Opinion Mining to Strengthen Teaching Learning Process

Sartaj Ahmad
KIET, Ghaziabad,
Affiliated to AKTU,
Lucknow(UP), India
sartajahmad2u@gmail.com

Ashutosh Gupta
School of Sciences,
U.P.Rajarshi Tandon Open
University, Allahabad(UP), India
ashutosh3333@gmail.com

Neeraj Kumar Gupta
KIET, Ghaziabad,
Affiliated to AKTU,
Lucknow(UP), India
neeraj.gupta@gmail.com

Abstract: *Online exchange of information is very common now a day. It offers opportunity & challenges to extract knowledge for individual use. There are different techniques to achieve this purpose. Therefore in this paper details about web mining especially about web content mining is presented. A comprehensive review about web mining techniques and their use in getting structured, semi structured and unstructured data from the ocean of information (World Wide Web) is also presented. Further opinion mining is explored as an application of web content mining for academic data to reinforce existing academic teaching learning process.*

Keywords: *Web Mining, Semi Structured Data, Structured Data, Unstructured Data, Web Crawler*

I. INTRODUCTION

Size of W3 (Information Space) is increasing very fast day by day because documents are connected through links. Reason of this is people awareness and dependency on the internet for the various purposes like business, shopping, education, banking, health, blogging, feedbacks etc. But this is also facts that major part of such information space is in unstructured in nature means in text form. Therefore major challenge is how to extract and use relevant information from such big information space.

This survey is conducted for study and finding problems to access relevant information from such growing rate of data which was 16 millions in 1995 and 3270 millions in June, 2015 on the web [35]. In this paper firstly a survey is provided on web mining then emphasis is given on web content mining. At the end unstructured content mining is used as an application for teaching learning process. Finally conclusion is given.

II. CLASSIFICATION OF WEB MINING

It is automatic covering and extracting patterns from the web through data mining techniques. It can be classified into three categories [1, 4, 5] as follows.

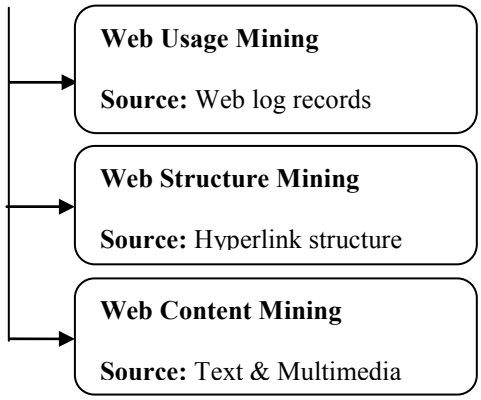

Fig.1: Web mining categories

Brief introduction about each category is given in [1]. But here focus is given to Web Content Mining.

2.1 Web Content Mining:

This includes knowledge discovery from web page's content such as text, images, videos, audios [4, 6, 7].

2.1.1 Problems in web content mining [12, 14]:

Extraction of data/information:

Most of the data on the web is unstructured (does not follow any data model). So it is challenge how to extract hidden information from such data. Software modules called wrapper are used to extract data from web sites. But problem is that wrapper is based on manual techniques and writing such wrapper is difficult process. An unstructured data management system to extract structures (e.g., person names, locations) from the raw text data is proposed to improve the extraction and integration methods, the quality of the resulting database, and the user services.

978-1-4673-9080-4/16/$31.00 ©2016 IEEE

Opinion extraction from online sources [27, 28, 30]:

Feedbacks regarding sold products or service are asked by the different sites. These feedbacks help the seller and new customer both. To extract and summarize opinions from these feedbacks is a challenge.

Knowledge creation [18, 19, 20]:

Picking willing information from web and convert into knowledge for the presentation is also a problem.

Detecting noise from different segments:

Web page content has different sub segments like advertisements, navigation links, copyright notices etc. So segmenting Web page automatically to extract the main content of the pages is interesting problem [13].

III. DATA ON WEB PAGE & ITS EXTRACTION

Data available on web is classified as structured data which is available as table, list and tree, semi structured data which does not have a predefined structure. It is in the form of Hierarchical structured like HTML and XML code and third one is unstructured data where data is in the form of text.

3.1 TECHNIQUES FOR EXTRACTING STRUCTURED DATA

Some of the techniques for extracting structured data are as follows.

3.1.1 Web Crawler [12, 14, 15, 16, 17]**:**

It is the very important source of information retrieval which traverses the Web and downloads web documents that suit the user's need. It is used by the search engine and other users to ensure that their database is up-to-date.

3.1.2 Wrapper Generation [12, 18, 26]:

It is a program that extracts content of a particular information source (Web pages) and translates it to the mediator into structured form. In this wrappers are built around individual information sources, which provide translation between the mediator query language and the individual source. This can be understood by the following figure.

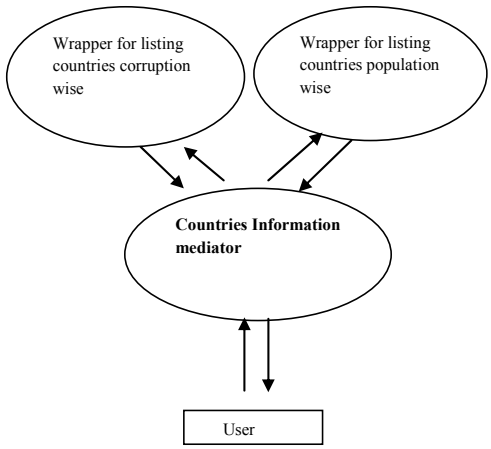

Fig.2. Use of Wrapper

3.2 TECHNIQUES FOR EXTRACTING SEMI STRUCTURED DATA

Some of the techniques are as described below.

3.2.1 Object Exchange Model and Schema Knowledge Mining:

In this method data is represented as a rooted tree (acyclic graph).Where objects and labels are represented as vertices and edges respectively. In this Schema knowledge mining is used. It helps user to understand the information structure of the web more deeply and thoroughly [12, 21, 22].

3.2.2. Top down Extraction:

In this approach distinguished objects and their attributes [12, 13] are extracted and inserted into a table that can be queried further.

3.3 TECHNIQUES FOR EXTRACTING UNSTRUCTURED DATA

Unstructured Data Mining covers many areas to explore for the problem identification as shown in the following figure.

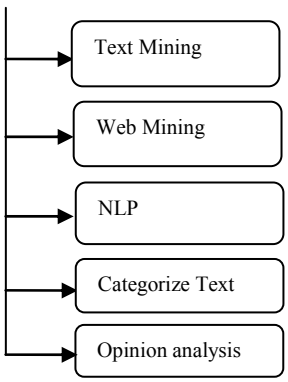

Fig.3. Unstructured Data Mining Components

Following combined techniques are used in such type of Data Mining

3.3.1 Bayesian Networks:

It is used to extract information from unstructured, ungrammatical and incoherent data sources from the web. It is combined with ontologies, machine learning and probabilistic reasoning techniques to extract structured and semi-structured information.

3.3.2 Text Analytics:

It covers different set of machine learning techniques like linguistic, statistical, that model and structure the information content of textual sources for business intelligence and research. Its subtasks include the followings.

- Identifying a set textual material from the Web.
- Named Entity Recognition means to identify named text features like people, organizations, place names, stock ticker symbols, certain abbreviations, and so on.
- Recognition of Features of entities such as telephone numbers, e-mail addresses, gender etc.
- Co reference: Identification of noun phrases, adjective and adverb phrases that refer to the same object.
- Relationship of information available in text.

Some Commercial text analytics software:
There are many text analytics research, commercial and open source software [14] options. Few of them are MeshLabs, Clarabridge, AeroText, SAS, Sysomos etc.

3.3.3 Natural Language Processing:

It is used to find some hidden patterns from the text. Unstructured information is passed to a parser to tokenize information. After this some algorithm is applied to find tokens like (Noun, Adjective, and Verb) for further use. This processing is explained as an application in the next section.

IV. OPINION MINING AS AN APPLICATION OF UNSTRUCTURED DATA MINING

Graduate students are asked to give online reviews in their style in the form of plain text about a particular subject named Computer programming in 'C'. They write freely and send it back. Number of reviews increases as time passes. And it becomes difficult to analyze such a big number of reviews manually. Therefore a mechanism is discussed to analyze these reviews and provide a summary to the teacher and new students. This summary will help the teacher to improve their teaching methodology and students will also increase their effort. Following steps are used to find summary about the subject.

1. Find subjective reviews means reviews having noun and adjective from collected reviews manually.

For Example:

Pointer is difficult for beginners.

Array is easy to understand.

I like loop very much.

All is easy but pointer is difficult.

I enjoyed c languge.

File handling is good concept.

Loop,Array and pointer is not easy.

Array is good and easy to understand.

I like array but pointer not easy.

c is popular programming language.

2. Pass filtered reviews to Stanford parser to collect sentences in the form of parts of speech tags as follows.

Pointer/NNP is/VBZ difficult/JJ for/IN beginners/NNS ./.

Array/NNP is/VBZ easy/JJ to/TO understand/VB ./.

I/PRP like/VBP loop/NN very/RB much/RB ./.

All/DT is/VBZ easy/JJ but/CC pointer/NN is/VBZ difficult/JJ ./.

I/PRP enjoyed/VBD c/SYM languge/JJ ./.

File/JJ handling/NN is/VBZ good/JJ concept/NN ./.

Loop/NNP ,/, Array/NNP and/CC pointer/NNP is/VBZ not/RB easy/JJ ./.

Array/NNP is/VBZ good/JJ and/CC easy/JJ to/TO understand/VB ./.

I/PRP like/VBP array/NN but/CC pointer/VBP not/RB easy/JJ ./.

c/SYM is/VBZ popular/JJ programming/NN language/NN ./.

3. Now applying selection rules collect the topic as Noun (NN/NNP tag) and

Adjective (JJ tag) and check the orientation (positive/negative) of each sentence as per the list for positive and negative list of words.

e.g. if NN is pointer and JJ is difficult then orientation is negative

if NN is array and JJ is easy then orientation is positive

4. Finally topic wise (Table No. 1) summary is provided to know the understanding of the students about the subject.

Table 1: Summary

S.No.	Topic	Positive	Negative
1	Looping	100	10
2	Array	210	20
3	Structure	50	20
4	Union	50	10
5	Pointer	100	200
6	File Handling	100	20

This summary can be represented as the following figure.

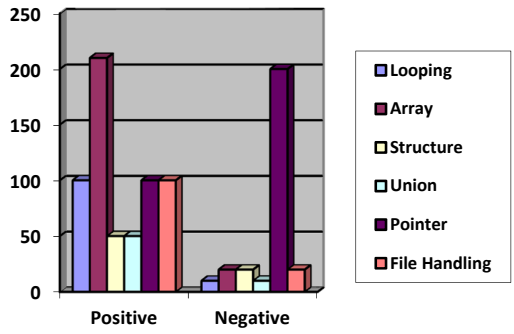

Fig.4. Topic vs. No. of opinions

***Summary Analysis*:**

In the graph positive side is heavy that shows overall response about subject is positive. But topic wise pointer's negative side is bigger one that shows pointer is less understandable by the students. It may be due to not good teaching or less focus of students. So there is need to give more focus on Pointer. Test 1 of 10 marks about pointer is also organized to confirm this summary. Following result is found for 1000 students.

Table 2: Result of Test 1

Marks >= 8	8 > Marks >= 5	Marks < 5
250	350	400

Result Interpretation: This result also supports feedback which is obtained earlier. So to come over this some preventive action is taken place.

***Corrective / Prevention Action*:**

Arrange remedial classes and give more exercises on this topic. We also revise our teaching methodology and study material to improve students understanding.

Action:

We organize test 2 to confirm our corrective/preventive action and get the following result as mentioned in table 3.

Table 3: Result of Test 2

Marks >= 8	8 > Marks >= 5	Marks < 5
300	600	100

Result comparison of Test 1 & Test 2:

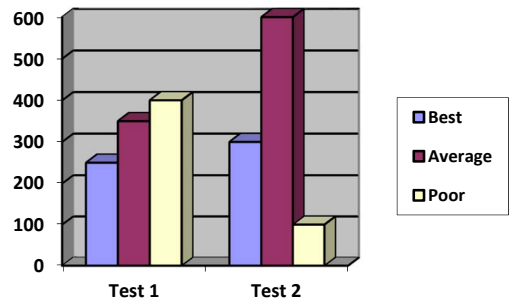

Fig.5. Test 1 vs. Test 2

Analysis:

Above graph shows that we reduce number of poor students as far as pointer is concerned.

V. FUTURE PERSPECTIVES

Web and its usage growing day by day, it causes to generate ever more content and usage data. This will keep increasing value of Web mining. Here are some research's future guidelines [14, 25] that may be pursued to ensure development in the Web mining technologies.

Semantic Web Mining [33]:

It is the second-generation WWW, enriched by machine processable information which supports the user in his tasks. Researchers are working on improving the results of Web Mining by exploiting semantic structures in the Web.

Fraud and threat analysis [34]:

There is significant increase in attempted frauds online for example Unauthorized use of credit cards after hacking into database for blackmail purposes. eBay like site also face Auction frauds. Web mining is the perfect analysis technique for detecting and preventing them. Research should focus on developing techniques to recognize, characterize and analyze such frauds.

Customer reviews Analysis [18, 19]:

This is also popular research area where customer's reviews about a product or service are analyzed. This analysis gives some important information to the customer and merchant both. Customer knows about the popularity and quality of product and merchant also knows about demand and weaknesses of the product.

Web services optimization:

Web is growing day by day therefore there is a need to make its services robust, scalable, efficient, etc. Web mining can be applied to better understand the behavior of these services, and the knowledge extracted can be useful for various kinds of optimizations [27]. Therefore research is also required to develop Web mining techniques to improve various other aspects of Web services.

VI. CONCLUSION

Survey is always important for new researcher in a particular area. This paper presents web mining, its categories, web data types, extraction techniques and finally opinion mining as an application to understand meaning of unstructured data mining. Future guidance is also given to the researcher for their work in web mining. Many challenging research problems due to different nature of data mainly unstructured are discussed in this paper. In future work one can plan and develop technique to automate unstructured data as per his/her requirements for later use.

REFERENCES

[1] Ahmad, Sartaj and Khan, M.Z. (2010), "Improving Effectiveness of Online Teaching (An Application of Web Usage Mining)", International Journal of Computer Applications, Vol. 8(13), pp. 21–24, October 2010.

[2] Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J. (2002), "Privacy Preserving Mining of Association Rules, in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", Edmonton, Alberta, Canada, pp. 217–228.

[3] Iyengar, V.S. (2002), "Transforming Data to Satisfy Privacy Constraints, in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining, Edmonton, Alberta, Canada, pp. 279–288.

[4] Kosala, Raymond and Blockeel, Hendrik (2000), "Web Mining Research: A Survey, ACM SIGKDD Explorations", Vol. 2(1), pp. 1–15, Jul. 2000.

[5] Srivastava, Jaideep, Cooley, Robert, Deshpande, Mukund and Pag-Ning, Tan (2000), "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, ACM SIGKDD Explorations", Vol. 1(2), pp. 12–23, Jan. 2000.

[6] Barfourosh, A.A., Nezhad, H.R. Motahary, Anderson, M.L. and Perlis, D. (2001), "Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition Technical Report".

[7] Wang, Jicheng, Huang, Yuan, Wu, Gangshan and Zhang, Fuyan (1999), "Web Mining: Knowledge Discovery on the Web", in Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC '99), Tokyo, Japan, Vol. 2, pp. 137–141, Oct. 12-15, 1999.

[8] Getoor, L. (2003), "Link Mining: A New Data Mining Challenge", ACM SIGKDD Explorations, Vol. 5(1), pp. 84–89.

[9] Page, L., Brin, S., Motwani, R. and Winograd, T. (1998), "The Pagerank Citation Ranking: Bring Order to the Web", Technical Report, Stanford University.

[10] Kleinberg, J.M. (1998), "Authoritative Sources in a Hyperlinked Environment", in Proceedings of ACM-SIAM Symposium on Discrete Algorithms, pp. 668–677.

[11] Brin, S. and Page, L. (1998), "The Anatomy of a Large-scale Hypertextual Web Search Engine", Elsevier Science Journal on Computer Networks and ISDN Systems, Vol. 30, pp. 107–117.

[12] Pol, K., Patil, N., Patankar, Shreya and Das, Chhaya (2008), "A Survey on Web Content Mining and Extraction of Structured and Semi Structured Data", in Proceedings of IEEE First International Conference on Emerging Trends in Engineering and Technology (ICETET'08), Nagpur, Maharashtra, India, pp. 543–546, Jul. 16-18, 2008.

[13] Callan, J.P. (1994), "Passage-Level Evidence in Document Retrieval", in Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94), Dublin, Ireland, pp. 302–310.

[14] Web Crawler [Online], Available: http://en.wikipedia.org/wiki/Web-crawler

[15] Ester, Martin, Kriegel, Hans-Peter and Schubert, Mattis (2004), "Accurate and Efficient Crawling for Relevant Websites", in Proceedings of the 30th VLDB Conference, Toronto, Canada, pp. 396–407.

[16] Tripathy, A. and Patra, P.K. (2008), "A Web Mining Architectural Model of Distributed Crawler for Internet Searches using Page Rank Algorithm", in Proceedings of IEEE Asia-Pacific Services Computing Conference (APSCC'08), Yilan, pp. 513–518, Dec. 9-12, 2008.

[17] Gupta, P. and Johari, K. (2009), "Implementation of Web Crawler", in Proceedings of IEEE 2nd International Conference on Emerging Trends in Engineering and Technology (ICETET'09), Nagpur, Maharashtra, India, pp. 838–843, Dec. 16–19, 2009.

[18] Hu, M. and Liu, B. (2004), "Mining and Summarizing Customer Reviews", in Proceeding of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), USA, pp. 168–177.

[19] Popescu, A.M. and Etzioni, O. (2005), "Extracting Product Features and Opinions from Reviews", in Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP'05), Canada, pp. 339–346.

[20] Lei, Xiang and Xin, Meng (2009), "A Data Mining Approach to Topic-Specific Web Resource Discovery", in Proceedings of IEEE 2nd International Conference on Intelligent Computation Technology and Automation (ICICTA'09), Zhangjiajie, China, Vol. 2, pp. 595–599, Oct. 10-11, 2009.

[21] Lv, Cheng, Wei, Chu-yuan and Hao, Ying (2009), "Schema Discovery of Semi-structured Hierarchical Data based on OEM Model and Hierarchical Transactional Database", in Proceedings of IEEE International Conference on Management of e-Commerce and e-Government (ICMECG'09), Nanchang, pp. 172–175, Sep. 16-19, 2009.

[22] Chen, Enhog and Wang, Xufa (1999), "Semi Structure Data Extraction and Schema Knowledge Mining, in Proceedings of 25th IEEE EUROMICRO Conference, Milan, Italy, Vol. 2, pp. 310–317, Sep. 08-10, 1999.

[23] Ribeiro-Neto, B., Laender, A.H.F. and Silva, A.S. da (1999), "Topdown Extraction of Semi-structured Data", in Proceedings of IEEE's the String Processing and Information Retrieval Symposium (SPIRE '99) & International Workshop on Groupware, Cancun, Maxico, pp. 176–183, Sep. 22-24, 1999.

[24] Li, Feifei, Liu, Zehua, Huang, Yangfreng, Ng, Wee-Keong (2001), "An Information Concierge for the Web", in Proceedings of IEEE 12th International Workshop on Database and Expert Systems Applications, Munich, Germany, pp. 672–676, Sep. 03-07, 2001.

[25] Ramakrishna, M.T., Gowdar, L.K., Havanur, M.S., Swamy, B.P.M., "Web Mining: Key Accomplishments, Applications and Future Directions", in Proceedings of IEEE International Conference on Data Storage and Data Engineering (DSDE), Bangalore, India, pp. 187–191, Feb. 09-10.

[26] Ashish, Naveen, and Craig, Knoblock (1997), "Semi-automatic Wrapper Generation for Internet Information Sources", Cooperative Information Systems, COOPIS'97, Proceedings of the Second IFCIS International Conference on IEEE.

[27] Walaa, Medhat, Ahmad, Hassan and Hoda, Korashy (2014), "Sentiment Analysis Algorithms and Applications: A Survey", Ain Shams Engineering Journal, Production and hosting by Elsevier.

[28] Cambria, Erik, Xia, Yunqing (2013), Intelligent Systems, IEEE, Vol. 28(2), pp. 15–21.

[29] Chitraa, V. and Davamani, Antony Seldoss (2010), "A Survey on Preprocessing Methods for Web Usage Data", IJCSIS, Vol. 7(3).

[30] Xu, Kaiquan, Liao, Stephen Shaoyi, Li, Jiexun, Song, Yuxia (2011), "Mining Comparative Opinions from Customer Reviews for Competitive Intelligence", Decision Support Systems 50, Elsevier, pp. 743–754.

[31] Gehrke, Nick, Werner, Michael, Dipl.-Wirt.-Inf. (2015), Process Mining, Article, Last seen on 25/10/2015, From http://www.wisu.de.

[32] Booth, Danielle and Jansen, Bernard J. (2015), "Chapter VIII, A Review of Methodologies for Analyzing Websites", Last seen on 25/10/2015, From http://faculty.ist.psu.edu/jjansen/academic/jansen_website_analysis.pdf

[33] Quboa, Qudamah K., Saraee, Mohamad (2013), Journal on Intelligent Information Management, Vol. 5, pp. 10–17.

[34] Srivastava, Jaideep, Desikan, Prasanna and Kumar, Vipin (2015), Web Mining-Concepts, Applications & Research Directions Seen on 10/11/2015 Available at http://www.internetworldstats.com/emarketing.htm