are equalization and smoothing. The correlation between color components is equalization. It suffers from one problem which is the sparse data problem where unwanted artifacts are present in equalized images. Smoothening solves this problem.

Multimedia miner: Here, the mage excavator extracts images and videos and the preprocessor extracts image features and stores them in a database. A search kernel matches queries with the images and videos in the database and the discovery module does image information mining routines so as to trace the patterns in images.

Shot boundary detection: This method helps to automatically detect the boundaries between shots in videos.

RESULTS AND DISCUSSION

Web usage mining: The process of examining the interaction of users with the Web by studying web logs in order to improve personalization and provide better search engines (Omar *et al*., 2014).

Steps used in web usage mining

Data gathering: Web logs are the records on which the server stores information about the user's activities on the web. It can be present on server, client or proxy

side and contains vital information such as user's domain, subdomain, hostname, resources accessed and so on.

Data preprocessing: It involves picking out and refurbishing desired user's entries and contents of their session.

Pattern discovery, analysis and visualizations: Log records are scrutinized to cognize the usage profile of an exact user.

Application: This knowledge can be applied to improve business in many e-Commerce sectors.

Techniques

Association rule mining: The algorithms in this category are applied in order to improve the design of web space and decide where to put hyperlinks which pages to connect and also to predict the next attention-grabbing page to the user and so on.

Clustering: These algorithms group the similar elements together and provides a clear distinction between divergent elements. Two varieties of clusters can be fashioned.

Page cluster: Groups pages of similar content together. Usage cluster, clusters a group of users having analogous surfing patterns.

Classification: It provides classes to web users based on their browsing behaviors. This classification is useful for sectors such as e-Commerce to further their businesses.

Web structure mining: Web structure mining is used to find connections between different web pages which are connected either by some information or by some direct connection or link. Connections are beneficial, so that search engines are able to extract the web pages from web sites based on the search query directly. This is done by spiders which scan web sites, get to the home page and thereon link information through these reference links to get to the particular page. Web structure mining uses graph theory to do so. Web structure mining involves two basic tasks: extraction of patterns from hyperlinks in the web and analysis of the tree like document structure.

Algorithms

Google's Page Rank algorithm: The web pages are ranked based on the number of backlinks pointing to them. A total page rank is assigned to all the pages based on the page ranks of the backlinks pointing to them (Sangeetha and Joseph, 2014). Essentially, page rank is a vote given to the webpage by all other webpages based on its importance (Page *et al*., 1999). Every link counts as a vote of support and an absence of a link means no support (which means there has been no vote, not that the page has been voted against) (Page *et al*., 1999). The page rank of a page A is given by:

$$PR(A) = (1-d)+d\,(PR(T1)/C(T1)+, ..., +PR(T_n)/C(T_n))$$

Google toolbar shows the page rank of your webpage (actually something like log base 10 of the actual page rank).

HITS algorithm: Used by the Ask.com search engine, this algorithm makes use of the link structure of the web in order to find and rank pages relevant to certain topics. First, the most relevant pages to the query are retrieved which can be done in many ways this set of most relevant pages is called the root set. Thereafter, the links of the webpages are explored and all the web pages that are linked from it and some of the web pages that link to it are added to the root set forming a base set. The web pages in the base set and the hyperlinks among those pages form a subgraph on which the HITS computation occurs (Devi *et al*., 2014).

Two values, namely authority and hub values are defined in terms of each other. Authority value is the sum of the scaled hub values that point to that page. Hub value is the sum of the scaled authority values of the pages that it points to. The algorithm is query based and iterative with each step involving two basic steps.

Authority update: Each node's authority score is updated to make it equal to the sum of the hub scores of each node that it points to. Hence, a node which is linked from pages recognized as hubs for information is given a higher authority.

Hub update: Each node's hub score is updated to make it equal to the sum of the authority scores of each node that it points to. Hence, a node is linked to nodes that are regarded as authorities on the subject to give it a high hub score.

Tools used in web mining: Different data mining tools have different features which is why companies have to consider a number of factors before deciding which tool needs to be installed.

Volume of data: Based on the amount of data that needs to be analyzed, the company has to decide how powerful it wants its data mining system to be the more powerful it is the more expensive it will be.

Amount of pre-processing: If the data is retrieved from relational databases, it is easier to analyze it for most data mining systems. However, in other cases, the data will first need to be processed in a manner that the system can understand and hence, analyze it.

Storage: The manner in which the data is stored needs to be considered if the data is stored in databases then the data mining system must be able to work with database or else more complex systems will be required to both retrieve and analyze the data from large data streams.

Analysis complexity: If the analysis is simple, then a simpler and affordable system can be put in place; if the complexity of the analysis is required to be more then a system with advanced features will be needed.

Tasks to be performed: Depends on what kind of operations need to be performed this include clustering, regression, association, classification, etc.

Scalability: The company's system should be able to handle larger volumes of data if its database needs to be expanded.

Flexibility: There exists many data mining algorithms that can be implemented for the same data mining task. The data mining system should be able to adapt to various types of analysis.

User-friendly: Not all users of the data mining system is well-versed with the technicalities which is why visualization tools help make the presentation of the result more appropriate and comprehensible. Table 1 and 2 depicts overview of all the tools in the area of web minning.

Table 1: Technical overview of results

Tool name	Technical overview	General features	Specialization/applications	Advantages	Limitations	Areas
WEKA (Rangra and Bansal, 2014)	Released in 1997; licensed by GNU; platform-independent; Java-compatible	Open source; variety of data mining and machine learning algorithms, provides three GUIs: Explorer, experimenter, knowledge flow	To mine association rules; machine learning techniques	Can develop new machine learning schemes; data file formats: binary, CSV, ARFF, C4.5; easy to integrate into other Java packages	Poor documentation; Poor connectivity to Excel spreadsheet and non-Java databases; weak in classical statistics; automatic parameter optimization for machine learning is unavailable	Web usage mining
KEEL (Rangra and Bansal, 2014)	Released in 2004; licensed by GNU; platform-independent; Java-compatible	Provides various machine learning tools; vast collection of libraries for pre-and post-processing techniques	Assessing evolutionary algorithms for data mining problems; suited for machine learning	Includes clustering, regression, classification, pattern mining; contains hybrid models, algorithms based on computational intelligence and rule learning	Supports limited number of algorithms compared to other tools	Web usage mining
R (Rangra and Bansal, 2014)	Released in 1997; licensed by GNU; platform-independent	Open source and free; well supported; for analyses and graphical and software development activities	Statistical computing; bio-informatics; social-science	Vast statistical library; easier to optimize machine learning code; better graphics; more transparent; easier import and export of data from spreadsheets	Hardly well-oriented with data mining; difficult to learn and progress further	Web usage mining
KNIME (Rangra and Bansal, 2014)	Released in 2004; licensed by GNU; compatible with windows, Linux and OS X; Java-compatible	Works in IBM's eclipse development environment; modular data exploration platform; incorporates >100	Chemical structures and compounds; data mining,enterprise reporting; business intelligence	Integrates all analyses modules of WEKA; allow R-scripts to run; easy to install; ability to interface	Limited number of error measurement methods; wrapper methods for descriptor methods are unavailable; automatic parameter	Web content mining

Table 1: Continue

Tool name	Technical overview	General features	Specialization/applications	Advantages	Limitations	Areas
		processing nodes for various applications (I/O, pre-processing, modeling, data mining)		with programs that visualize and analyze molecular data	optimization for machine learning is unavailable	
Rapidminer (Rangra and Bansal, 2014)	Released in 2006; licensed by AGPL; proprietary; platform-independent; language independent	Uses a modular operator concept that allows complex design of nested operator chains; uses XML for representation if needed; supports about 22 file formats; includes learning algorithms from WEKA; reads and writes Excel files	Predictive analysis; statistical computing	Effective model evaluation using cross validation and independent validation sets; over 1500 methods for data transformation, analysis, modeling and visualization; offers numerous procedures in the areas of attribute selection and outlier detection	Mostly suited for users who are able to work with database files	Web content mining
Orange (Rangra and Bansal, 2014)	Released in 2009; licensed by GNU; compatible with Python, C and C++	Component-based software for data mining and machine learning; data mining is done through visual programming or python scripting	Open source data visualization; text mining; bioinformatics; data analytics	Can be used as a script; works well with an ETL work flow GUI; easiest tool to learn; better debugger; simpler scripting of data categorization problems; GUI is cross-platform	Large installation; limited number of machine learning algorithms; weak in classical statistics; provides no widgets for statistical testing; reporting capability is restricted to exporting visual representation of models	Web content mining
Tableau	Display of any unicode character set	Trend lines: regression analysis of linear, polynomial, logarithmic and exponential data functions; Forecasting: prediction of time series based on historical data	Data visualization	User friendly: drag-and-drop interface; can be easily integrated with R programming; works well with other data mining tools like KNIME	Not intended for data mining or predictive analysis	Web usage mining
Scrapy	Supports windows, Linux, MAC OS; compatible with Python	Open source and free; non-commercial; written in Python	Extraction of structured data	Useful for testing web pages; helps in monitoring	Use of this tool is difficult compared to other tools	Web content mining
Web Information Extractor (WIE) (Herrouz *et al*., 2013)	Supports windows 2000/ XP/Vista OS	Commercial tool; extraction of structured and non-structured data; data export formats are Excel (CSV) and text (TXT)	Web content extraction	Monitors the web page constantly to detect any changes; ability to multi-task; Supports recursive task definition	Loading of website can be time-consuming; not possible to record the data	Web content mining
Web Data Extractor (WDE)	Supports windows 95/98/2000/XP; default export file format is Excel (CSV)	Commercial tool; extraction of URLs, phone, meta tags, e-Mail addresses, etc.	Web content extraction	Easy to use and relatively comprehensive; settings can be changed according to the user preference	Highly automated: extensive training required	Web content mining
Mozenda; (Shanmugapriya and Kiruthika, 2014)	Platform-independent (note: mozenda agent builder runs only on windows)	Commercial tool; web console section: allows users to run agents and publish results	Mine and manage data	Easy to use; smart filtering of user's text; rotating IP prevents user identification	Not possible to record the data	Web content mining

Table 1: Continue

Tool name	Technical overview	General features	Specialization/applications	Advantages	Limitations	Areas
		of extracted data; agent producer section: windows application to construct project associated with data extraction				
Web Content Extractor (WCE) (Herrouz *et al*., 2013)	Supports windows OS; export formats: MS Excel (CSV), Access, TXT, HTML, XML, SQL, MySQL	Commercial tool; known for crawling and web spiders; collect data from password protected sites	Real estate data; online auctions Job seeking	User friendly wizard interface; easy to create crawling rules; ability to download data as a multi-subject	Not possible to record the data	Web content mining
Screen scraper; Shanmugapriya and Kiruthika (2014)	Can be integrated with languages like PHP, Java, ASP, NET	Commercial tool; can search for content from databases; (Malarvizhi and Saraswathi, 2013) extracted data can be downloaded into a spreadsheet	Meta-search engines	Easy automation of website tasks (filling a form, open links, etc.)	Not possible to record the data	Web content mining
Automation anywhere (Shanmugapriya and Kiruthika, 2014)	Export formats: XML, Excel, TXT, MySQL	Commercial tool; ability to repeat an action during hours, minutes or seconds; ability to specify the rate of the required action; Scheduler: schedules an action at a particular time	Web record and web data extraction	Easy to use and faster; it can record the data	Highly automated: extensive training to required	Web mining mining

Other data mining tools are listed as follows: Chaudhary and Gupta (2013), Thakkar *et al*. (2016)

Table 2: Features of tools

Tool name	Features	Areas
Web miner	Helps in the mining of useful patterns, provides user-specific information	Web usage mining
Import.Io	Relatively better GUI than that of commercial tools, various crawler options are available, structured data can get extracted, access to integrated data can be done easily online	Web content mining
I-miner	Discovers data clusters, uses fuzzy clustering algorithm and fuzzy inference system	Web usage mining
Speed tracer	Used in the mining of web server logs, reconstructs user navigational path for session identification	Web usage mining
Web log miner	Helps in the extraction and presentation of various kinds of reports, supports the extended W3C log format data is stored in the PostgreSQL database	Web usage mining
Koinotites	A personalization tool, used for the construction of user communities in the web	Web usage mining
Context miner	Free tool for mining online web content, export options: XML and CSV	Web content mining
Irobotsoft	Robot performs website related activities, multiple automatic data extraction from different websites	Web content mining
Mining mart	Information is processed from relational databases, supports PostgreSQL, MySQL and Oracle	Web content mining
TraMineR	An R-package for mining, describes and visualizes sequences of states or events	Web usage mining

CONCLUSION

Web mining is a branch of data mining which deals with mining heterogeneous and vast data available in the gold mine of information, i.e., the world wide web. Web mining helps the user to scrutinize and filter out the data useful to the users in an effective manner. This study incorporates a detailed study of various tools and techniques involved in mining the data in the web. Future work with respect to web content mining would be personalization of web or predicting user needs effectively by proper content interpretation and selection of appropriate data to satisfy user needs.

REFERENCES

Chaudhary, K. and S.K. Gupta, 2013. Web usage mining tools and techniques: A survey. Intl. J. Sci. Eng. Res., 4: 1762-1768.

Devi, P., A. Gupta and A. Dixit, 2014. Comparative study of HITS and page rank link based ranking algorithms. Intl. J. Adv. Res. Comput. Commun. Eng., 3: 5749-5754.

Gupta, V. and G.S. Lehal, 2009. A survey of text mining techniques and applications. J. Emerg. Technol. Web Intell., 1: 60-76.

Herrouz, A., C. Khentout and M. Djoudi, 2013. Overview of web content mining tools. Intl. J. Adv. Res. Comput. Sci. Software Eng., 3: 375-385.

Jayalatchumy, D. and P. Thambidurai, 2013. Web mining research issues and future directions: A survey. IOSR. J. Comput. Eng., 14: 20-27.

Johnson, F. and S.K. Gupta, 2012. Web content mining techniques: A survey. Intl. J. Comput. Appl., Vol. 47,

Malarvizhi, R. and K. Saraswathi, 2013. Web content mining techniques tools & algorithms: A comprehensive study. Intl. J. Comput. Trends Technol., 4: 2940-2945.

Omar, R., A.O.M. Tap and Z.S. Abdullah, 2014. Web usage mining: A review of recent works. Proceedings of the 5th IEEE International Conference on Information and Communication Technology for The Muslim World (ICT4M), November 17-18, 2014, IEEE, Kuching, Malaysia, ISBN:978-1-4799-6243-3, pp: 1-5.

Page, L., S. Brin, R. Motwani and T. Winograd, 1999. The pagerank citation ranking: Bringing order to the web. Technical Report Stanford InfoLab. http://ilpubs. stanford.edu:8090/422/.

Rangra, K. and K.L. Bansal, 2014. Comparative study of data mining tools. Intl. J. Adv. Res. Comput. Sci. Software Eng., 4: 216-223.

Saini, S. and H.M. Pandey, 2015. Review on web content mining techniques. Intl. J. Comput. Appl., Vol. 118,

Sangeetha, M. and K.S. Joseph, 2014. Page ranking algorithms used in Web Mining. Proceedings of the IEEE International Conference on Information Communication and Embedded Systems (ICICES2014), Feburary 27-28, 2014, IEEE, Chennai, India, ISBN:978-1-4799-3698-4, pp: 1-7.

Shanmugapriya, T. and P. Kiruthika, 2014. Survey on web content mining and its tools. Intl. J. Sci. Eng. Res., 2: 1-4.

Singh, B. and H.K. Singh, 2010. Web data mining research: A survey. Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research, December 28-29, 2010, IEEE, Coimbatore, India, ISBN:978-1-4244-5965-0, pp: 1-10.

Thakkar, P., G. Bhatt, A. Kurtkoti, S. Shah and C. Joshi, 2016. A survey: Web mining tools and technique. Intl. J. Latest Trends Eng. Technol., 7: 212-217.