International Journal of Advance and Applied Research (IJAAR)
Peer Reviewed Bi-Monthly

ISSN – 2347-7075
Impact Factor – 7.065
Vol.8 No.1 Sept – Oct 2020

WEB MINING USING SOFT COMPUTING METHODOLOGY: A SURVEY

Dr. A.M. Sote
Assistant Professor,
Department of Computer Science
Arts, Commerce and Science College, Arvi.

ABSTRACT:

In the current era worldwide user are connected to each other through internet and their computer they can share information globally with help of medium which is known as World Wide Web. It is very important to get a correct and meaningful information to each user thus gives rise to new term web mining. Web mining is the application of data mining techniques to the purpose of learning or extracting knowledge from online data. Web mining included with number of techniques, in which one of them is soft computing. In this paper we are explaining some soft computing methodology which is helpful to the researcher in the field of AI and online technologies.

Keyword: Web mining, KDD, Soft computing.

INTRODUCTION:

The web and its usage continuously growing, so too grows the opportunity to analyse web data and extract all manner of useful knowledge from it. The discovery of useful, user information and server access patterns allow web based organizations to mining user access patterns and helps in future developments specially knowledge discovery analysis. The Web mining technologies are the right solutions for knowledge on the web data. Web mining is broadly defined as the application of data mining techniques to extract knowledge from web data, where at least one of the web structure (hyperlink) or web usage (Weblog) data is used in the mining process.

Soft computing tools can be used to observe patterns and evolution in the web usage data with respect to target metrics. However preparing and transforming the data for analysis and translating the findings into actionable insights require domain knowledge and human expertise. This human expertise is often the best tool in an analytics task, and this is extremely crucial, if not impossible, to automate. As a result, soft computing techniques usually mean by presenting interesting patterns to the analyst in that manner helpful for interpretation.

Organizing soft computing tools in a web environment thus places a high requirement on the technical skill as well as the business expertise of users. As a result, rapid, successful organizations across domains have been a significant.

Fuzzy logic sets (FL), artificial neural networks (ANNs), genetic algorithms (GAs), and rough set (RS) theory are the fundamental methodologies of soft computing. Fuzzy sets dealing with uncertainty in natural framework. Neural networks (NNs) are provide learning and generalization tool. RSs help in granular computation and knowledge discovery. These methodologies of soft computing is useful in finding and extracting an accurate and meaningful data from online resources which is explain latter on in this paper.

In this paper we make survey on some soft computing methodology which is useful in web mining application. We manage this paper as in next section we reviewed some literature relating with web mining and soft computing and its application in later section we explained some soft computing methodology which is applicable in web mining and finally we conclude this paper.

LITERATURE SURVEY:

Soft computing is the association of methodologies which is helpful to finding a web data precisely. In this section we make a survey of literature which is applicable in research work of soft web mining.

Lee and Kim developed a fuzzy web information retrieval system [1]. K. Nowacka et al. proposed a comprehensive model of information retrieval (IR) based on Zadeh's linguistic statements[2]. Recent research on fuzzy quantification for information retrieval has proposed the application of semi-

fuzzy quantifiers for improving query languages were discussed in [3]. O. Zamir and O. Etizioni[4] listed key requirements of web document clustering as measure of relevance, browsable summaries, and ability to handle overlapping data, snippet tolerance, speed and incremental characteristics. In [5], fuzzy c medoids (FCNdd) and fuzzy c Trimmed medoids (FCTMdd) are used for clustering of web documents and snippets. In [6], the use of soft computing comprising fuzzy logic (FL)is discussed with the present web mining techniques.

In [7], most popular algorithms for accurate clustering had been described i.e. K-means[8], CURE[9], BIRCH[10] and ROCK [11], which is used by various authors in their research works. BIRCH (Balance and Iterative Reducing and Clustering Hierarchies) is useful algorithm for data represented in vector space. It also works well with outliers like CURE

M. Dimitrijevic and Z. Bosnjak [12], proposed to improve the problem of web usage association rule over-generation by pruning the rules that contain directly linked pages out of the rule set. A hybrid learning system should learn more effectively than systems that use only one of the information sources. KBANN(Knowledge-Based Artificial Neural Networks)is a hybrid learning system built on top of connectionist learning techniques[13].

Mercure[14], described another information retrieval system based on a connectionist approach and modelled by a three layered network. The network is composed of a query layer (set of query terms),a term layer representing the indexing terms and a document layer.

In [15], the self organizing map (SOM) algorithm is used to automatically organize very large and high dimensional collections of text documents onto two dimensional map displays known as “WEBSOM”. In a clustering problem, it is always difficult to determine the number of clusters. H. Rana and M. Patel [16], discussed the auto clustering feature of SOM is more effective and objective than the K-means method. M. Pazzani et al.[17], described Syskill & Webert, a software agent that learns to rate pages on the World Wide Web (WWW), deciding what pages might interest a user. The user rates explored pages on a three point scale, and Syskill & Webert learns a user profile by analyzing the

information on each page. In [18], C. Drummond et al. Described a technique termed as Active Browsing.

A.F. Ali with M.H. Marghny [19], proposed a steady state genetic algorithm (GA) which evolve a population of pages is presented. Asllani and Lari [20], proposed a multiple web-site optimizations using GA. Jin Cheng et al.[21], proposed an improved genetic algorithm which solves the issues in two generation competitive genetic. In[22], K. P. Rao and G. K. Chakravarthy had proposed an efficient web association rule mining approach with Apriori Algorithm and Genetic Algorithm features. C. Romero et al. [23] proposed, a dynamic elaboration methodology, where the evaluation information is used to modify the course and to improve its performance for better student's learning. J. Usharani, and K. Iyakutti [24], proposed a novel approach - genetic based Apriori algorithm for web crawling.

In RS theory, Information granules refer to homogeneous blocks/clusters of documents as described by a particular set of features which may vary over the clusters. Wong et al.[25], suggested reducing the dimensionality of terms by constructing a term hierarchy in parallel to a document hierarchy. In [26] U. Straccia presented a logic-based framework in which multimedia objects medium dependent properties (objects low level features) and multimedia objects medium independent properties (abstract objects features, or objects semantics) are addressed in a principled way.

In [27], a new philosophical view and methodology called "Explanation Oriented Data Mining" is introduced.. H. H. Inbarani and K. Thangavel et al. [28], focused on obtaining user profiles based on intelligent rough clustering techniques. Different problems can be addressed though Rough Set Theory. In S. Rissino and G. L. Torres [29], described relationship between Rough Set Theory and the Dempster-Shafer Theory and between rough sets and fuzzy sets. Sung-Kwun Ohet al.[30], introduced an advanced architecture of genetically optimized Hybrid Fuzzy Neural Networks (gHFNN) and develop a comprehensive design methodology supporting their construction..

Dynamic Evolving Neural Fuzzy Network(dmEFuNN) [31], is a modified version of the EFuNN with the idea of not only the winning rule node's

activation is propagated but a group of rule nodes that is dynamic selected for every new input vector and theiractivation values are used to calculate the dynamical parameters of the output function. In [32], K. K. Ang and C. Quck, presented a novel hybrid intelligent Rough set-based Neuro-Fuzzy System (RNFS). U. Keerthika et al. [33], constructed a framework for Fuzzy Temporal Rule Based Classifier that uses fuzzy rough set and temporal logic in order to mine temporal patterns in medical databases. K. Y. Shen and G. H. Tzeng [34], proposed an integrated inference system to predict the financial performance of banks. In [35], S. Ding et al. summarized the review on the recent research development of RNNs.

SOFT COMPUTING METHODOLOGY:

Soft computing is an association of methodologies that works more effectively in collaboration with and provides, in one form or another, flexible information processing capability for handling real-life ambiguous situations. Soft computing main aim is to take more advantage of the progressiveness for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve tractability, robustness, low-cost solutions, and close resemblance to human-like decision making. In other words, it provides the foundation for the conception and design of high machine IQ (MIQ) systems, and, therefore, forms the basis of future generation computing systems.

Fuzzy Logic (FL), Rugh Sets(RS), Artificial Neural Networks(ANN), and Genetic algorithm(GA). As are the principal components, where FL provides algorithms for dealing with imprecision and uncertainty arising from vagueness rather than randomness, RS for handling uncertainty arising from limited discriminately of objects, ANN the machinery for learning and adaptation, and GA for optimization and searching[9].

For handling issues related to incomplete/imprecise data/query, approximate solution, human interaction and understandability of patterns and deduction, and mixed media information Fuzzy Logic methodology are used [8]. Neural Networks methodologies are used for modeling highly nonlinear decision

boundaries, generalization and learning, self organization, rule generation, and pattern discovery.

Genetic Algorithms methodology is to be useful for prediction and description, efficient search, and adaptive and evolutionary optimization of complex objective functions in dynamic environments. Rough Set theory methodology applicable in "fast" algorithms for extraction of domain knowledge in the form of logical rules. In recent times, various combinations of these tools have been made in soft computing paradigm, among which neuro-fuzzy integration is the most powerful and efficient technology which is mostly used by various research in the field of Artificial Intelligence.

Fuzzy-granular which is computation may be done with perception, in nature. Web data, being fundamentally unlabeled, imprecise/incomplete, heterogeneous, and dynamic, are found to be a very good applicant for finding a meaningful and efficient data using the soft computing methodologies [10]. Human interaction is a main factor in web mining technology, to solving an issues such as context-sensitive and approximate equerries, summarization and deduction, and personalization and learning are of tremendous importance where soft computing methodologies are found to be the most appropriate platform for providing effective and powerful solutions to extracting and retrieval the web data.

In or around 1996, researcher has focused their attention in the field of soft computing technology and AI to develop the new field of computing and giving a name "soft web mining" systems in parallel with the conventional web mining technology. This combine concept of soft web mining is very much helpful to develop very effective and attractive system with online resources.

CONCLUSION:

Web mining is the application of data mining techniques to extract knowledge from online resources. Web mining has many advantages which makes web technology attractive. Soft computing is the combination of various methodologies which is very helpful to found out appropriate and meaningful data. Web technology and Soft computing methodologies in combination gives

the power in the AI research field. We explained and focused on some useful methodologies of soft computing i.e. fuzzy logic, neural network, genetic algorithm and rough set theory which are helpful in research activities and learners in the field of web technologies and artificial intelligence.

REFERENCES:

[1] J. Lee and E. Kim, "Fuzzy Web Information Retrival System".

[2] K. Nowacka, S. Zadrozny and J. Kacprzck, "A new fuzzy logic based information retrival model," in *Proceedings of IPMU'08*, Torremolinos(Malaga), Jun 22 -27, 2008.

[3] David E. Losada, F. Daiz-Hermida and. A. Bugarin, "Semi-fuzzy Quantifiers for Information Retrieval," in *Soft Computing in Web Information Retrieval Models and Applications*, Enrique Herrera-Viedma, Gabriella Pasi, Fabio Crestani (Eds.), 2006, pp. 195 -220.

[4] O. Zamir and O. Etzioni, "Web document clustering: A feasibility demonstration," in *21st Annual International ACM SIGIR Conference*, Melbourne,Australia, 1998.

[5] R. Krishnapuram, A. Joshi and L. Yi, "A fuzzy relative of the k-medoids algorithm with application to document and snippet clustering," in *IEEE International Conference on fuzzy System-FUZZIEEE 99*, korea, 1999.

[6] M. Kathuria, N. Duhan and C. K. Nagpal, "Application of Fuzzy Logic in Web Mining Domain: A survey," *International Journal of Advanced research in IT and Engineering IJARIE,* vol. 1, no. 3, 2012.

[7] S. Shehata, F. Karrey and M. S. Karnel, "An Efficient Concept-Besed mining model for Enhancing Text Clustering," *IEEE Transacation on Knowledge and Data Engineering ,* vol. 22, no. 10, 2010.

[8] T. Kanungo, D. M. Mount, N. Netanyahu, C. D. Pitako, R. Silverman and A. Y. Wu, "An Efficient K-means Clustering Algorithm: Analysis and Implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 7, 2002.

[9] L. Baltruns and J. Gordevicius, "Implementation of CURE Clustering Algorithm," *SIKMOD Seattle ACM,* 2005.

[10] T. Zhang, R. Ramkrishnan and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," *SIGMOD, ACM,* 1996.

[11] S. Song and C. Li, "Improved ROCK for text Clustering using Asymmetric Proximity," *SOFSEM - 2006,* pp. 501 - 510, 2006.

[12] M. Dimitrijevic and Z. Bosnjak, "Web Usage Association Rule Mining System," *Interdisciplinary Journal of Information, Knowledge, and Management,* vol. 6, no. 2, pp. 137 - 150, 2011.

[13] J. Shavlik and G. G. Towell, "Knowledge-based artificial neural networks," *Artificial Intelligence,* vol. 70, no. 1/2, pp. 119-165, 1994.

[14] M. Boughanem, T. Dkaki, J. Mothe, and C. Soule-Duppy, "Mercure at trec7," in 7th International Conference on Text Retrieval, TREC7, Gaithrsburg,MD, 1998.

[15] T. Kohonen, "Self-organizing maps for large documents," IEEE Transaction on Neural Networks, vol. 11, no. Special issue on Data Mining , pp. 574 - 589, 2000.

[16] H. Rana and M. Patel, "A study of web log analysis using clustering techniques," International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, no. 4, pp. 925 - 929, 2013.

[17] M. Pazzani, J. Muramatsu and D. Billsus, "Syskill and webert: idetifying intresting web sites," in Thirteenth National Conference on AI, 1996.

[18] C. Drummond, D. Ionescu and R. Holte, "A learning agent that assists the browsing of software libraries," University of Ottawa, 1995.

[19] A. F. Ali and M. H. Marghny, "Web mining based on genetic algorithm," in AIML 05 Conference, Cairo, Egypt, 2005.

[20] A. Asllani and A. Lari, "Using genetic algorithm for dynamic and

multiple criteria website optimization," European Journal of Operation Search, pp. 1767 - 1777, 2007.

[21] Jin Cheng, W. Chen, L. Chen and Y. Ma, "The improvement of Genetic algorithm searching performance," in First International Conference on Machine Learning and Cybernetics, Beijing, 2002.

[22] K. P. Rao and G. K. Chakravarthy, "AIntelligence Service of Web Mining with Genetic Algorithm," International Journal of Engineering Trends and Technology (IJETT), vol. 4, no. 10, 2013.

[23] C. Romero, S. Ventura, C. deCasto, W. Hall and M. H. Ng, "Using Genetic Algorithms for Data Mining in web Based Educational Hypermedia Systems," in AH - 2002 Workshop on Adaptive System for Web based Education, 2002.

[26] U. Straccia, "A framework for the retrieval of multimedia objects based on four-valued fuzzy description logics," in Soft Computing in Information Retrieval: Techniques and Applications, Heidelberg, Physica Verlag, 2000, pp. 332 - 357.

[27] Y. Y. Yao, Y. Zhao and R. B. Magire, "Explanation Oriented Association Mining using Rough Set Theory," Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada, 2003.

[28] H. H. Inbarani and K. Thangavel, "Rough set based User profiling for Web Personalization," International Journal of Recent Trends in Engineering, vol. 2, no. 1, pp. 103 - 107, 2009.

[29] S. Rissino and G. L. Torres, "Rough Set Theory - Fundamental Concepts, Principal, Data Extraction and applications," in Data Mining and knowledge Discovery in Real Life Applications, Vienna, Austria, I - Tech Education and Publishing, 2009, pp. 35 - 60.

[30] D. Nauck and R. Kurse, "Neuro-Fuzzy Systems for Function Approximation," in 4th International Workshop on Fuzzy- Neuro System, 1997.

[31] D. Nauck, F. Klawn and R. Kruse, Foundations of Neuro-Fuzzy Systems, J. Wiley & Sons, 1997.

[32] K. K. Aug and C. Quck. [Online]. Available: www.igi-global.com/chapter/rough-set-based-neuro-fuzzy/10422.

[33] U. Keerthika, R. Sethukkarai and A. Kannan, "A rough set based fuzzy inference system for mining temporal medical databases," International Journal on Soft Computing (IJSC), vol. 3, no. 3, pp. 41 - 54, 2012.

[34] K. Y. Shen and G. H. Tzeng, "DRSA based neuro-fuzzy Inference Systems for the Financial Performance prediction of Commercial Banks," International Journal of Fuzzy Systems, vol. 16, no. 2, pp. 173 - 183, 2014.

[35] S. Ding, J. Chen, X. Xu and J. Li, "Rough Neural Networks: A Review," Journal Of Computational Information Systems, vol. 7, no. 7, pp. 2338 - 2346, 2011.