Bio-Inspired Fuzzy Expert system for Mining Big data

Mr. A. SENTHIL KARTHICK KUMAR[1] , Dr. M. THANGMANI[2] & Dr. A.M.J MOHAMED ZUBAIR RAHMAN[3]
[1]Assistant Professor, Department of Computer Applications, Nehru Institute of Information Technology & Management, Coimbatore-641 105, Tamilnadu, India.
[2]Assistant Professor, Department of Computer Science, Kongu Engineering College, Perundurai -638 052, Erode District, Tamilnadu, India.
[3]Principal, Al-Ameen Engineering College, Erode, Tamilnadu, India. Pin code-638104
karthickmcamba@gmail.com, manithangamani2@gmail.com, mdzubairrahman@gmail.com

Abstract: - The increase in number of documents worldwide increases the difficulty for classifying those documents according to these needs. Cluster availability of large quantity of text documents from the World Wide Web and business document management forms has made the dynamic separation of texts into new categories as a very important task for every business intelligence systems. But, present text clustering algorithms still suffer from problems of practical applicability. Recent studies have shown that, in order to improve the performance of document clustering, ontologies are useful. Expert system (ontology) is nothing but the conceptualization of a domain into an individual identifiable format, but machine-readable format containing entities, attributes, relationships and axioms. By analyzing all types of techniques for document clustering, a clustering technique depending on Genetic Algorithm (GA) is determined to be better. The evaluation result shows that the proposed approach is very significant in clustering the documents in the distributed environment.

Key-Words: - Fuzzy ontology, clustering, distributed clustering, Peer to peer network.

1 Introduction

Genetic Algorithm (GA) is considered to provide better clustering results. The convergence time for the usage for GA is more and also the number of iterations required for GA is more when compared to other techniques. For enriching the performance measure, in this paper, the fuzzy ontology is applied to the database to reduce the convergence time and number of iterations before using GA. The usage of fuzzy ontology will provide better classification of a large, vague database. This has motivated the usage of fuzzy ontology and GA for clustering. Hence, in this approach, fuzzy ontology is combined with the GA to yield better classification accuracy for large databases.

In this paper, ontology [1] is introduced as a modeling technology for structured metadata definition within document clustering system. Documents can be clustered with the metadata obtained using the Genetic Algorithms (GAs) [2]. GA is a search technique based on natural genetic, selection and merging of survival of the fittest with structured interchanges. It conserves the attributes of finest exponents of a generation for use in the next generation; additionally introducing the variations in the new generation composition with the help of cross over and mutation function. GA [3] is a famous technique for handling complex search problems through implementing an evolutionary stochastic search because GA can be very effectively applied to various challenging optimization problems.

2. Related Works

Lena Tenenboim *et al,* [4] proposed ontology based classification for document clustering. The author recommended classification of news items in an ePaper, a prototype system of a future personalized newspaper service on a mobile reading device. The ePaper system comprises news items from different news suppliers and distributes to each subscribed user a personalized electronic newspaper, making use of content-based and collaborative filtering techniques. As classical Euclidean distance metric could not create a suitable separation for data lying in manifold, a GA based clustering method based on geodesic distance measure was proposed by Gang Li *et al,* [5]. In the proposed method, a prototype-

based genetic illustration is used, where every chromosome is a sequence of positive integer numbers that indicate the k-medoids.

Casillas *et al,* [6] also put forth a concept of document clustering using GA. It deals with document clustering that computes an approximation of the optimum k value and resolves the best clustering of the documents into k clusters. It is experimented with sets of documents that are the output of a query in a search engine. Andreas *et al,* [7] advocated text data clustering technique. Text clustering, usually, involves clustering in a high dimensional space that appears complex with regard to all virtual practical settings. Additionally, a scrupulous clustering outcome is provided.

Word sets based document clustering algorithm for large datasets was proposed by Sharma *et al.,* [8]. Document clustering is a significant tool for use in search engines and document browsers. It facilitates the user to have a better overall observation of the data available in the documents. There is also a strong requirement for hierarchical document clustering [9] where clustered documents can be browsed based on the increasing specificity of topics. Frequent Itemset Hierarchical Clustering (FIHC) is used for hierarchical grouping of text documents. This technique does not provide consistent clustering results when the number of frequent sets of terms is large. In this paper, the authors proposed Wordsets-based Clustering (WDC), an efficient clustering technique based on closed words sets. WDC makes use of hierarchical technique to cluster text documents having common words.

Cao *et al.,* [10] provided fuzzy named entity-based document clustering. Conventional keyword-based document clustering methods have restrictions because of simple treatment of words and rigid partition of clusters. Named entities are introduced as objectives into fuzzy document clustering, the important elements of defining document semantics and in many cases are of user concerns. Zhang *et al.,* [11] gave clustering aggregation based on GA for documents clustering. A technique based on GA for clustering aggregation difficulty, named as GeneticCA, is provided to approximate the clustering performance of a clustering division. In this case, clustering precision is defined and features of clustering precision are considered. Web document clustering using document index graph is put forth by Momin *et al.,* [12]. Document clustering methods are generally based on single term examination of document data set. To attain more precise document clustering, more informative features like phrases are essential.

Muflikhah *et al.* [13] proposed a document clustering based on concept space and cosine similarity measurement. This technique aims at incorporating the information retrieval and document clustering into concept space approach. It is known as Latent Semantic Index (LSI) because it uses Singular Vector Decomposition (SVD) or Principle Component Analysis (PCA). Its objective is to decrease the matrix dimension by identifying the pattern in document collection with reference to the terms. Affinity-based similarity measure for Web document clustering is presented by Shyu *et al.,* [14]. Document clustering is extended into Web document clustering by establishing affinity based similarity measure, It makes use of the user access patterns in finding the similarities among Web documents through a probabilistic model. Various experiments are conducted for evaluation with the help of real data set. The experimental results illustrate the fact that similarity measure outperforms the cosine coefficient and the Euclidean distance technique under various document clustering techniques.

ELdesoky *et al.,* [15] gave a similarity measure for document clustering based on topic phrases. In the conventional vector space model (VSM), researchers have used unique word available in the document set as the candidate feature. Currently, phrase based informative feature is considered because it contributes to enhancing the document clustering accuracy and effectiveness. Similarity measure of the traditional VSM is evaluated by considering the topics phrases of the document as the comprising terms instead of the conventional term. Thangamani et. al examined document clustering [16,17] in individual and peer to peer environment and also developed the system for automatic extraction and classification of document using multi domain ontology.

A document clustering method based on hierarchical algorithm with model clustering is presented by Haojun *et al*., [18]. It analyzes and makes use of cluster overlapping to design cluster merging criterion. Document clustering with fuzzy c-mean algorithm is proposed by Thaung *et al.,* [19]. Most traditional clustering technique allocates each data to exactly single cluster, therefore creating a crisp separation of the data provided. However, fuzzy clustering permits for degrees of membership to which data fit into various clusters. Xindong et. al [20] investigated as data mining techniques can applied to big data set for information extraction.

ISBN: 978-960-474-372-8

3. Expert system for distributed textual clustering

Initially, ontology generation using fuzzy logic is implemented to the database containing a large amount of documents. This technique generates the ontology for the given database. With this ontology, the next step is the application of GA. GA is used for clustering the documents in the database with the help of ontology generated by fuzzy logic technique. The combination of expert system generation using Fuzzy Logic and GA helps to increase the accuracy of clustering. It consists of the following modules.

Formal concept analysis using fuzzy: In fuzzy formal concept analysis integrates fuzzy logic into formal concept analysis to represent vague data. A fuzzy formal context shown in Table 1 consists of three objects which denote three documents. Those objects are named as D1, D2 and D3. Moreover, it has three attributes such as Data Mining, Clustering and Fuzzy Logic indicating the three titles. A membership value between 0 and 1 denotes the relationship between an object and an attribute. To remove the relationships that have low membership values, a confidence threshold T is introduced. Table 2 represents the fuzzy formal context provided in Table 1 with confidence threshold T as 0.5. Usually, the attributes of a formal concept can be considered as the description of the concept. Thus, the relationships between the object and the concept must be the separation of the relationships between the objects and the attributes of the concept. A membership value in fuzzy formal context denotes all the relationship between the object and an attribute. Then based on fuzzy theory, the intersection of these membership values must be the minimum of these membership values. Figure 1 one shows the automatic generation of expert system for analyzing distributed textual clustering.

Table 1 Fuzzy formal context

	Data Mining	Clustering	Fuzzy Logic
D1	0.75	0.3	0.5
D2	1	0.75	0.25
D3	0.25	0.25	0.75

Table 2 Fuzzy formal context for table 1 with T =0.5

	Data Mining	Clustering	Fuzzy Logic
D1	0.75	-	0.5
D2	1	0.75	-
D3	-	-	0.75

Expert system generation: While the formal concepts are also generated mathematically, distinct formal concepts are created on the basis of the difference in terms of attribute object and the traditional concept lattice. This produces the effect of concepts as interpreted by humans. Based on this observation, a cluster formal concept is infused into conceptual clusters of fuzzy conceptual clustering. **Class Mapping**: In this process, the extent and intent of the fuzzy context are mapped into the extent and intent classes of the ontology. It requires supervised training to name the label for the extent class. Keyword attributes can be represented by appropriate names and they are used to label the intent class names also.

Taxonomy relation generation: With concept hierarchy in place, this phase produces the intent class of the ontology as a hierarchy of classes. The step can be considered as an isomorphic mapping from the concept hierarchy into taxonomy classes of the ontology. **Non-taxonomy relation generation**: This step involves generating the similarities among the extent class and intent classes with no hierarchy between classes. The generation will mean an equivalent class with no sub class or super class. **Instances generation**: In this process, instances for the extent class are generated. Each instance indicates an object in the initial fuzzy context. Depending on the data existing on the fuzzy concept hierarchy, instances attributes are automatically furnished with suitable values. For example, each instance of the class document, related to an actual document, will be associated with the appropriate research areas. After the ontology is generated, GA is used to cluster the documents. The usage of ontology helps in determining the best classification for clustering using GA.

ISBN: 978-960-474-372-8

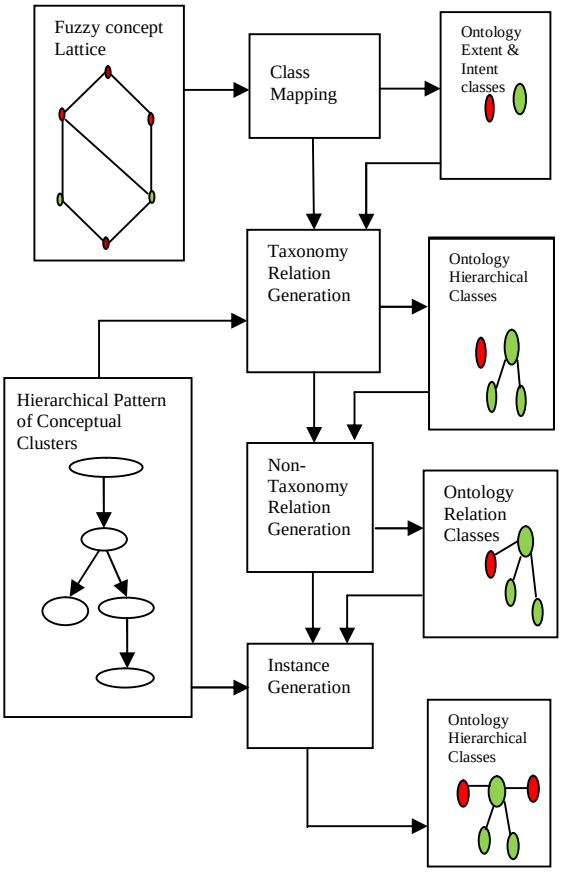

Figure.1: Expert system for distributed document clustering

4. Experiment discussions

This technique avoids getting stuck into a local maximum from which one cannot escape reaching a global maximum. This is one of the main benefits of GA in opposition to the conventional search techniques as the gradient technique. Another advantage is the utility of GA for real time applications, in spite of its inability to offer the optimal solution to the problem. However, it provides almost a better solution in a shorter time, including complex problems.

5. Conclusion and future work

In this work, expert system is created with help of bio inspired method for large data. In further investing the expert system to be applied to the different dataset to test how much this system will effective. Also system can be applied in cloud environment with e-Learning activities.

References:

[1] Andreas Hotho., Alexander Maedche. and Steffen Staab., "Ontology-based Text Document Clustering".

[2] Banerjee, A. and Louis, S.J., "A Recursive Clustering Methodology using a Genetic Algorithm", IEEE Congress on Evolutionary Computation, 2007, Pp. 2165-2172.

[3] Murthy, C.A. and Chowdhury, N., "In Search of Optimal Clusters using Genetic Algorithms", Pattern Recognition Letters, 1996, Pp. 825–832.

[4] Lena Tenenboim., Bracha Shapira. and Peretz Shoval., "Ontology-Based Classification of News in an Electronic Newspaper", International Book Series Information Science and Computing, 2008, Pp: 89-98.

[5] Gang Li., Jian Zhuang., Hongning Hou. and Dehong Yu., "A Genetic Algorithm based Clustering using Geodesic Distance Measure", IEEE International Conference on Intelligent Computing and Intelligent Systems, 2009, Pp: 274 – 278.

[6] Casillas, A., Gonzalez de Lena, M.T. and Martínez, R., "Document Clustering into an Unknown Number of Clusters Using a Genetic Algorithm", Lecture Notes in Computer Science, Vol. 2807, 2003, Pp. 43-49.

[7] Andreas Hotho., Alexander Maedche. and Steffen Staab., "Ontology-based Text Document Clustering", Journal on Kunstliche Intelligenz, Vol. 4, 2002, Pp. 48-54.

[8] Sharma, A. and Dhir, R., "A Wordsets based Document Clustering Algorithm for Large datasets", Proceeding of International Conference on Methods and Models in Computer Science, 2009.

[9] Koller, D. and Sahami, M., "Hierarchically Classifying Documents using Very Few Words", Proceedings of the 14th International Conference on Machine Learning (ML), 1997, Pp. 170-178.

[10] Cao, T.H., Do, H.T., Hong, D.T. and Quan, T.T.; "Fuzzy Named Entity-Based Document Clustering", IEEE International Conference on Fuzzy Systems, 2008, Pp. 2028 – 2034.

[11] Zhenya Zhang., Hongmei Cheng., Shuguang Zhang., Wanli Chen. and Qiansheng Fang., "Clustering Aggregation based on Genetic Algorithm for Documents Clustering", IEEE

Congress on Evolutionary Computation, 2008, Pp. 3156 – 3161.
[12] Momin, B.F., Kulkarni, P.J. and Chaudhari, A., "Web Document Clustering Using Document Index Graph", International Conference on Advanced Computing and Communications, 2006, Pp. 32 – 37.
[13] Muflikhah, L. and Baharudin, B., "Document Clustering Using Concept Space and Cosine Similarity Measurement", International Conference on Computer Technology and Development, Vol.1, 2009, Pp. 58-62.
[14] Shyu, M.L., Chen, S.C., Chen, M. and Rubin, S.H., "Affinity-based similarity measure for Web document clustering", IEEE International Conference on Information Reuse and Integration, 2004, Pp. 247 – 252..
[15] ELdesoky, A.E., Saleh, M. and Sakr, N.A., "Novel Similarity Measure for Document Clustering based on Topic Phrases", International Conference on Networking and Media Convergence, 2009, Pp. 92-96.
[16] Thangamani .M and Thangaraj .P,"Survey on Text Document Clustering", International Journal of Computer Science and Information Security, Vol.8(4),2010.
[17] Thangamani.M and Thangaraj.P "Effective fuzzy semantic clustering scheme for decentralized network through multidomain ontology model", International Journal of Metadata, Semantics and Ontologies, Interscience Vol.7, Issue 2, 2012, pp.131-139, Interscience publication
[18] Haojun Sun., Zhihui Liu. and Lingjun Kong., "A Document Clustering Method Based on Hierarchical Algorithm with Model Clustering", 22nd International Conference on Advanced Information Networking and Applications, 2008, Pp. 1229 – 1233.
[19] Thaung Win. and Lin Mon., "Document clustering by fuzzy c-mean algorithm", 2nd International Conference on Advanced Computer Control (ICACC), 2010, Pp.239 – 242.
[20] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE, "Data Mining with Big Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, 2014, Pp. 97- 107.

About Authors:

A.Senthil Karthick Kumar is a. Research Scholar in Bharathiar University, Coimbatore. He completed his B.Sc in Information Technology from Madurai Kamaraj University in 2003. Did his MCA from Bharathiar University in 2006; Completed his M.Phil in Computer Science in 2009 and E.M.B.A in Human Resource Management, from MS University 2012. Prior to joining in NIITM he worked for 3years as a Human Resource Executive (Technical) in various companies like Perot Systems, Bangalore. Currently he is working as an Assistant Professor in Nehru Institute of Information Technology and Management, Coimbatore Affiliated to Anna University. He enrolled his Life time Membership with ISTE, Member in CSI and IAENG. He has published and presented around 10 papers in National and International level Seminars and Journals. His area of interest in research includes Cloud computing, Software Engineering, Data mining and E-learning.

Dr. M. Thangamani completed her B.E., from Government College of Technology, Coimbatore, India. She completed her M.E., and PhD (Computer Science and Engineering) from Anna University, Chennai, India. Currently, she is working as Assistant Professor in the Department of Computer Science and Engineering, Kongu Engineering College, Tamil Nadu, India. She has published 20 articles in International journals and presented papers in 46 National and International conferences. She has published 11 books for polytechnic colleges and also guided many UG projects. She has delivered more than 30 Guest Lectures in reputed engineering colleges on various topics. She has organized many self supporting and sponsored National Conference and Workshop in the field of Data mining and Cloud computing. She also seasonal reviewer in IEEE Transaction on Fuzzy System, International journal of advances in Fuzzy System and Applied mathematics and information journals. She is also the editorial member for many International Journals and organizing chair for International conferences in India and other countries. Her research interests include Data mining; Cloud computing, Ontology development, Web Services and Open Source Software.

ISBN: 978-960-474-372-8

Dr. A.M.J Mohamed Zubair Rahman, Principal, Al-Ameen Engineering College, Erode. He is a Person with 22 Years of Teaching Experience and He was awarded with Ph.D from Anna University, Chennai in the year 2009. Add on to his academics excellence he completed his M.S. Software Systems from BITS-Pilani in the year 1996, He completed his B.E Computer Science Engineering from IRTT in 1989, further in continuation of his education he did his M.E Computer Science Engineering from Bharathiar University in 2002. To his credit he has attended several National and International Seminars and presented more than 20 papers in various conferences. He enrolled his Life time Membership with ISTE, and Member in CSI. He has published and presented around 20 papers in National and International Journals. His area of interest in research includes Data mining, Network Security, Software Engineering and E-learning.

ISBN: 978-960-474-372-8