RESEARCHERID
THOMSON REUTERS

[DOKANIA * *et al.*, 7(3): March, 2018]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A STUDY ON MINING PRODUCT OPINIONS AND REVIEWS ON THE WEB USING WEB SCRAPING

MR. Nilesh Kumar Dokania[*1] & MS. Jaspreet Kaur[2]
[*1]Assistant Professor, Guru Nanak Institute of Management
[2]Research, MCA, Guru Nanak Institute of Management

DOI: 10.5281/zenodo.1199202

ABSTRACT

As enterprises expand and post increasing information about their business activities on their websites, website data promises to be a valuable source for investigating innovation. This article examines the practicalities and effectiveness of web mining as a research method for innovation studies. We use web mining to explore the R&D activities of 296 UK-based green goods small and mid-size enterprises. We find that website data offers additional insights when compared with other traditional unobtrusive research methods, such as patent and publication analysis. We examine the strengths and limitations of enterprise innovation web mining in terms of a wide range of data quality dimensions, including accuracy, completeness, currency, quantity, flexibility and accessibility. We observe that far more companies in our sample report undertaking R&D activities on their web sites than would be suggested by looking only at conventional data sources. While traditional methods offer information about the early phases of R&D and invention through publications and patents, web mining offers insights that are more downstream in the innovation process. Handling website data is not as easy as alternative data sources, and care needs to be taken in executing search strategies. Website information is also self-reported and companies may vary in their motivations for posting (or not posting) information about their activities on websites. Nonetheless, we find that web mining is a significant and useful complement to current methods, as well as offering novel insights not easily obtained from other unobtrusive sources.

Web Mining is extracting information from the web re-sources and finding interesting patterns that can be useful from ever expanding database of World Wide Web. Whenever we talk about data, we conclude that there is a huge range of data on World Wide Web. Due to heterogeneity and unstructured nature of the data available on the WWW, Web mining uses various data mining techniques to discover useful knowledge from Web hyperlinks, page content and usage log. Web Content Mining is a component of Data Mining. The main uses of web content mining are to gather, categorize, organize and provide the best possible information available on the Web to the user requesting the information. This paper deals with a preliminary discussion of Web content mining, contributions in the field of web mining, the prominent successful tools and algorithms.

Keywords: Web mining, Web scraping, Innovation, R&D, Web content mining, structured data mining, unstructured data mining, semi-structured data mining.

I. INTRODUCTION

Enterprises use their publicly-viewable websites for a variety of reasons, including promoting their products and services, directly selling those products and services, presenting information about their development, capabilities and credentials, documenting their achievements, and expanding their customer base, especially in export markets (Fisher et al. 2007). Enterprise websites often also contain valuable information about the company's location(s) and facilities, specifications of products and services offered, the orientation and attitude of the firm, key personnel, and strategies and relationships with other firms and organizations.

The ever-growing amount of information that is available through enterprise websites offers significant opportunities for researchers. With the understanding that websites are self-reports, website information has advantages in that it is readily and publicly available, is cost-effective to obtain, and can be extensive in terms of coverage and the amount of data contained. In an era where corporate response rates to voluntary academic

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

research questionnaire surveys are frequently rather low, those same corporations maintain a relatively high website presence. For example, just under three-quarters of UK companies with at least one employee indicate that they maintain a website—a figure that rises to 85 % for companies (all sectors) with 10 or more employees and 91 % for manufacturing companies with 10 or more employees.[1]

We recognize that the information that is available on enterprise websites is not standardized, varies according to the company and how it wishes to present itself, and is typically in an unstructured format. Yet, notwithstanding these and other caveats, we suggest that the data that can be found on enterprise websites is an additional and important source of information and intelligence, particularly in addressing questions related to innovation where other data sources are less effective in gathering sufficient and relevant information. Technical advances in handling and analyzing unstructured data as well as current interest in the use of "big data" in discovering patterns and trends make it timely to investigate the appropriateness of using enterprise websites as a data source. However, while there are significant benefits to using website data through methods such as web scraping or web mining in innovation research, the literature on the use and validity of these approaches is relatively underdeveloped. This article aims to address this issue by analyzing the usefulness of website data in comparison with other data sources. We explore a web-derived dataset to discuss methodological issues related to the processes of conceptualizing, retrieving, structuring, cleaning, manipulating and interpreting website data in understanding company innovation strategies, focusing on enterprise research and development (R&D) activity.

In the next section, we review the available literature on the use of website data in innovation studies and wider social science applications. This is followed by discussion of our sample dataset and the methodologies used to obtain and analyze the data. In the ensuing section, there is an empirical demonstration of how the web mining process is operationalized in identifying R&D activities, as well as how this data correlates with other data sources. We then present a conceptual discussion of the relative qualities of website data for evaluating the enterprise R&D activities in comparison with other data sources. A discussion of the results and conclusions, as well as limitations, is contained in the final section.

II. LITERATURE REVIEW

Researchers often select methods such as web mining due to their "unobtrusiveness". Webb et al. (1966) first coined the term "unobtrusive measures" in reference to methods of data collection that do not require direct contact with research subjects. Conversely, obtrusive methods can be regarded as those that require direct contact with the population studied.

These methods are each suitable to different circumstances, depending on what is being studied. For example, the opinions and beliefs of individuals are often best explored through interviews or questionnaires—obtrusive methods. However, if the research concerns real actions and behaviors, these may best be observed from a distance—using unobtrusive methods. Unobtrusive methods are a way of collecting data about a subject without their direct knowledge or participation **(Cargan** 2007**).** Unobtrusive methods can be less expensive in that they do not involve the costs of training and placing researchers in the field and following up directly with respondents. Additionally, as **Lee (**2000**)** discusses, one major advantage of using "non-reactive" approaches **(Webb et al.** 1981**)** is that they avoid problems caused by the researcher's presence. In the case of obtrusive methods, the respondents are aware of the researcher and may alter their response to these research methods in light of this. Unobtrusive methods are also not limited to those who are accessible and cooperative **(Webb et al.** 1966**). Lee (**2000**)** also outlines the opportunity that internet data presents in unobtrusive research.

In the field of innovation studies, there has long been the use of a combination of obtrusive methods (such as innovation surveys of firms or business case studies) and unobtrusive methods (such as analyzing databases of patents and publications). More recently, innovation researchers have demonstrated increasing creativity in developing more diverse unobtrusive methods, many of which use website or social media data. **Robson (**2002**)** distinguishes between the more-traditional unobtrusive approaches already discussed and another—content analysis. The author describes content analysis as that conducted on a written document, such as books, letters and newspapers. However, we can see how this can be extended to analyzing the textual content of a website, through the process of web mining. Robson states that such an approach is different from other unobtrusive methods as the observation itself is indirect (i.e. there is no need to observe the participants – in our case, a group of companies—directly).

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

There are three general categories of web mining **(Miner et al.** 2012**).** *Web content mining* involves the analysis of unstructured text data within webpages to extract structured information. *Web structure mining* focuses on analyses of the hyper-linked structure of a set of webpages, typically using methods of network analysis. *Web usage mining* is the data mining process involving the usage data of webpages. All three types of web mining have been used in innovation studies.

An example of web structure mining in innovation studies is offered by **Katz and Cothey (**2006**)** who investigate relationships between the internet and innovation systems by utilizing website-based indicators from webpage counts and links. Another instance of web structure mining is from **van de Lei and Cunningham (**2006**),** who employ website data in a future-oriented technology analysis, where it is used to identify existing networks that are concerned with technological change. In this research, a web crawling process is used to identify linkages between nanotechnology web portals, creating a network of activity between parties across many sectors. **Ladwig et al. (**2010**)** use web structure mining to study the landscape of online resources in emerging technologies by identifying the top search terms and resulting top-ranked webpages from Google. Similarly, **Ackland et al. (**2010**)** use web crawling to capture hyperlinks: examining the relationships between, and prominence of, actors engaged in nanotechnology. The use of metrics based on web presence in measuring scientific performance ("webometrics") has widely been discussed in science policy literature **(see Thelwall (**2012**)** for an overview). Webometrics approaches use both web structure mining and web usage mining.

More recently, innovation scholars have been applying web content analysis in their research. **Veltri (**2013**)** carried out semantic analysis on 24,000 tweets from Twitter to understand the public perception of nanotechnology. **Libaers et al. (**2010**)** examine keyword occurrence in company websites from a cross-industry sample of small and medium-size enterprises to identify commercialization-focused business models among highly-innovative firms. Hyun Kim (2012) conducted both web-content and web-structure analysis of nanotechnology websites across the "Triple Helix" **(Etzkowitz and Leydesdorff** 2000**)** of university, government and enterprise relationships. The former allowed the author to discern different lexicons from three sectors, while the latter offered an understanding of which organizations played key roles in the development of an emerging technology.

Two recent studies are notable for examining the commercialization of emerging technologies by small and medium-sized firms through web content analysis. **Youtie et al. (**2012**)** examine current and archived website data of nanotechnology small and medium-sized enterprises, with a particular focus on the transition of such technologies from discovery to commercialization. The authors note the problems of coverage, timeliness, and response rate in commonly used sources of information such as patent databases and surveys in understanding enterprise innovation in rapidly transforming domains. A new approach—one which uses current and archival website data—is proposed. This method involved identifying and mining content information found on the websites of a pilot sample of 30 small and medium-sized enterprises from the United States, then analyzing the unstructured data in order to draw findings. The authors note that smaller firms tend to have smaller websites, therefore making the web mining process and subsequent analysis more manageable in such cases. From their analysis of the website data, the authors were able to identify the occurrence of various innovation stages and production transitions in the development of their sample of enterprises. The paper also discusses the role of government research grants and venture capital investment in bringing a technology to market.

The second study by **Arora et al. (**2013**)** undertakes a similar web content analysis method to examine the activities of small and medium size enterprises in the US, UK and China commercializing emerging graphene technologies. The authors again discuss the limitations of conventional methods, including issues of response rate and bias in surveys, and coverage and time lag in bibliometric and patent data. The study employs a web crawling technique of searching for keywords across all webpages of the sample firms' websites. This allowed the authors to not only draw conclusions on the degree of innovation employed by the sample firms but also the extent to which these activities were globalized and in partnerships, using such analysis to characterize three different types of emerging technology SME.

Web mining has also been used in other areas of social science. **AleEbrahim and Fathian (**2013**)** develop a method to summarize customer online reviews from websites. **Al-Hassan et al. (**2013**)** investigate whether the North American Industry Classification System code (NAICS) effectively shows the true industrial sectors of Fortune 500 firms by analyzing their websites. **Battistini et al. (**2013**)** present a technique to map geo-tagged geo-hazards, such as landslides, earthquakes and floods, by analyzing online news. **Hoekstra et al. (**2012**)**

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

investigate the feasibility and desirability of the automated collection of official statistics, such as consumer price index, from websites. There is a stream of publications concerning the mining of political opinions from websites, forums and social media **(Sobkowicz et al.** 2012**; Sobkowicz and Sobkowicz** 2012**).** There are also attempts to use web mining in health research: for instance content mining of website discussion forums to detect concern levels for HIV/AIDS (Sung et al. 2013) and mining social media to discover drug adverse effects **(Yang et al.** 2012**).**

III. WEB SCRAPING

Web scraping is a technique used to harvest information from web pages based on script routines. Web pages are documents written in Hypertext Markup Language (HTML), and more recently XHTML which is based on XML. Web documents are represented by a tree structured called the Document Object Model, or simply the DOM tree and the goal of HTML is to specify the format of text displayed by Web browsers as shown in figure 1

Since web scraping uses data from other online sources, it is important to consider the deliberately use of this technique, as the content from one web site is usually copied and republished by another one. Usually is sufficient to read their terms of service to ensure that data reuse is allowed, or to simply ask permission for that. From the operation perspective, a web scraping resembles in many ways a manual copy and paste task. The difference here is that this job is done in a organized and automatic way, by a virtual computer agent. When an agent is following each link of a web page, it is actually performing the same operation that a human being would normally do when interacting with a web site. This agent can follow links (by issuing HTTP GET requests) and submit forms (through HTTP POST), browsing through many different web pages. While a computer would perform manual tasks at the speed of a computer instruction, a human would have to think, grab the mouse, point to the link and finally click on it. Now the benefit seems clear when a user has to click on several links before getting to the actual desired page. However, not surprisingly, the benefit of issuing requests at a script speed, also brings a problem. If one uses web scraping without a policy for limiting requests, the requested server may find that someone is trying a Denial-of-Service attack, due to the great amount of requests triggered in a short period of time.

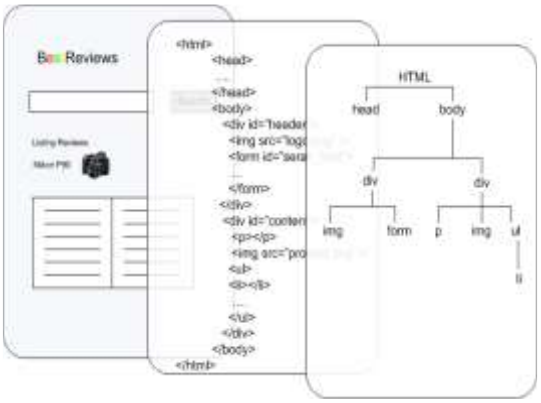

Figure 1: A web document from three different perspectives - From left to right: The document presentation, the HTML code and the DOM

After retrieving the target web document, the parser follows user-specified paths inside the document to retrieve the desired information. These paths are specified by CSS selectors or XPATHs. They use both relative and absolute paths (based on the DOM tree) to point the parser to a specific element inside a web document. After locating the desired information, normally web scraping operations uses also regular expressions to narrow or prune the located information, in order to retrieve data with an user- specified granular size. This process is illustrated in figure 2. A big shortcoming of web scraping, is the dfficulty to generalize extraction scripts. The script is usually tied up to the DOM model of a given page, therefore the dependence introduced by XPATHs or CSS Selectors, make it not easily reusable through different web sites. Also, considering performance web scraping may not be an optimal solution for retrieving information, specially when using it in larger scales or for commercial solutions. The request of an entire document when just a small part of it is actually necessary, makes it a very expensive process from the performance point of view as illustrated in figure 2. However, still with the mentioned shortcomings, web scraping can be a very powerful technique (an possibly the only public

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

known method), when no other option for retrieving information on an user- specified size is available. commercial solutions. The request of an entire document when just a small part of it is actually necessary, makes it a very expensive process from the performance point of view as illustrated in figure 3. However, still with the mentioned shortcomings, web scraping can be a very powerful technique (an possibly the only public known method), when no other option for retrieving information on an user- specified size is available.

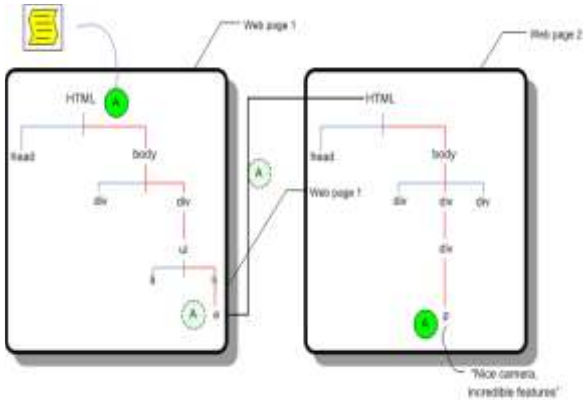

Figure 2: A web scraping agent gathering information from web pages – The dotted circle represents a web scraping agent traversing a DOM tree. The red lines are XPATHs to a desired element within the document. The agent reaches the hyperlink in web page 1 and proceeds to web page 2 until it finds the information enclosed by element p (paragraph)

IV. WEB CONTENT MINING ALGORITHMS

There are two common tasks involved in web mining through which useful information can be mined. They are Clustering and Classification. Here various classification algorithms used to fetch the information are described

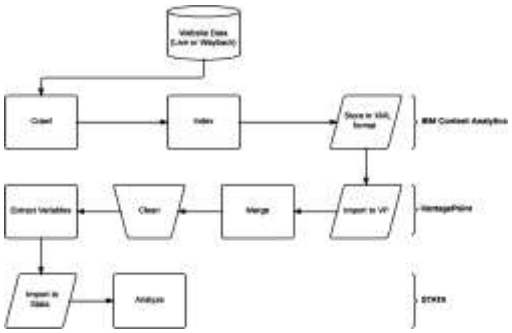

Fig. 3: Web content analysis process

4.1 Decision Tree: The decision tree is one of the powerful classification techniques. Decision trees take the input as its features and output as decision, which denotes the class information. Two widely known algorithms for building decision trees are Classification and Regression Trees and ID3/C4.5.

The tree tries to infer a split of the training data based on the values of the available features to produce a good generalization. This split at each node is based on the feature that gives the maximum information gain. Each leaf node corresponds to a class label. The leaf node reached is considered the class label for that example. The algorithm can naturally handle binary or multiclass classification problems. The leaf nodes can refer to either of the K classes concerned [12].

4.2 k-Nearest Neighbour: KNN is considered among the oldest nonparametric classification algorithms. To classify an unknown example, the distance (using some distance measure e.g. Euclidean) from that example to every other training example is measured. The k smallest distances are identified, and the most represented class in these k classes is considered the output class label. The value of k is normally determined using a validation set or using cross-validation [15].

4.3 Naive Bayes: Naive Bayes is a successful classifier based upon the principle of Maximum A Posteriori (MAP).Given a problem with K classes {C1, ..., CK} with so called prior probabilities P (C1), ... , P(CK),

can assign the class label c to an unknown example with features such features x=(x1,.....,xN) such that **c=argmaxcP(C=ckx1.....xN),** is choose the class with the maximum a posterior probability given the observed data. This posterior probability can be formulated, that is choosing the class with the maximum a posterior probability given the observed data. This posterior probability observed data. This posterior probability can be formulated,

P(C=ckx1,..,xN)=p(C=c)P(x1,....,xNkC=c)P(x1.....xN)

As the denominator is the same for all classes, it can be dropped from the comparison. Now, we should compute the so-called class conditional probabilities of the features given the accessible classes. This may be quite difficult taking into account the dependencies between features. This approach is to assume conditional independence i.e. x1,. , xN are independent. This simplifies numerator as P(C = c) P (x1kC = c). P (xNkC = c), and then choosing the class c that maximizes this value over all the classes c = 1. . . K [12].

4.4 Support Vector Machine: Support Vector Machines are among the most robust and successful classification algorithms. It is a new classification method for both linear and nonlinear data and uses a nonlinear mapping to transform the original training data into a higher dimension. Among the new dimension, it searches for the linear optimal separating hyper plane (i.e., "decision boundary"). With an appropriate nonlinear mapping to a adequately high dimension, data from two classes can be partitioned by a hyper plane [15].

4.5 Neural Network: The most popular neural network algorithm is back propagation which performs learning on a multilayer feed forward neural network. It contains an input layer, one or more hidden layers and an output layer. The basic unit in a neural network is a neuron or unit. The inputs to the network correspond to the attributes measured for each training tuple. The inputs fed simultaneously into the units making up the input layer. It will be weighted and fed simultaneously to a hidden layer. Number of hidden layers is arbitrary, although usually only one. Weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction [12]. As network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer.

4.6 Cluster Hierarchy Construction Algorithm (CHCA)

The algorithm takes a binary matrix (a table) as input. The rows of the table correspond to the objects we are clustering. Here we describing this algorithm with web pages, but the method is applicable to other domains as well. The columns correspond to the possible attributes that the objects may have (terms appearing on the web pages for this particular application). When row i has a value of 1 at column j, it means that the web page corresponding to i contains term j. From this table, which is a binary representation of the presence or absence of terms for each web page, we create a reduced table containing only rows with unique attribute patterns (i.e., duplicate rows are removed). Using the reduced table, we create a cluster hierarchy by examining each row, starting with those with the fewest terms (fewest number of 1's); these will become the most general clusters in our hierarchy.

The row becomes a new cluster in the hierarchy, and we determine where in the hierarchy the cluster belongs by checking if any of the clusters we have created so far could be parents of the new cluster. Potential parents of a cluster are those clusters which contain a subset of the terms of the child cluster. This comes from the notion of inheritance discussed above. If a cluster has no parent clusters, it becomes a base cluster. If it does have a parent or parents, it becomes a child cluster of those clusters which have the most terms in common with it. This process is repeated until all the rows in the reduced table have been examined or we create a user specified maximum number of clusters, at which point the initial cluster hierarchy has been created. The next step in the algorithm is to assign the web pages to clusters in the hierarchy. In genera there will be some similarity comparison between the terms of each web page (rows in the original table) and the terms associated with each cluster, to determine which cluster is most suitable for each web page. Once this has been accomplished, the web pages are clustered hierarchically. In the final step we remove any clusters with a number of web pages assigned to them that is below a user defined threshold and re-assign the web pages from those deleted clusters.

V. WEB CONTENT MINING TOOLS

Web content mining tools helps to download the essential information. Some of them are Screen-scraper, Automation Anywhere 6.1, Web Info Extractor, Mozenda and Web Content Extractor, Rapid Miner.

5.1 Rapid Miner: Rapid Miner is open source software and it is a tool for extracting information from web, Contains inbuilt algorithm. It can generate algorithm by itself.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

Features:

- Easy to use.
- Reduce time.
- Open source software.

5.2 Screen-scaper: Screen-scraping is a tool for extracting/ mining information from web sites [11]. It can be used for searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements. The programming languages like Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper. Features: Screen-scraper present a graphical interface allowing the user to designate URL's, data elements to be extracted and scripting logic to traverse pages and work with mined data. Once these items have been created, from external languages such as .NET, Java, PHP, and Active Server Pages, Screen-scraper can be invoked. This also facilitates scraping of information at periodic intervals. One of the most regular usages of this software and services is to mine data on products and download them to a spreadsheet. A classier example would be a meta-search engine where in a search query entered by a user is concurrently run on multiple web sites in real-time after which the results are displayed in a single interface.

5.3 Automation Anywhere: It is a Web data extraction tool
used for retrieving web data, screen scrape from Web pages or use it for Web mining [14].

Features:

- Unique SMART Automation Technology for fast automation of complex tasks.
- Record keyboard and mouse or use point and click wizards to create automated tasks quickly. Web record and Web data extraction.

5.4 Web Info Extractor: This is a tool for data mining, extracting Web content, and Web content analysis. It can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server.

Features:

- No need to learn boring and complex template rules and it is easy to define extract tool.
- Extract tabular as well as unstructured data to file or database.
- Monitor Web pages and extract new content when update.
- Can deal with text, image and other link file
- Can deal with Web page in all language
- Running multi-task at the same time
- Support recursive task definition.

5.5 Mozenda: This tool enables users to extract and manage Web data [15]. Users can setup agents that routinely extract, store, and publish data to multiple destinations. Once information is in Mozenda systems, users can format, repurpose, the data to be used in other applications or as intelligence. There are two parts of Mozenda's scraper tool:

i. **Mozenda Web Console:** It is a Web application that allows user to run agents, view & organize results, and export publish data extracted.

ii. **Agent Builder:** It is a Windows application used to build data extraction project.

Features:

- Easy to use.
- Platform independency. However, Mozenda Agent Builder only runs on Windows.
- Working place independence.

5.6 Web Content Extractor: It is a powerful and easy to use data extraction tool for Web scraping, data mining or data extraction from the Internet[13].It offers a friendly, wizard-driven interface that will help through the

process of building a data extraction pattern and creating crawling rules in a simple point-and-click manner. This tool allows users to extract data from various websites such as online stores, online actions, shopping sites, real estate sites, financial sites, business directories, etc. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT,
HTML, XML, SQL script, MySQL script and to any ODBC data source.

Features:

- Helps to extract/collect the market figures, product pricing data, or real estate data.
- Helps users to extract the information about books, including their titles, authors, descriptions, ISBNs, images, and prices, from online book sellers.
- Assists users in automate extraction of auction information from auction sites.
- Assists to Journalists extract news and articles from news sites.
- Helps people seeking a job extract job postings from online job websites. Find a new job faster and with minimum inconveniences
- Extract the online information about vacation and holiday places, including their names, addresses, descriptions, images, and prices, from web sites.

VI. CONCLUSION

The web continues to increase in size and complexity with time hence making it difficult to extract relevant information. The mining of web data still be present as a challenging research problem in the future. Because the web documents possess numerous file formats along with its knowledge discovery process. There are many concepts available in web content mining but this paper tried to expose the various web content mining strategy and explore some of the techniques. Then we described some tools web content mining..

VII. REFERENCES

[1] Herrouz, A., Khentout, C., Djoudi, M. Overview of Visualization Tools for Web Browser History Data, IJCSI International Journal of Computer Science Issues, Vol.9, Issue 6, No3, November 2012, pp. 92-98, (2012).
[2] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey,ACM SIGKDD Explorations Newsletter, June 2013, Volume 2 Issue 1.
[3] Han, J., Kamber, M. Kamber. “Data mining: concepts and techniques”. Morgan Kaufmann Publishers, 2014.
[4] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In Proc. of ACM- SIAM Symposium on Discrete Algorithms, pages 668–677,2014
[5] R.Cooley,. B.Mobasher,.; J.Srivastava,.; “Web mining: information and pattern discovery on the World Wide Web”. In Proceedings of Ninth IEEE International Conference. pp. 558 – 567, 3-8 Nov. 2015.
[6] Inamdar, S. A. and shinde, G. N. 2010. An Agent Based Intelligent Search Engine System for Web Mining. International Journal on Computer Science and Engineering, Vol. 02, No. 03.
[7] V. Bharanipriya & V. Kamakshi Prasad, Web Content Mining tools: A Comparative Study in International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215.
[8] Johnson, F., Gupta, S.K., Web Content Minings Techniques: A Survey, International Journal of Computer Application. Volume 47 – No.11, p44, June (2012).
[9] Ackland R, Gibson R, Lusoli W, Ward S. Engaging With the Public? Assessing the Online Presence and Communication Practices of the Nanotechnology Industry. Social Science Computer Review. 2010;28(4):443–465. doi: 10.1177/0894439310362735.
[10] AleEbrahim N, Fathian M. Summarising customer online reviews using a new text mining approach. International Journal of Business Information Systems. 2013;13(3):343–358. doi: 10.1504/IJBIS.2013.054468.
[11] Al-Hassan AA, Alshameri F, Sibley EH. A research case study: Difficulties and recommendations when using a textual data mining tool. Information & Management. 2013;50(7):540–552. doi: 10.1016/j.im.2013.05.010.
[12] Arora SK, Youtie J, Shapira P, Gao L, Ma TT. Entry strategies in an emerging technology: A pilot web-based study of graphene firms. Scientometrics. 2013;95(3):1189–1207. doi: 10.1007/s11192-013-0950-7.
[13] Batini C, Scannapieco M. Data quality: Concepts, methodologies and techniques. New York: Springer; 2006.

[14] Battistini A, Segoni S, Manzo G, Catani F, Casagli N. Web data mining for automatic inventory of geohazards at national scale. Applied Geography. 2013;43:147–158. doi: 10.1016/j.apgeog.2013.06.012.
[15] Etzkowitz H, Leydesdorff L. The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. Research Policy. 2000;29(2):109–123. doi: 10.1016/S0048-7333(99)00055-4.
[16] FAME. (2014). Detailed information on UK and Irish companies. Bureau van Dijk Electronic Publishing. Accessed via the University of Manchester Library.
[17] Fisher J, Craig A, Bentley J. Moving from a Web Presence to e-Commerce: The importance of a business—Web strategy for small-business owners. Electronic Markets. 2007;17(4):253–262. doi: 10.1080/10196780701635864.
[18] Hoekstra R, Ten Bosch O, Harteveld F. Automated data collection from web sources for official statistics: first experiences. Statistical Journal of the IAOS. 2012;28(3–4):99–111.
[19] Hyun Kim J. A hyperlink and semantic network analysis of the triple helix (University-Government-Industry): The interorganizational communication structure of nanotechnology. Journal of Computer-Mediated Communication. 2012;17(2):152–170. doi: 10.1111/j.1083-6101.2011.01564.x

CITE AN ARTICLE

It will get done by IJESRT Team