RESEARCH ARTICLE OPEN ACCESS

Machine Learning Spectrum for Web Data Analytics

Sunny Sharma[1]
Department of Computer Science, Guru Nanak Dev University, Amritsar, Punjab

--************************----------------------------------

Abstract:

There are numerous ways to analyse the web information, generally web substance are housed in large information sets and basic inquiries are utilized to parse such information sets. As the requests expanded with time, mining web information amended to meet challenging task in a web analysis. Machine learning methodologies are the most up to date one to go into these analysis forms. Different approaches like decision trees, association rules, Meta heuristic and basic learning methods are embraced for making web data appraisal and mining data from various web instances. This study will highlight these approaches in perspective of web investigation. One of the prime goals of this exploration is to investigate more data mining approaches alongside machine learning systems, and to express emerging collaboration of web analytics with artificial intelligence.

Keywords **— Data Analytics, Sentiment Mining, Machine Learning**

--************************----------------------------------

I. INTRODUCTION

Throughout the years Web Analytics have developed to quantify and help business requests and research queries. Analytics assumes to be essential part in the present time for the advancement procedure of any association. Presently a day's organizations are concentrating more towards learning revelation & separating the concealed data from vast information sets. While doing the online shopping, swipe our card for installment, shop something in bigger market places, messaging/emailing, conferences and performing different business assignments online over the web we generate huge or we can say generating big data, and the hidden part is that people are unaware of that. Organizations use such kind of data to do the market analysis and sets the trends as well as change their policies such as changing cost, customer satisfaction etc. usually they consider the sentiment of number of visitors on site towards particular block [1]. So such data helps the companies to gather the statistics and do the market analysis.

II. MULTISTEP APPROACH FOR WEB ANALYTICS PROCESS

The Web Analytics Process uses the multistep approach in its first step collection of data is done, and this task is done by performing association of data items to get similar data of interest together. This step provides the rules on which relations between data items are established. Secondly, preprocessing of data in information is done and the data is transformed into information on the basis of classification of data items. Variety of classification techniques are existing like linear programming, decision trees, neural networks etc. in the third step the development of key performance indicators is carried out, based on this heuristic information, regression analysis done which sets the ground for alternative business strategies. Finally formulation of online strategy is done by evaluating performance indicators. Web analytics assist the formulation process by various models of machine vision in web analytics. It can be expressed in Figure 1.

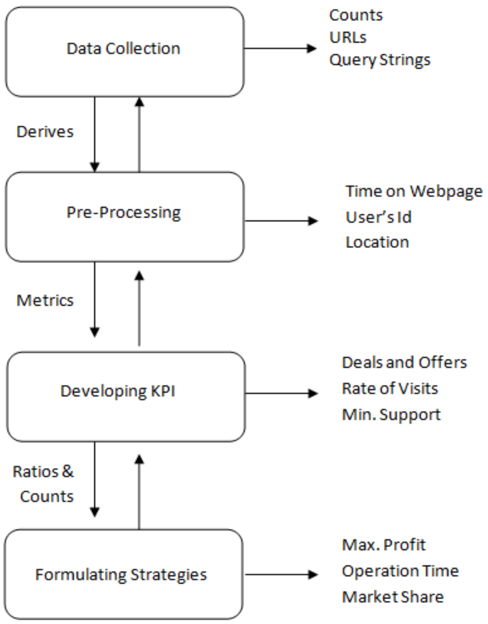

Fig. 1 Web Data Analytics Process Model

III. MACHINE LEARNING APPROACHES TO WEB ANALYTICS

The main objective of machine learning is that, to increase the accuracy of the algorithms and the objectives of data mining are not different from machine learning from practical point of view. Both do not make efforts for collecting or managing or pre-processing data, here the emphasis is just on expanding the precision. Distinction is that, in data mining extends above all else machine learning calculations that are appropriate and versatile to tremendous and monstrous datasets without considering exactness are talked about. At the end of the day in data mining ventures pertinence and adaptability are more critical than exactness. The various approaches are used to do such task these are as follow:

- Decision Tree Learning
- Association Rules Learning
- Artificial Neural Networks
- Meta Imperative Learning
- Support Vector Machines
- Analysis through Boosted Trees
- Clustering Process in Analysis

IV. LITERATURE SURVEY

In 2001 G. Adomavicius et. al. [2] observed personalization community by taking customers profiles into account for finding out behavior of customers and nature of their interactions. As a model they have developed two parts of the whole framework. One part contains data sets from customer profiles and other contains rule base for fitting these profiles under some logical processing. In 2001 J. Su. et. al. [3] found medical data discovery using three basic techniques named Bayessian Network, Decision Trees and Back Propagation Neural Networks. According to this approach mined medical information is not only classified for research perspectives but also made available to the physician to improve his practices. Six steps of knowledge discovery include data collection firstly followed by data filtration, data enhancement, data encoding, data mining and at last knowledge representation. High correlation parameter is chosen for construction of Bayessian networks. While learning rate of 0.5 with learning time of 1000 nanoseconds is taken for back propagation technique. After generating decision tree the results of three algorithms were discussed by considering average of high accuracy of data sets. In 2001 Y. Leung et. al.[1] proposed regression class mixture decomposition method for mining regression classes in large data sets. They have used iterative and genetic based optimized objective function to guide regression curves. They have found that it is not a real matter that only a regression analysis can solve purpose of identification in large data sets. For more accurately modeling this identification process multiple approaches have to be fit to explore classes from large data sets. In November 2003, B. Fan et. al. [4] proposed a spatial data mining method for web analytics for evaluating customer's intelligence. The idea was to separate non spatial data of customer from spatial data items. Non spatial data of customer such as age of customer and the population of a particular region whereas spatial data comprises location of customer with locality as a region. On the basis of data classification techniques like k means clustering, spatial overlapping method is designed for detecting potential customers. In 2005 J. Xing et. al. [6] used GIS as spatial data analysis tool which can automatically discover implicit knowledge from spatial data. Their idea was to develop web based data mining system by integrating state of the art GIS and data mining functionality in a closely coupled open and extensible system architecture. The idea was to extend predictive model markup language (PMML) with spatial data mining to get new version named SPMML. In 2006 W. Niblack et. al. [8] allow uniform access to wide variety of sources like bulletin boards, news feeds, reports by analyst, a platform for very large scale text applications that is known as web fountain. By using set of hosted web services allows easy development of documents in the form of text that are useful for deriving end user applications. With the help of application programming interfaces the users are able to develop remote analytical components, most useful application of web fountain is reputation management. In 2006, S. Yoon et. al. [9] [10] found usefulness of biochips for acquisition of biological data with high throughput. The main advantages of micro fluidic lab on a chip include ease of use, speed of analysis, low sample and reagent consumption, and high reproducibility due to standardization and automation. Without effective data-analysis methods, however, the merit of acquiring massive data through biochips will be marginal. In 2011 C. Yang at. al. [7] found that analysis of developing web opinions and social interaction is potentially valuable for discovering ongoing topics of interests of the public like terrorist and crime detection, understanding how topics evolve together with the underlying social interaction between participants, and identifying important participants who have great influence in various topics of discussions. In 2011, C. Wang et al. [5] observed that with the expansion of web on the internet, it becomes the great demand of times to analyse user's sentiments or opinions for the betterment of society and humanity. In 2015 N. Koul et. al. [11] performs survey of machine learning approaches to opinion mining & sentiment analysis in text data. In 2016 I. Cho et al [12] observed the task of identifying relevant textual information on places, time periods, people, and events is challenging since much information on locations and times as well as their relationships are described in many separate texts with no clear linkage or structure. They have proposed a visual analytics approach to help people gain knowledge through making connections on places, times, and events. The comparative analysis of few literature surveys is expressed in Table1.

V. CONCLUSION

Data mining alongside with machine learning systems for sentiment analysis is beginner approach. This paper has observed that mining accuracy through regression class mixture decomposition over Gaussian mixture decomposition the results would be more significant in customer profile building, mining medical databases and intelligence mining, if sentiment analysis of available data sets is done. This paper also highlighted useful application of data analytics for reputation management in web fountain, and importance of web opinion mining, limitations of existing methods of web text mining. Focus was to highlight various applications of web analytics.

TABLE I: Comparative Study Literature Survey

Methodology & Tool	**Year of Publication**	**Data Set Type/ Investigation Area**	**Author**	**Results Discussion**
Rule Based Classification	2001	**Building Customer Profiles**	G. Adomavicius et. al [2]	In 2001 author observed personalization community by taking customer's profiles into account for finding out behavior of customers and nature of their interactions. As a model they have developed two parts of the whole framework. One part contains data sets from customer profiles and other contains rule base for fitting these profiles under some logical processing. These two parts are named by them as data model and profile models.
ANN & Fuzzy Logic , Decision tree Induction	2001	**Mining Regression Classes**	J. L. Su. Et. al.[3]	In 2001 Author found medical data discovery using three basic techniques named Bayessian Network, Decision Trees and Back Propagation Neural

				Networks. According to this approach mined medical information is not only classified for research perspectives but also made available to the physician to improve his practices. Six steps of knowledge discovery include data collection firstly followed by data filtration, data enhancement, data encoding, data mining and at last knowledge representation. High correlation parameter is chosen for construction of Bayessian networks. While learning rate of 0.5 with learning time of 1000 nanoseconds is taken for back propagation technique. After generating decision tree the results of three algorithms were discussed by considering average of high accuracy of data sets.
Regression analysis	2001	**Mining Medical Databases**	Y. Leung et. al.[1]	Author proposed Regression Class Mixture Decomposition (RCMD) method for mining regression classes in large data sets. They have used iterative and genetic based optimized objective function to guide regression curves. They have found that it is not a real matter that only a regression analysis can solve purpose of identification in large data sets. For more accurately modeling this identification process multiple approaches have to be fit to explore classes from large data sets

REFERENCES

[1] Yee Leung, Jiang-Hong Ma, & Wen-Xiu Zhang. (2001). A new method for mining regression classes in large data sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(1), 5-21. doi:10.1109/34.899943.

[2] Adomavicius, G., & Tuzhilin, A. (2001). Using Data Mining Methods to Build Customer Profiles. *Data Mining and Knowledge Discovery*, *5*(1/2), 33-58. doi:10.1023/a:1009839827683.

[3] Jenn-Lung Su, Guo-Zhen Wu, I-Pin Cha. (2001, October). *THE APPROACH OF DATA MINING METHODS FOR MEDICAL DATABASE*. Paper presented at IEEE, Istanbul, Turkey.

[4] BO FAN, YI-JUN LI, XIANG-BIN YAN. (2003, November). *SPATIAL DATA MINING METHOD FOR CUSTOMER INTELLIGENCE*. Paper presented at IEEE, China.

[5] Wang, C., Xiao, Z., Liu, Y., Xu, Y., Zhou, A., & Zhang, K. (2013). SentiView: Sentiment Analysis and Visualization for Internet Popular Topics. *IEEE Transactions on Human-Machine Systems*, *43*(6), 620-630. doi:10.1109/thms.2013.2285047.

[6] Jeonghee Yi Wayne Niblack. (2005). Sentiment Mining in WebFountain. San Jose, CA.

[7] Yang, C. C., & Dorbin Ng, T. (2011). Analyzing and Visualizing Web Opinion Development and Social Interactions With Density-Based Clustering. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *41*(6), 1144-1155. doi:10.1109/tsmca.2011.2113334.

[8] Jin Xingxing, Cai Yingkun, *Xie Kunqing, Ma Xiujun, Sun Yuxiang, Cai Cuo. (2006, January). *A Novel Method to Integrate Spatial Data Mining and Geographic Information System*. Paper presented at IEEE, Beijing, China.

[9] Wang, C., Xiao, Z., Liu, Y., Xu, Y., Zhou, A., & Zhang, K. (2013). SentiView: Sentiment Analysis and Visualization for Internet Popular Topics. *IEEE Transactions on Human-Machine Systems*, *43*(6), 620-630. doi:10.1109/thms.2013.2285047.

[10] Yang, C. C., & Dorbin Ng, T. (2011). Analyzing and Visualizing Web Opinion Development and Social Interactions With Density-Based Clustering. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *41*(6), 1144-1155. doi:10.1109/tsmca.2011.2113334.

[11] N. Koul, S. Hassan," A Survey of Machine Learning Approaches to Opinion Mining & Sentiment Analysis in Text", International Conference on Communication, Information and Computing Technology, 12-13 May, 2015.

[12] Cho, I., Dou, W., Wang, D. X., Sauda, E., & Ribarsky, W. (2016). VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History. *IEEE Trans. Visual. Comput. Graphics*, *22*(1), 210-219. doi:10.1109/tvcg.2015 .2467971

[13] S. Sharma, " A Review on Efficacy of Artificial Neural Networks in Medical & Business Areas ", International Journal of Recent Trends in Engineering & Research, Volume 02, Issue 04, April – 2016

[14] S. Sharma, " Cervical Cancer stage prediction using Decision Tree approach of Machine Learning ", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 4, April 2016.

[15] S. Sharma, "A Study on Data Mining Horizons", International Journal of Recent Trends in Engineering & Research, Volume 02, Issue 04; April – 2016.

[16] http://en.wikipedia.org/wiki/Data_mining