International Conference on Recent Advancement in Science & Technology- 2020
(ICRAST-2020)

A Review on Website Quality Estimation Techniques

[1]Sudha Gautam, [2]Anurag Jaiswal
Dept of Computer Science
Goel Institute of Technology and Management, Lucknow, India
sudha.pdh90@gmail.com, anurag.jaiswal@goel.edu.in

Abstract: Web Mining has emerged as an important research area and association rule mining becomes one of the most attractive tasks of data mining, nevertheless, there are still some unsolved problems. A problem of classical association rule mining algorithms is that they are mostly limited to deal with a relative small size of databases, in the number of records and in the number of attributes as well. However, the rapid development in the database technology, especially in the storage capacity has realized the databases with the size of terabytes, consequently the current databases may also have a huge number of the attributes.

Keywords: Genetic Algorithm, Web Mining, Estimation analysis, Data prediction.

1. Introduction:

Web mining is the automated extraction of predictive information from large databases. It is an interdisciplinary field involving databases, machine learning, statistics, artificial intelligence, information retrieval and visualization in order to extract high level knowledge from real-world data sets [1]. Data mining is the core step of the knowledge discovery process. It has attracted a lot of interest in the database community mainly motivated by the explosive growth and availability of huge amount of data in all fields of science, business, medicine, government, etc. Through data mining, it is possible to find novel and useful patterns from databases. Many data mining techniques have been developed since the last decade, that go beyond simple analysis, contributing greatly to business strategies, knowledge bases and scientific research. Furthermore, data mining has become a key topic in database technology.

Association Rule mining has received great interest by the scientific community since it is very convenient to find patterns or rules on which attributes that occur frequently together in a given dataset. Association rule mining has been introduced in 1993 by Agrawal et al. [2] and it is continuously been applied to several fields such as medicine, finance, stock market, manufacturing, e-business, intrusion detection, bio-informatics, etc.

In the last decade, several techniques have been applied to association rule mining such as artificial neural networks (ANN), expert systems (ES), linear programming (LP), database systems (DS), and evolutionary computing (EC). EC is a computational technique inspired by the natural evolution process that imitates the mechanism of natural selection and survival of the fittest. When EC is applied to data mining, they provide robust search methods which perform a global search in the space of candidate solutions. In contrast, most association rule induction methods perform a local, greedy search in the space of candidate rules. Intuitively, the global search of evolutionary algorithms can discover interesting rules and patterns that would be missed by the greedy search performed by most rule induction methods [3].

Web Mining has emerged as an important research area and association rule mining becomes one of the most attractive tasks of data mining, nevertheless, there are still some unsolved problems. A problem of classical association rule mining algorithms is that they are mostly limited to deal with a relative small size of databases, in the number of records and in the number of attributes as well. However, the rapid development in the database technology, especially in the storage capacity has realized the databases with the size of terabytes, consequently the current databases may also have a huge number of the attributes. The aim of our work is to propose an improved method for mining and pruning association rules based on evolutionary computational methods such as Genetic Algorithms (GA) and Genetic Relation Algorithm (GRA). Furthermore, to demonstrate the usefulness of GA as a selection method, GA is applied to Web mining, where one of the major problems is the poor quality of the information retrieved because of the lack of performing additional computation.

As the usage of web started to increase, so does the demand of data mining. Web mining is the application of data mining techniques to discover usage patterns from large Web repositories. It reveals interesting and unknown knowledge about both users and websites which can be used for analysis. It is used to understand customer behaviour, evaluate the effectiveness of a particular website and help quantify the success of a marketing campaign. Web mining can be classified into three types based on the type of data:

Web content mining – it is the process of extracting useful information and knowledge from the web

contents/data/documents. Content may consist of text, images, audio, video or
structured records such as lists and tables. Web content mining is differentiated from two different points of view: Information Retrieval View and Database View.

Web structure mining – it is the process of using graph theory to analyse the node and connection structure of a website. It tries to discover the underlying link structures of the web. It can be used to generate information on the similarity or the difference between different websites.

Web usage mining – it attempts to discover useful knowledge from the data obtained from web user sessions. It tries to find usage patterns from the web data to understand and better serve the needs of Web-based applications. Some applications of web usage mining are adaptive websites, web personalization and recommendation, business intelligence.

2. Applications Of Web Mining

- It has its great use in e-commerce and eservices
- In e-learning
- Self-organizing websites
- Digital libraries
- E-government
- Security and crime investigation.

3. Web Mining Uses

This technology has enabled ecommerce to do personalized marketing, which eventually results in higher trade volumes. The predicting capability of the mining application can benefit the society by identifying criminal activities. The companies can establish better customer relationship by giving them exactly what they need. Companies can understand the needs of the customer better and they can react to customer needs faster. The companies can find, attract and retain customers, they can save on production costs by utilizing the acquired insight of customer requirements. They can increase profitability by target pricing based on the profiles created. They can even find the customer who might default to a competitor the company will try to retain the customer by providing promotional offers to the specific customer, thus reducing the risk of losing a customer.

4. Related Work:

It includes Web-commerce applications, Intelligent web search, Hypertext classification and Categorization, Information/trend monitoring, Analysis of online communities, Improving the relationship between the website and the user, Recommendations to modify the web site structure and content Web personalization. Web mining applications in E-commerce and E-services is a new research direction in the area of web mining. Among all of the possible applications in web research, e-commerce and e-services have been identified as important domains for Web-mining techniques. Web-mining techniques also play an important role in e-commerce and eservices, proving to be useful tools for understanding how ecommerce and e-service Web sites and services are used.

Past few years have witnessed an explosive growth in the information available on the World Wide Web (WWW). Today, web browsers provide easy access to myriad sources of text and multimedia data. More than 1 000 000 000 pages are indexed by search engines, and finding the desired information is not an easy task. This profusion of resources has prompted the need for developing automatic mining techniques on the WWW. Below is given a brief literature review of the work done by various authors in the firld of web mining using different techniques.

Ee-Peng Lim and Aixin Sun represented ontology as a set of concepts and their inter-relationships relevant to some knowledge domain. The knowledge provided by ontology is extremely useful in defining the structure and scope for mining Web content. Reviewed Web mining and described the ontology approach to Web mining. The application of these Web mining techniques to digital library systems was also discussed.

B. RAJDEEPA, DR. P. SUMATHI, (2014), proposed the Web Structure according to an effectiveness criterion. The frequencies of procedure and time division to click on a link were utilized for constructing the web user utility. Showd that researching probabilistic behaviours have permitted quality development of the hyperlink structure, using a Hybrid GA/PSO. The planned methodology is tested on a real web site and the results are interpreted using the web user benefits by links. The simplicity and effectiveness of Particle Swarm Optimization technique based on the idea of Swarm cleverness was executed in high-dimensional order clustering investigation for web usage mining. As the Hybrid GA/PSO algorithm has fast convergence and simple achievement, numerous enhanced versions of Hybrid GA/PSO algorithm have been developed to resolve that difficulty of GA and PSO.

Soumadip Ghosh, Amitava Nag, Debasish Biswas, Arindrajit Pal, (2011), have dealt with a challenging association rule mining problem of finding frequent itemsets from XML database using proposed GA based method. The method, described here is very simple and efficient. It was successfully tested for different XML databases. The results reported are correct and appropriate. Also tried to compare the performance of GA based method proposed with the FP-tree algorithm.

V.V.R. Maheswara Rao, Dr. V. Valli Kumari and Dr. K.V.S.V.N Raju, (2010), introduced the Advanced Optimal Web Intelligent Model with granular computing nature of Genetic Algorithms, IOG. The IOG model is designed on the semantic enhanced content data, which works more efficiently than that on normal data. The unique parameters of the IOG fitness function generates balanced weights, to represent the significance of characteristics, which yields the dissimilar characteristics of page vector. The evaluation function of IOG improves the page quality and reduces the execution time to a specified number of iterations as it considers practical measures like page, link and mean qualities. The genetic operators are intended in drawing the stickiness among the characteristics of a page vector. By integrating all the above the web intelligent model IOG, it significantly improves the investigation of web user usage behaviour.

Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar, (2010), dealt with a challenging association rule mining problem of finding frequent itemsets using proposed GA based method. The method, described is very simple and efficient. This is successfully tested for different large data sets. The results reported are correct and appropriate. However, a more extensive empirical evaluation of the proposed method will be the objective of future research. Compared the performance of GA based method proposed with the FP-tree algorithm.

C.J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M.J. del Jesus, S. García, (2012), presented the methodology used in an e-commerce website of extra virgin olive oil sale called www.OrOliveSur.com. Described the set of phases carried out including data collection, data preprocessing, extraction and analysis of knowledge. The knowledge is extracted using unsupervised and supervised data mining algorithms through descriptive tasks such as clustering, association and subgroup discovery; applying classical and recent approaches. The main purpose is to apply a set of descriptive data mining techniques to obtain rules that allow to help to the webmaster team to improve the design of the website. Techniques used are clustering, association rules and subgroup discovery. Therefore, here web mining study can be analysed using different knowledge with respect to these techniques used, i.e. a variety of groups and/or rules associated for each technique.

Kunjal Mankad, Priti Srinivas Sajja and Rajendra Akerkar, (2011) presented design and development of a system to automatically evolve rules through genetic-fuzzy approach. Highlighted the advantages of genetic and fuzzy hybridization and proposed a framework to evolve rules automatically. The objective of the framework is to reduce the development effort and guide the development procedure in user friendly fashion. The architecture presents a novel approach which integrates fuzzy logic with system design as well as back end of the framework to evolve the fuzzy rules using genetic algorithm approach. To experiment the working of the proposed framework, a system to measure multiple intelligence is selected. They described the detail methodology including encoding strategy of rules, suitable genetic operators and fitness function to evolve fuzzy rules for the selected system. The initial population of rules, evolved generations and output results are also described. Concluded with the scope and applications of the work to other domains.

Deepak Kumar Niware, et. al., (2014), presented use of Genetic algorithm in Fuzzy c-means algorithm to select initial center point for clustering in FCM. Worked to provide optimum initial solution for FCM with the help of genetic algorithm to reduce the error rate in pattern creation. Concluded that the cluster created using Fuzzy c-means technique have high error-rate as compared to the cluster created through Genetic algorithm based Fuzzy c-means. Error-rate shows that the data loss is occurred in techniques. But, in comparison of FCM, the proposed GFCM technique has less data loss. From the experimental results it is concluded that GFCM creates rich cluster, more cluster and efficient cluster than FCM technique.

Arben Asllani, Alireza Lari, (2007), introduced a genetic algorithm, to be used in a model-driven decision-support system for web-site optimizations. The algorithm ensures multiple criteria web-site optimizations, and the genetic search provides dynamic and timely solutions independent of the number of objects to be arranged.

Vikrant Sabnis, R. S. Thakur, (2013), proposed a genetic based model for mining frequent patterns in web content data. In the proposed genetic operator, crossing over method leads to offspring which must survive the certain fitness test or conditions to become frequent pattern and ancestor for next patterns. In this way the useless individuals or candidates are pruned out thereby reducing the number of candidates for next test. Also this approach requires only one scan of database. Thus this model is able to address the issues of large number of candidate generation and number of database scans. This genetic algorithm based approach needs only one dataset scan and during scanning the datasets are converted into chromosome. It generates less number of candidates or offsprings and does not require level by level candidate generation done in traditional approaches. This is good approach to get maximal itemsets. It is easy to implement as parallel process to get frequent itemsets. In this case each pair of chromosomes can run on individual processor and perform logical AND operations. In next pass only surviving offsprings are used to repeat same process to get higher size offspring.

Ranu Singhal, Nirupama Tiwari, (2013), proposed a modified FCM algorithm using multi-objective genetic algorithm for web log clustering. Showed that to a certain degree, which overcomes the defects that FCM is sensitive to initial value and it is easily able to be trapped in a local optimum. The practicability of this approach is analyzed in principle, and its practical effect is confirmed by experiment in technical. The experiment results show that the global distribution characteristic of the space clustering centers which are found during the process of the fuzzy clustering analysis by utilizing improved GA is suitably kept, so clustering effect is more rational.

Stephen G. Matthews, Mario A. Gongora, Adrian A. Hopgood, Samad Ahmadi, (2013) have demonstrated the problem of losing temporal fuzzy association rules on real-world Web log data for the first time and presented a novel solution. Previous approach of using a GA and the 2-tuple linguistic representation has been improved by transforming the dataset to a graph, which ensures valid itemsets are discovered, and modifying the fitness function. The execution time of the GA-based approach is longer, however, the contribution to knowledge is that it can discover rules that a traditional approach cannot, and the rules have higher temporal fuzzy support. The GA-based approach is recommended as complementary to existing algorithms, because it discovers extra rules that a traditional algorithm does not. The decision to use this complementary approach can rely on understanding what temporal changes may be present in the application domain (e.g., seasonal and/or scheduled events). Showed that lowering minimum support/confidence would overcome the problem of losing rules with traditional approaches, however, the number of rules increases, which is undesirable in association rule mining. Further work will explore different enhancements, and different approaches to tackle the same problem.

Pramod Vishwakarma, Yogesh Kumar, Rajiv Kumar Nath, (2012), used Genetic Algorithm to mine the real world dataset in medical domain. The GA discovered the optimal features or peaks in mass spectrometry data. On the basis of the specific features extracted from mass spectrometry data by using genetic algorithm, distinguished the cancer patients from control patients. Finally results of the simulation were discussed and graphs of the result were plotted in for the better understandings of the findings. The work is only the stating of exploration of the clinical data or the medical datasets. Many other issues are still to be resolved and warrant further investigation.

Romero et. al., (2013), showed how web usage mining can be applied in e-learning systems in order to predict the marks that university students will obtain in the final exam of a course. Developed a specific Moodle mining tool oriented for the use of not only experts in data mining but also of newcomers like instructors and courseware authors. The performance of different data mining techniques for classifying students are compared, starting with the student’s usage data in several Cordoba University Moodle courses in engineering. Several well known classification methods have been used, such as statistical methods, decision trees, rule and fuzzy rule induction methods, and neural networks. Carried out several experiments using all available and filtered data to try to obtain more accuracy. Discretization and rebalance pre-processing techniques have also been used on the original numerical data to test again if better classifier models can be obtained. Finally, examples of some of the models discovered have been given and explained that a classifier model appropriate for an educational environment has to be both accurate and comprehensible in order for instructors and course administrators to be able to use it for decision making.

5. Conclusion:

The rapid advances in data generation, availability of automated tools in data collection and
continued decline in data storage cost enable with high volumes of data. In addition, the data is non scalable, highly dimensional, heterogeneous and complex in its nature. This situation creates inevitably increasing challenges in extracting desired information. Thus, web mining evolves into a fertile area and got the focus by many researchers and business analysts. Web mining is a methodology that blends conventional techniques with incremental algorithms. Web mining is the application of data mining techniques to discover and retrieve useful information and patterns from the WWW documents and services. Web usage mining is the process of automatic discovery and investigation of patterns in click streams and associated data collected or generated as a result of user interactions with web resources on web sites. The main goal of web usage mining is to capture, model and analyze the behavioural patterns and profiles of users interacting with web sites. The discovered patterns are usually represented as collection of pages, objects or resources that are frequently accessed by groups of users with common needs or interests. The primary data resources used in web usage mining are log files generated by web and application servers.

References:

[1] M. Berry and G. Lino_, Data Mining Techniques. For Marketing, Sales and Customer Support, John Wiley & Sons, Inc. 1997.

[2] R. Agrawal, T. Imielinski and A. Swami, \Mining Association Rules between Sets of Items in Large Databases", In Proc. of the 1993 ACM SIGMOD International Conference on Management of Data, 1993.

[3] A. Freitas, \Data Mining and Knowledge Discovery with Evolutionary Algorithms", Springer-Verlag, New York Inc., 2002.

[4] A. A. Freitas, \Data Mining and Knowledge discovery with Evolutionary Algorithms", IEEE Transactions on Evolutionary Computation, Vol. 7, No. 6, pp. 517-518, 2003.

[5]. Ee-Peng Lim and Aixin Sun, WEB MINING - THE ONTOLOGY APPROACH, pp. 1-8.

[6]. B. RAJDEEPA, DR. P. SUMATHI, (2014), “WEB STRUCTURE MINING FOR USERS BASED ON A HYBRID GA/PSO APPROACH”, Journal of Theoretical and Applied Information Technology, Vol.70 No.3, pp. 573-578.

[7] Soumadip Ghosh, Amitava Nag, Debasish Biswas, Arindrajit Pal, (2011), “ XML MINING USING GENETIC

ALGORITHM", Journal of Global Research in Computer Science, Volume 2, No. 5, pp. 86-90.

[8] V.V.R. Maheswara Rao, Dr. V. Valli Kumari and Dr. K.V.S.V.N Raju, (2010), "An Advanced Optimal Web Intelligent Model for Mining Web User Usage Behavior using Genetic Algorithm", ACEEE Int. J. On Network Security, Vol. 01, No. 03, Dec 2010, pp. 44-49.

[9] Anurag, R. Sharma, " Load Forecasting by using ANFIS", International Journal of Research and Development in Applied Science and Engineering, Volume 20, Issue 1, 2020.

[10] R. Sharma, Anurag, " Load Forecasting using ANFIS A Review", International Journal of Research and Development in Applied Science and Engineering, Volume 20, Issue 1, 2020.

[11] R. Sharma, Anurag, " Detect Skin Defects by Modern Image Segmentation Approach, Volume 20, Issue 1, 2020.

[12] Anurag, R. Sharma, " Modern Trends on Image Segmentation for Data Analysis- A Review", International Journal of Research and Development in Applied Science and Engineering, Volume 20, Issue 1, 2020.

[13] C.J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M.J. del Jesus, S. García, (2012), "Web usage mining to improve the design of an e-commerce website: orolivesur.com", Expert Systems with Applications 39, 11243–11249.