The International Journal Of Engineering And Science (IJES)
|| Volume || 3 || Issue || 4 || Pages || 21-27 || 2014 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Data Harvesting through Web Mining: A Survey

Prakul Gupta[1] Amit Sharma[2] Dr. Sunil Kr Singh[3]
[1, 2,] UG research Scholar, Department of CSE, Bharati Vidyapeeth College of Engineering, New Delhi, India
[3,] Professor, Department of CSE, Bharati Vidyapeeth College of Engineering, New Delhi, India

--ABSTRACT--

Web mining is one of the fastest growing technology. Experts believe that it will aid business houses in making better decisions. However, even after an extensive research in this field, there is an uncertainty regarding the usage of this term and it is often confused with Data mining. In this paper, we will be focussing on shedding light on such doubts and pointing out the similarities and differences between the two synonymously used words "Data Mining" and "Web Mining". We'll be addressing the sundry categories of Web Mining and its pros and cons. We'll also compare the latest tools available in the market which perform Web mining. Finally we'll delineate a strategy, for beginners, to develop a web mining tool which will help them in understanding the framework of Web mining.

Keywords: *Data mining, Web Content Mining, Web mining, Web Structure Mining, Web Usage Mining.*

Date of Submission: 18 April 2014 Date of Publication: 05 May 2014

I. INTRODUCTION

World Wide Web, being the largest repository of information, influences the everyday life of most of the people as we cannot only find the required information but also can easily share our knowledge and information with others. In just over two decades, the Web has become a virtual society (a fundamental research, marketing and communication vehicle) from a university curiosity. Due to this wide availability of huge amounts of information and the imminent need for turning it into meaningful information [1], we need to use Web mining.

Maintaining the quality and the accuracy of the data is a critical task and in an ever expanding universe of mammoth data there would be a high demand of dedicated data harvesting and managing tools to build an advanced analysis. To extract mammoth data from the internet, Web mining is certainly the technique to be worked upon. It has allured a great deal of heed in the information industry and in the society as a whole in recent years, due to the wide availability of bulk of data. This meaningful data gained can be used for applications ranging from Enterprise Applications like context-aware advertising, database building, business intelligence, competitive intelligence, comparison shopping etc. to Social Web Applications like Extracting data from a single and multiple Online Social Web platforms [12].

The key challenges [4,7] we can encounter in the design of a Web Data Harvesting system and its techniques can be summarized as follows:

- Need the help of human experts
- Large amount of data should be processed in relatively short time
- Solid privacy must be provided by applications dealing with human related data (eg: Applications in the field of Social Web)
- Large training set of Web pages which are manually labelled is required by approaches dependent on Machine Learning
- Evolution of web data source over time is also a hurdle for Web Data Extraction tools which needs to extract data routinely
- Maintaining the integrity of the specifications is another task due to explosive growth of internet in recent time which has rendered user to get effective information
- Time loss experienced by users
- Consumption of a lot of System Resources for Knowledge discovery
- Caching Schemes fails in certain Conditions.
- pre- fetching techniques result in over congestion of Network traffic

This paper is structured as follows. In section 2, we have given an overview of Data mining and Web Mining and pointed out differences between the two. In section 3, we provided a classification of Web Mining, and highlighted the pros and cons of it as well. In section 4, we provided a comparison between the latest available tools present in the market focusing on the basic features they possess. In section 5, we provided a strategy to help beginners for developing a web harvesting tool and finally we concluded in section 6.

II. DATA MINING AND WEB MINING

2.1 Data Mining

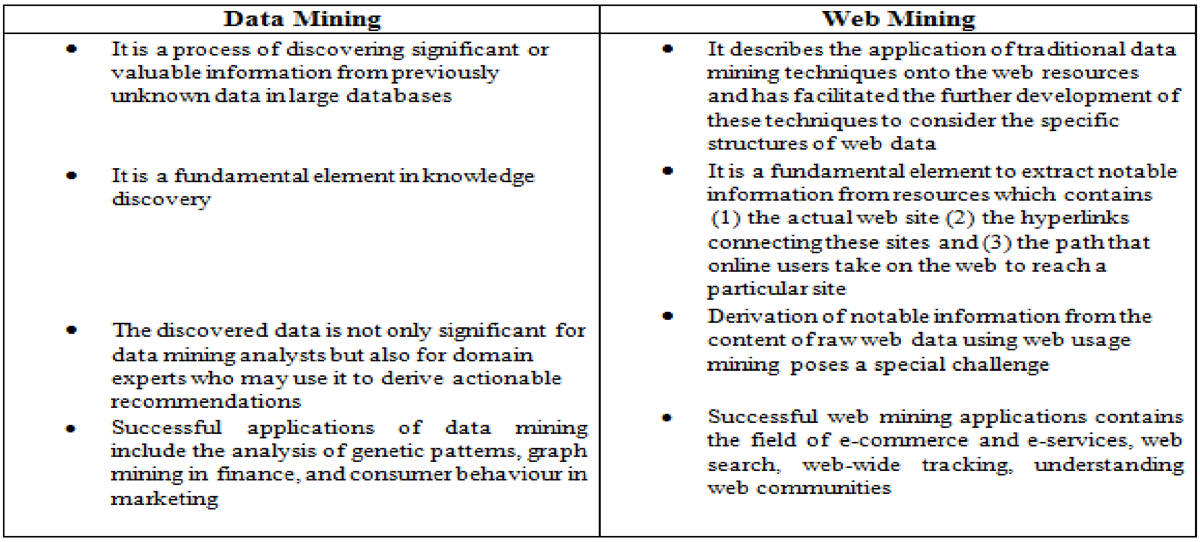
Mining is a vivid term distinguishing the process that finds a small set of precious nuggets from a great deal of raw material [3]. A good definition of Data Mining is that in Principles of Data Mining by David Hand : "Data mining is the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [11].

Data mining is also called knowledge discovery in databases (KKD) [3]. It is the process of identifying useful patterns and gathering of data from numerous data sources like disparate databases, texts, images, the Web, .etc into a single database from which it can be re-published in a unified manner [6]. The patterns must be valid, understandable and potentially useful. Data mining is a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval [2,8], induction, neural networks and visualization.

2.2 Web Mining

Web mining is an application of data mining techniques to explore exciting and potentially useful information from Web data. It is generally expected that either the hyperlink structure of the Web or the Web log data or both have been used in the mining process. We can also state Web mining as the discovery and analysis of meaningful information from Web pages and services. This describes the automatic search of information resources available on-line, i.e. Web content mining, and the discovery of user access patterns from Web servers, i.e., Web usage mining [3].

With the remarkable growth of the Web, there is an explosive increase in the amount of data and information published in various Web pages. The research in Web mining strives for new techniques to effectively extract and mine meaningful information from these Web pages [8]. Due to the diversity and structure of Web data, automated realization of targeted information is a tough task. The different Web mining techniques could be used to solve the information overload problems, like Finding relevant information or creating new knowledge out of the information available on the Web or Personalization of the information or Learning about consumers or individual users, directly or indirectly. By the direct approach we mean that the application of the Web mining techniques directly addresses the above problems. However, we do not claim that Web mining techniques are the only tools to solve those problems. Other techniques and works from different research areas, such as database (DB), information retrieval (IR), natural language processing (NLP), and the Web document community, could also be used [8]. In Fig.1 we have summarised the differences between the two.



Data Mining	Web Mining
• It is a process of discovering significant or valuable information from previously unknown data in large databases	• It describes the application of traditional data mining techniques onto the web resources and has facilitated the further development of these techniques to consider the specific structures of web data
• It is a fundamental element in knowledge discovery	• It is a fundamental element to extract notable information from resources which contains (1) the actual web site (2) the hyperlinks connecting these sites and (3) the path that online users take on the web to reach a particular site
• The discovered data is not only significant for data mining analysts but also for domain experts who may use it to derive actionable recommendations	• Derivation of notable information from the content of raw web data using web usage mining poses a special challenge
• Successful applications of data mining include the analysis of genetic patterns, graph mining in finance, and consumer behaviour in marketing	• Successful web mining applications contains the field of e-commerce and e-services, web search, web-wide tracking, understanding web communities

Fig.1: data mining vs. web mining

III. 3. Web Mining Taxonomy

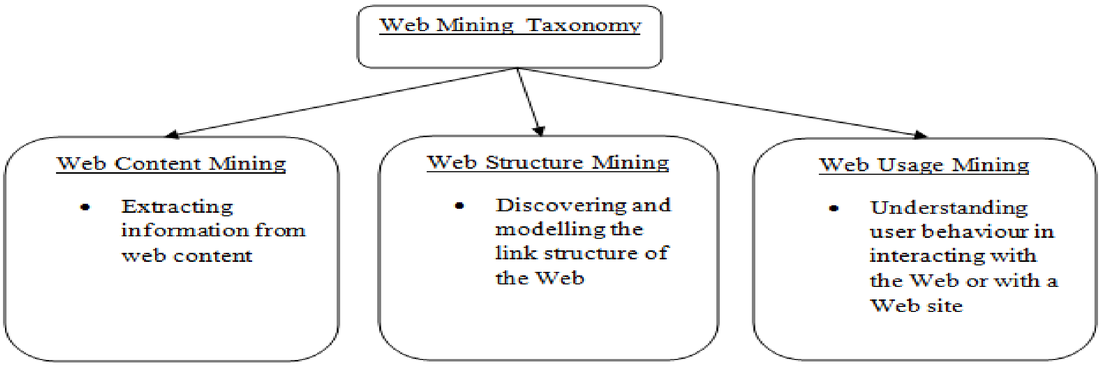

Web Mining Taxonomy
Web Content Mining
• Extracting information from web content
Web Structure Mining
• Discovering and modelling the link structure of the Web
Web Usage Mining
• Understanding user behaviour in interacting with the Web or with a Web site

Fig.2: web mining taxonomy

The web mining techniques can broadly be classified into three categories and are represented in Fig.2 and detailed differences amongst them in Fig.3, which are namely:

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

3.1 Web Content Mining

It deals with extracting valuable information from Web page contents which is well beyond using keywords in a search engine. In contrast to Web usage mining and Web structure mining, Web content mining mainly focuses on the Web Page content rather than the links. Web content is a very rich information resource consisting of many types of information, for example unstructured free text, images, audio, video and metadata as well as hyperlinks. The content of Web pages includes no machine readable semantic information. Search engines, subject directories, intelligent agents, cluster analysis, and portals are employed to find what a user might be looking for. It has been suggested that users should be able to pose more sophisticated queries than just specifying the keywords.

3.2 Web Structure Mining

It deals with discovering and modelling the link structure of the Web. Work has been carried out to model the Web based on the topology of the hyperlinks. This can assist in discovering resemblance between sites or in exploring important sites for a specific topic or discipline or in exploring Web communities.

3.3 Web Usage Mining

It deals with understanding user behaviour in interacting with the Web or with a Web site. One of the objective is to obtain information that may assist Web site reorganization or assist site adaption to better suit the user [11]. The mined data often contains data logs of users' interactions with Web. The logs include the Web server logs, proxy server logs, and browser logs. The logs include information about the referring pages, user identification, time a user spends at a site and the sequence of pages visited. Information is also collected via cookie files. While Web structure mining shows that page X has a link to page Y, Web usage mining shows who or how many people took that link, which site they came from and where they went when they left page Y.

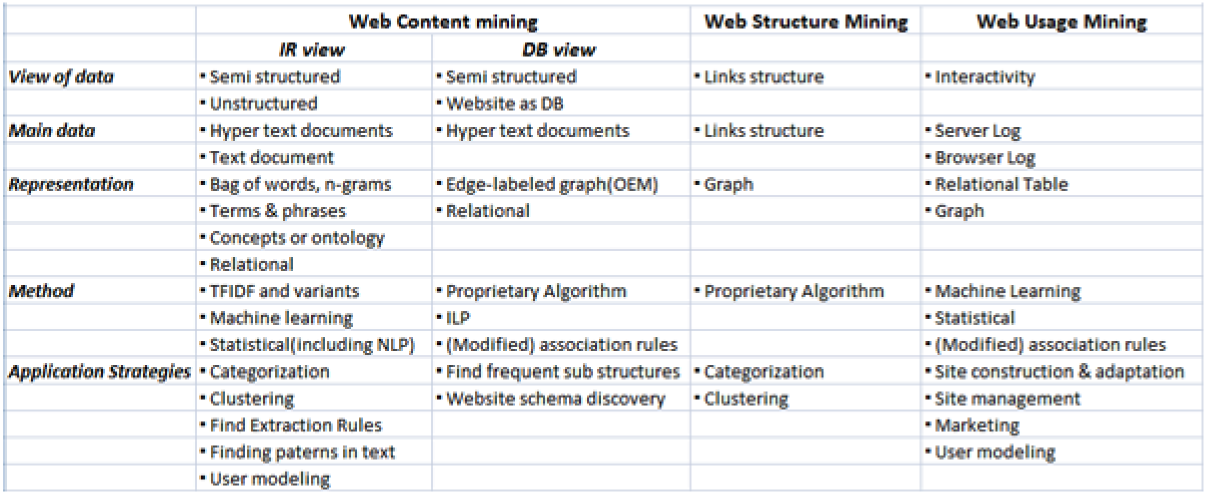
3.4 Pros and cons of Web Mining

Pros

- Enables e-commerce for personalized marketing which results in higher trade volumes
- Classifies threats and fight terrorism
- Identifies criminal activities
- Establishes better customer relationship by responding to customer needs faster and fulfilling their requirements efficiently
- Profitability can be increased by target pricing which would be based on different profiles created after minin**g**

Cons

- Invasion of privacy: Issues related to data of personal nature
- De-individualizing users: Harvesting tools judge users by their mouse clicks. De-individualization means a tendency of judging and treating people on the basis of group characteristics instead of on their own individual characteristics and merits [10]
- Infringement of User's interests: Companies can extract the data for one particular purpose but they might use the harvested data for a totally different purpose
- Trading personal data: Since, there is no law which can prevent website owners to trade data and hence, there is a growing trend of selling personal data obtained from their sites as a commodity

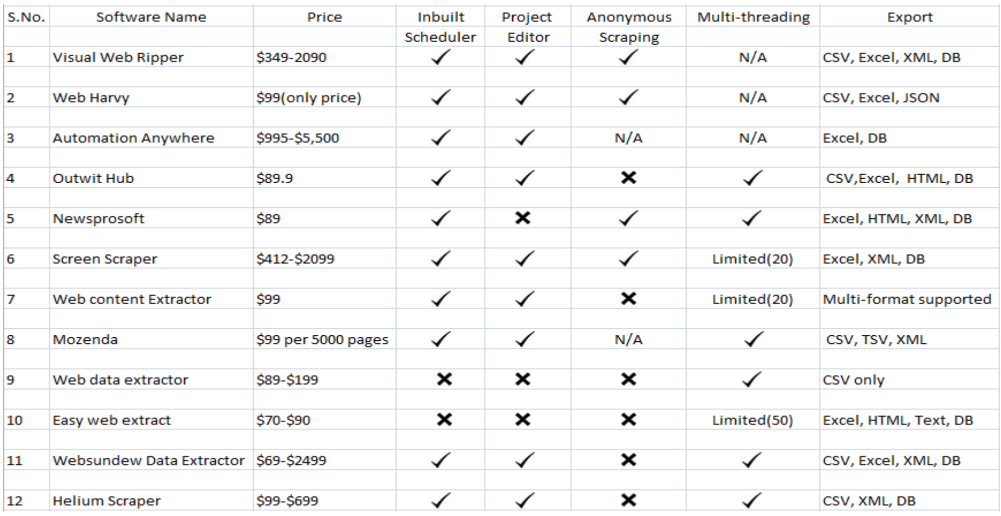


	Web Content mining		Web Structure Mining	Web Usage Mining
	IR view	*DB view*		
View of data	• Semi structured	• Semi structured	• Links structure	• Interactivity
	• Unstructured	• Website as DB		
Main data	• Hyper text documents	• Hyper text documents	• Links structure	• Server Log
	• Text document			• Browser Log
Representation	• Bag of words, n-grams	• Edge-labeled graph(OEM)	• Graph	• Relational Table
	• Terms & phrases	• Relational		• Graph
	• Concepts or ontology			
	• Relational			
Method	• TFIDF and variants	• Proprietary Algorithm	• Proprietary Algorithm	• Machine Learning
	• Machine learning	• ILP		• Statistical
	• Statistical(including NLP)	• (Modified) association rules		• (Modified) association rules
Application Strategies	• Categorization	• Find frequent sub structures	• Categorization	• Site construction & adaptation
	• Clustering	• Website schema discovery	• Clustering	• Site management
	• Find Extraction Rules			• Marketing
	• Finding paterns in text			• User modeling
	• User modeling			

Fig.3: comparison of web mining categories

IV. COMPARISON OF DIFFERENT HARVESTERS

In the market there are various web harvesting software available and we did the comparison amongst 12 such tools developed by various firms on the basis of their cost and features. The detailed comparison can be seen from the Fig.4. The features taken into account for comparison are availability of inbuilt scheduler, project editor, anonymous scraping, multi-threading and different file formats used to export data. This comparison will help any beginner to look forward to different features which can be included in the build-up of a web harvester.



S.No.	Software Name	Price	Inbuilt Scheduler	Project Editor	Anonymous Scraping	Multi-threading	Export
1	Visual Web Ripper	$349-2090	✓	✓	✓	N/A	CSV, Excel, XML, DB
2	Web Harvy	$99(only price)	✓	✓	✓	N/A	CSV, Excel, JSON
3	Automation Anywhere	$995-$5,500	✓	✓	N/A	N/A	Excel, DB
4	Outwit Hub	$89.9	✓	✓	✗	✓	CSV,Excel, HTML, DB
5	Newsprosoft	$89	✓	✗	✓	✓	Excel, HTML, XML, DB
6	Screen Scraper	$412-$2099	✓	✓	✓	Limited(20)	Excel, XML, DB
7	Web content Extractor	$99	✓	✓	✗	Limited(20)	Multi-format supported
8	Mozenda	$99 per 5000 pages	✓	✓	N/A	✓	CSV, TSV, XML
9	Web data extractor	$89-$199	✗	✗	✗	✓	CSV only
10	Easy web extract	$70-$90	✗	✗	✗	Limited(50)	Excel, HTML, Text, DB
11	Websundew Data Extractor	$69-$2499	✓	✓	✗	✓	CSV, Excel, XML, DB
12	Helium Scraper	$99-$699	✓	✓	✗	✓	CSV, XML, DB

Fig.4: comparison of web harvesting tools

V. STRATEGY

In this part, we'll try to give an overview of what a Web Harvester can do and how it will function. The description is such that it'll help any beginner [7] to start with their Web Harvester. The algorithm for our proposed Web Harvester will be presented in the next paper. The preliminary step to design any software [5] is to firstly create the overview of the system, Fig.5 represents Level-0 DFD for a Web Harvesting software.

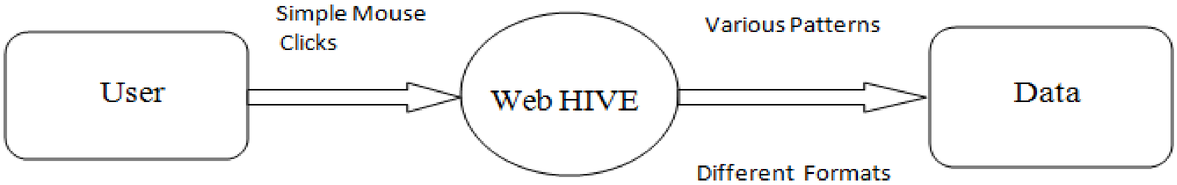


Fig.5: level-0 DFD

Once the overview of the system is understood, one can focus on how to elaborate it. The software (Web HIVE) is basically divided and implemented , as depicted in Fig.6, according to the following modules:

- GUI implementation
- Basic Scraping/Harvesting
- Addition of extra features
- Security Implementation

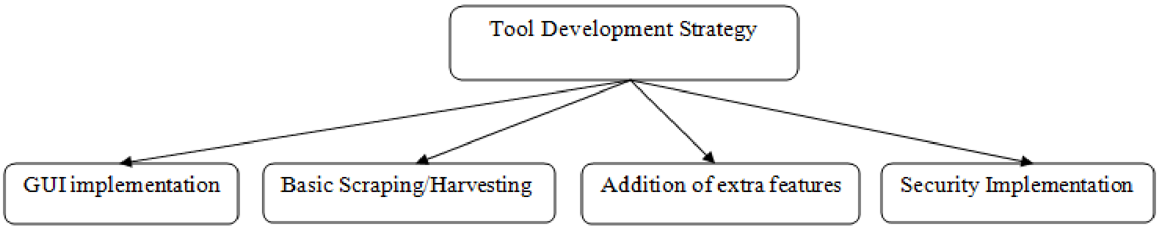

Tool Development Strategy
GUI implementation
Basic Scraping/Harvesting
Addition of extra features
Security Implementation

Fig.6: tool development strategy

5.1 GUI Implementation

A very efficient and user friendly GUI should be implemented in a software. It should be easy to interact and provide many features to its users. The Web Harvesting software must be a point and click web harvester (visual web harvester) which lets its user scrape data from the web with ease. It should provide an inbuilt web browser to navigate to any webpage. It could be configured to extract data from websites with a mere mouse click thereby minimizing the use of keyboard and other input devices. The user just needs to select the data to be extracted by pointing the mouse and clicking on it.

5.2 Basic Harvesting

For proper functioning of the tool, a lot of websites of different categories were studied in order to understand the pattern of data that is present on the web. Based on the study, different patterns, which would be able to fetch data not only from a few selected websites but from a large pool of various categories of websites, were used to scrape the data. So the algorithm used to harvest data should revolve around some active element that the user will click and based on which similar data from the current page and other pages can be extracted [9]. The main task in this module was to describe how patterns can be used to extract meaningful data. To harvest data the user need to be in Config mode which provides the user to highlight the data items which are to be captured. This mode also displays a Capture Window when data elements, which are to be harvested ,are clicked. Now the user could select what to extract by choosing the appropriate options, like link, image, url, html code, regular expressions etc, available. These are some basic harvesting strategies.

5.3 Addition of extra features

Apart from providing basic scraping facilities, the main task is to scrape data across multiple pages as data displayed by websites spans over multiple pages and hence, an extra feature for this should be provided. Also a facility to range the number of pages from where data has to be scraped should be provided. In addition to this, if a user wishes to follow certain link, i.e. if the user wants to scrap data from the links which are similar to the selected followed link, and then scrape data, such feature should also be provided. The scrap data can be saved for scheduling purpose. The harvested data can be exported in .csv and .txt format. One more convenient feature can be added which allows the user to enter the changing field in a url manually, so that the data can be extracted until the number of pages specified by the user is reached. Also, if the user has a list of links (all belonging to the same domain, which shares the same page layout) , then the user can be provided with another special feature to include all those links using a single configuration.

5.4 Security Implementation

In order to maintain a level of anonymity while extracting data from websites, there should be an option which allow the user to pause the miner periodically while harvesting data. This prevents the harvester from making data requests continuously(long-time) to the website, resulting in minimization of the chances of the user's IP from being blocked by the website. Also, an option should be there to prevent the data loss. Using this option the miner could automatically export the extracted data to a file on the user's computer periodically. This option is an optimum way to prevent loss of data due to unexpected problems during over long mining sessions.

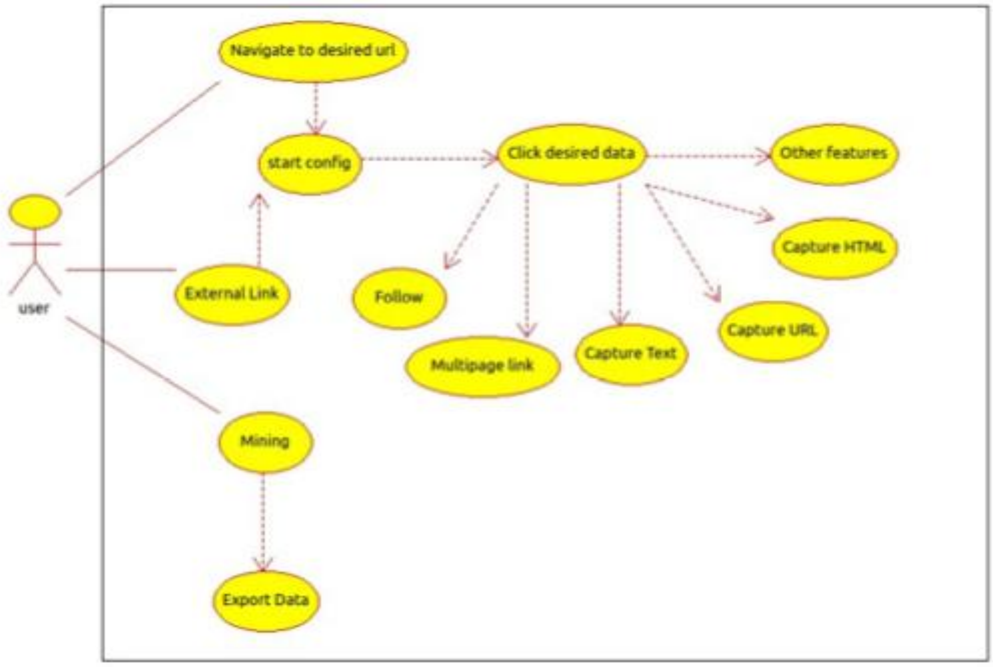


Fig.7: use case diagram

VI. CONCLUSION

As the usage of Web continues to grow, so does the opportunity to analyze data on the Web and to extract all possible useful knowledge from it is getting threefold. We hope that this paper proves to be a starting point for profound discussions not only for beginners but also for developers and researchers.

We have tried to provide a clear cut view of different aspects of Web Mining and pointed out and cleared basic confusions regarding the usage of the term Web mining and Data Mining. In order to reach out to a larger audience, we have tried to explain Web Mining in a simple way. The paper also provides basic information regarding Web Mining through various diagrams and tables and also the potential , this technology has in future, which is essential for beginners to understand its framework.

We did a comparison between different tools available in the market and provided a strategy for beginners to develop a tool for harvesting data through Web. The Use Case Diagram is depicted in Fig.7. The optimized algorithm for the same will be provided by us in our next paper.

REFERENCES

[1] M. Berthold and D.J. Hand, *Intelligent Data Analysis: An Introduction* (Springer- Verlag New York, Inc., Secaucus, NJ, USA, 1999).

[2] Sarawagi. Information extraction, Found, *Trends databases, 1*(3):261–377, 2008 DOI: 10.1561/1500000003.

[3] Jiawei Han and Micheline Kamber, *Data mining: concepts and techniques* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000).

[4] Ferrara, E., de Meo, P., Fiumara, G., and Baumgartner, R. 2012. Web Data Extraction, Applications and Techniques: A Survey, *arXiv*:1207.0246v2 [cs.IR] 7 Mar 2013.

[5] Laender, A. H. F., Ribeiro-Neto, B. A. and da Silva, A. S. 2001. DEByE – Data Extraction by *Example. Data and Knowledge Engineering*, 40(2), 121-154.

[6] White Paper on *Data Harvesting On-time, accurate and easy delivery of data* by Snowflake Software team.

[7] Mrs.Bhanu Bhardwaj, Asst Proff DCE G.Noida, Extracting Data through Webmining, *International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 3*, May – 2012, ISSN: 2278-0181

[8] R. Kosala, H. Blockeel "Web mining research: A survey," *ACM SIGKDD Explorations, Vol. 2 No. 1*, pp. 1-15, June 2000.

[9] Yanhong Zhai and Bing Liu. Structured Data Extraction from theWeb Based on Partial Tree Alignment. *IEEE Transactions on Knowledge and Data Engineering, Vol. 18*, No. 12,Dec 2006.

[10] Lita Van Wel and Lambèr Royakkers, Ethical issues in web data mining. *Ethics and Information Technology, v.6* n.2, p.129-140, 2004.

[11] Gupta, G.K.: *Introduction to Data Mining with Case Studies* (Phi Learning, 1st edn. 2008).

[12] S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Crawling facebook for social network analysis purposes. *In Proc. International Conference on Web Intelligence, Mining and Semantics*, page 52, Sogndal, Norway, 2011. ACM.