Neuro Fuzzy Based Clustering For XML Document Mining: A Review

Ms. Vaishali S. Sarode[1], Prof. R. P. Sonar[2]
Department of computer science and engineering

Abstract

XML has increasingly become the language of choice for data representation, storage and exchange in many domains. With the standardization of XML as an information exchange language over the net, a huge amount of information is formatted in XML documents. As many paper presents different approaches for clustering and mining an XML documents. But in many cases, it does not provide efficient clustering as it requires much amount of time and Extracting information from documents is a very hard task, and is going to become more and more critical as the amount of digital information available on the Internet grows. In this paper we discuss different XML documents clustering algorithms and XML document mining algorithms. Then we discuss neuro fuzzy technique and how it provides better result for clustering an XML document.

1. Introduction

Extensible Mark-up Language (XML) has been recognized as a standard data representation for interoperability over the Internet and it is a flexible hierarchical model suitable to represent huge amounts of data with no absolute and fixed schema, and a possibly irregular and incomplete structure. Web pages formatted in XML have started to appear. Besides flat file storage, object-oriented databases, and native XML databases, developers have been using the more mature relational database technology to store data. Keyword search is proposed as an alternative means for querying XML data, which is simple and yet familiar to most Internet users as it only requires the input of keywords. Keyword search is a widely accepted search paradigm for querying document systems and the World Wide Web. One important advantage of keyword search is that it enables users to search information without knowing a complex query language such as XPath or XQuery, or having prior knowledge about the structure of the underlying data.

In this paper we propose a neuro fuzzy clustering algorithm for clustering XML documents, with the help of neuro fuzzy clustering algorithm documents will be group together into cluster of similar behaviour. Neuro fuzzy refers to the combination of artificial neural networks and fuzzy logic. Modern neuro-fuzzy systems are usually represented as special multilayer

feed forward neural networks. Both neural networks and fuzzy systems have some things in common. They can be used for solving a problem (e.g. pattern recognition, regression and density estimation). A neuro-fuzzy system based on an underlying fuzzy system is trained by means of a data-driven learning method derived from neural network theory. This heuristic only takes into account local information to cause local changes in the fundamental fuzzy system. It can be represented as a set of fuzzy rules at any time of the learning process, i.e., before, during and after. A neuro-fuzzy system is represented as special three-layer feed forward neural network. The first layer corresponds to the input variables. The second layer symbolizes the fuzzy rules. The third layer represents the output variables. The fuzzy sets are converted as (fuzzy) connection weights.

In this paper, we proposed neuro fuzzy clustering algorithm for clustering an XML documents. Document Mining algorithm for mining XML document. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. The remainder of this paper is organized as follow: section 2. Discuss previous work on XML document mining & clustering. Section 3 Section 4 Discuss the different XML document clustering techniques. Section 5 Discuss available document mining techniques. Finally, section 6 concludes this paper with some suggestions for further improvement.

2. RELATED WORK

In [1] presents a novel fuzzy clustering algorithm that operates on relational input data; i.e., data in the form of a square matrix of pair wise similarities between data objects. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework in which the graph centrality of an object in the graph is interpreted as likelihood. Results of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying

overlapping clusters of semantically related sentences, and that it is therefore of potential use in a variety of text mining tasks. The FRECCA algorithm was motivated in fuzzy clustering of sentence-level text, and the need for an algorithm which can accomplish this task based on relational input data.

In [3] data mining is commonly used in attempts to induce association rules from transaction data. They show that using the clustering technique to speed up the evaluation process can not only get nearly the same fitness values but also greatly reduce the execution time. The proposed approach can thus get a good trade-off between accuracy and execution time.

In [7] author introduces mining of XML data in which XML mining includes mining both the structure and the contents from XML documents. It determines different approaches for mining of static XML document and clustering of XML document. Mining of structure, which is essentially mining the XML schema, includes intra-structure mining (mining the structure inside an XML document) and inter-structure mining (mining the structures between XML documents).

In [8] author introduces learn about mining association rules from XML documents. This article also introduces the notion of dynamic association rules.

In [9] it introduces the task of mining association rules from an XML document. It also determines Clustering of dynamic (multi-version) XML documents. It described a method for clustering dynamic XML documents using a distance-based technique.

In [17] author proposed a hierarchical algorithm (S-GRACE) for clustering XML documents based on structural information in the data & proposed a framework for clustering XML document . It shows that clustering a large collection of XML documents by structure can alleviate the fragmentation problem of storing them into relational tables. S-GRACE algorithm on the real DBLP data set cannot be easily spotted by manual inspection.

In [15] author Proposed the technique is mainly based on the idea of representing an XML document as a time series. The structural similarity between two documents can be computed by exploiting the Discrete Fourier Transform of the associated signals. It determines similarity between xml documents in faster manner. It does not provide accuracy.

In [20] author proposed a novel clustering methodology based on the notion of keywords matching pattern, which clusters results according to the way they match the given query. Author investigates the problem of returning cluster-based search results for XML keyword search. Author does not measure the delays of clustering directly because there are no separate phases of clustering in our algorithms; instead, the clustering is pushed into the search process.

In [18] author proposed an efficient data mining algorithm i.e. a three-phase algorithm, which finds the minimal infrequent structures. But this technique operates slowly. By indexing the occurrences of MIS, author can efficiently locate the high-selective substructures of a query, improving search performance significantly. Time complexity is the main problem.

In [13] author describes an approach based on Tree-Based Association Rules (TARs): mined rules. it provides the method for deriving intentional knowledge from XML document. It provides approximate, intentional information on both the structure and the contents of Extensible Mark-up Language (XML) documents, and can be stored in XML format as well. As the number of nodes within document increases, the time required to extract the data also increases.

In [4] author proposed LCA (lowest common ancestor) based fuzzy type a head search method to retrieve information from XML documents. Author has implement method on real data sets, and experimental results show that method achieves high search efficiency and result quality. Time complexity is the main problem.

In [5] author presented a framework for data structure-guided association rules extraction from XML trees. Author can efficiently locate the high-selective substructures of a query, improving search performance significantly**.**

In [14] this paper conceptualizes a novel data mining technique, genetically guided cluster based fuzzy decision tree (GCFDT), is introduced for the mining task.

In [16] author proposes fuzzy c-means (FCM) clustering to VL data. Clustering algorithms that scale well to VL data are important and useful.

3. ARTIFICIAL NEURAL NETWORK

Artificial neural network (ANN) is an information processing method inspired by biological nervous systems. An ANN uses interconnected processing nodes computationally linked to solve problems.

Neural networks are frequently used for pattern recognition and document classification and learn by using training data to adjust the weights between connecting nodes. Some research has applied artificial neural networks to text classification. ANNs are constructed with layers of units and thus are termed multilayer ANNs. A layer of units in such an ANN is composed of units that perform similar tasks. First layer of a multilayer ANN consists of input units. These units are as independent variables in statistical literature. Last layer contain output units. in statistical nomenclature, these units are known as dependent or response variables. All other units in the model are called hidden units and constitute hidden layers. In figure 1, all signals and weights are real numbers. The input neurons do not change the input signals so their output is the same as their input.

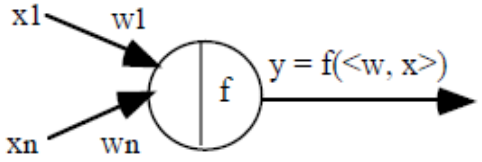

Figure 1: A simple neural network

It employs multiplication, addition, and sigmoid *f*, will be called as regular (or standard) neural net. Sigmoid function used to compute the output.

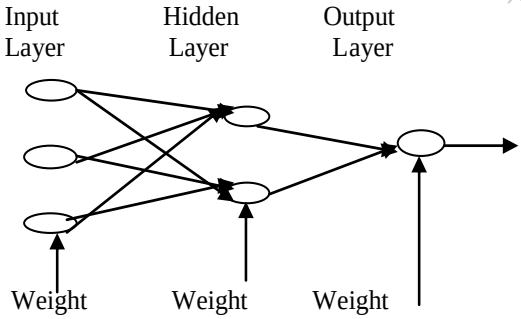

Figure 2: The architecture of a neuro- fuzzy system

A neuro-fuzzy system is represented as special three-layer feed forward neural network as it is shown in Figure 2. The first layer corresponds to the input variables. The second layer symbolizes the fuzzy rules. The third layer represents the output variables. The fuzzy sets are converted as (fuzzy) connection weights. Some approaches also use five layers where the fuzzy sets are encoded in the units of the second and fourth layer, respectively. However, these models can be transformed into three-layer architecture. A neuro-fuzzy system based on an underlying fuzzy system is trained by means of a data-driven learning method derived from neural network theory. This heuristic only takes into account local information to cause local changes in the fundamental fuzzy system. It can be represented as a set of fuzzy rules at any time of the learning process, i.e., before, during and after. Thus the system might be initialized with or without prior knowledge in terms of fuzzy rules. The learning procedure is constrained to ensure the semantic properties of the underlying fuzzy system. A neuro-fuzzy system approximates a n-dimensional unknown function which is partly represented by training examples. Properties are as follows

1. A neuro-fuzzy system is based on a fuzzy system which is trained by a learning algorithm derived from neural network theory. The (heuristically) learning procedure operates on local information, and causes only local modifications in the underlying fuzzy system.

2. A neuro-fuzzy system can be viewed as a 3-layer feed forward neural network. The first layer represents input variables, the middle (hidden) layer represents fuzzy rules and the third layer represents output variables. Fuzzy sets are encoded as (fuzzy) connection weights. It is not necessary to represent a fuzzy system like this to apply a learning algorithm to it. However, it can be convenient, because it represents the data flow of input processing and learning within the model.

3. A neuro-fuzzy system can be always (i.e.\ before, during and after learning) interpreted as a system of fuzzy rules. It is also possible to create the system out of training data from scratch, as it is possible to initialize it by prior knowledge in form of fuzzy rules.

4. The learning procedure of a neuro-fuzzy system takes the semantically properties of the underlying fuzzy system into account. This results in constraints on the possible modifications applicable to the system parameters.

4. XML DOCUMENT CLUSTERING TECHNIQUE

The structure of an XML document is characterized by elements, which are denoted by start-tags and end-tags. The basic idea of the clustering methodology is to cluster search results based on the ways they match keyword queries. Wang Lian, David Wai-lok Cheung, Nikos Mamoulis, and Siu-Ming Yiu proposed hierarchical algorithm (S-GRACE) for clustering XML documents based on structural information in the data. The notion of structure graph (s-graph) is proposed, supporting a computationally efficient distance metric defined between documents and sets of documents.

This simple metric yields our new clustering algorithm which is efficient and effective, compared to other approaches based on tree-edit distance. For two XML documents C1 and C2, the distance between them is defined by

$$dist(C_1, C_2) = 1 - \frac{|sg(C_1) \cap sg(C_2)|}{max\{|sg(C_1)|, |sg(C_2)|\}}$$

It is straightforward to show that distðC1; C2Þ is a metric. If the number of common sub element relationships between C1 and C2 is large, the distance between the s-graphs will be small, and vice versa. It shows that clustering a large collection of XML documents by structure can alleviate the fragmentation problem of storing them into relational tables. S-GRACE algorithm on the real DBLP data set cannot be easily spotted by manual inspection.

Sergio Flesca, Giuseppe Manco, Elio Masciari, Luigi Pontieri, and Andrea Pugliese present an approach for detecting structural similarity between XML documents which significantly differs from standard methods based on graph-matching algorithms, and allows a significant reduction of the required computation costs. Author proposed the technique is mainly based on the idea of representing an XML document as a time series. The structural similarity between two documents can be computed by exploiting the Discrete Fourier Transform of the associated signals. Structure of an XML document represented as time series in which each occurrence of a tag in a given context correspond to an impulse. It determines similarity between xml documents in faster manner. It does not provide accuracy.

Xiping Liu, Changxuan Wan, and Lei Chen first proposed a novel semantics for answers to an XML keyword query. The core of the semantics is the conceptually related relationship between keyword matches, which is based on the conceptual relationship between nodes in XML trees. Then, propose a new clustering methodology for XML search results, which clusters results according to the way they match the given query. Author investigates the problem of returning cluster-based search results for XML keyword search**.** Author does not measure the delays of clustering directly because there are no separate phases of clustering in our algorithms; instead, the clustering is pushed into the search process.

Andrew Skabar and Khaled Abdalgader present a novel fuzzy clustering algorithm that operates on relational input data; i.e., data in the form of a square matrix of pairwise similarities between data objects. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework in which the graph centrality of an object in the graph is interpreted as likelihood. Results of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences, and that it is therefore of potential use in a variety of text mining tasks. Author proposed Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA).

5. XML DOCUMENT MINING TECHNIQUE

Jianhua Feng and Guoliang Li study fuzzy type-ahead search in XML data, a new information-access paradigm in which the system searches XML data on the fly as the user types in query keywords. Author proposed LCA (lowest common ancestor) based fuzzy type ahead search method to retrieve information from XML documents. It allows users to explore data as they type, even in the presence of minor errors of their keywords. Study research challenges in this new search framework. Author proposes effective index structures and top-k algorithms to achieve a high interactive speed and examine effective ranking functions and early termination techniques to progressively identify the top-k relevant answers. Implemented method on real data sets and the experimental results show that method achieves high search efficiency and result quality. First find the trie node corresponding to this keyword by traversing the trie from the root. Then locate the leaf descendants of this node, and retrieve the corresponding predicted words and the predicted XML elements on the inverted lists. Time complexity is the main problem.

Wang Lian, Nikos Mamoulis, David Wai-lok Cheung and S.M. Yiu proposed an efficient data mining algorithm, which finds the minimal infrequent structures. Their occurrences in the XML data are then indexed by a lightweight data structure and used as a fast filter step in query evaluation. Author validates the efficiency and applicability of methods through experimentation on both synthetic and real data. Three phase data mining algorithm achieves a significant improvement compared to the direct use of the Apriori algorithm. By indexing the occurrences of MIS, author can efficiently locate the high-selective substructures of a query, improving search performance significantly. Time complexity is the main problem.

Mirjana Mazuran, Elisa Quintarelli, and Letizia Tanca describe an approach based on Tree-Based Association Rules (TARs): mined rules, which provide approximate, intentional information on both the structure and the contents of Extensible Mark-up Language (XML) documents, and can be stored in XML format as well. Once the mining process has finished and frequent TARs have been extracted, they are stored in XML format. This decision has been taken to allow the use of the same language for querying both the original data set and the mined rules. When users specify queries without knowing the document structure, they may fail to retrieve information which was there. As the number of nodes within document increases, the time required to extract the data also increases**.**

6. CONCLUSION AND FUTURE WORK

XML document clustering is important for speed evaluation. In this paper, we provide an overview of existing work, introduce artificial neural network with their properties. Next we discuss XML document clustering techniques and then XML document mining techniques. In future, we implement artificial neural network for clustering an XML documents and keyword pattern matching algorithm for XML document mining. Further study is required to improve performance.

References

[1]Andrew Skabar, and Khaled Abdalgader, January 2013. Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm, IEEE Transaction on knowledge and data engineering, VOL. 25, NO. 1.

[2]Asli Celikyilmaz, and I. Burhan Turksen, *Fellow*, June 2008. Enhanced Fuzzy System Models with Improved Fuzzy Clustering Algorithm, IEEE Transaction on fuzzy system, VOL. 16, NO. 3.

[3] Chun-Hao Chen, Vincent S. Tseng, Tzung-Pei Hong, February 2008. Cluster-Based Evaluation in Fuzzy-Genetic Data Mining, IEEE Transaction on fuzzy system, VOL. 16, NO. 1.

[4] Jianhua Feng and Guoliang Li, May 2012. Efficient Fuzzy Type-Ahead Search in XML Data, IEEE Transaction on knowledge and data engineering, VOL. 24, NO. 5.

[5] Juryon Paik, Junghyun Nam, Won Young Kim, Joon Suk Ryu, Ung Mo Kim, November 2009. Mining Association Rules in Tree Structured XML Data, ACM.

[6] Kyong-Ho Lee, Yoon-Chul Choy and Sung-Bae Cho, August 2004. An Efficient Algorithm to Compute Differences between Structured Documents, IEEE Transaction on knowledge and data engineering, VOL. 16, NO. 8.

[7] Laura Irina RusuPhD, November 2011. XML data mining, part-1: Survey several approaches to XML data mining**.**

[8] Laura Irina RusuPhD, November 2011. XML data mining, Part 2: Mining XML association rules.

[9] Laura Irina RusuPhD, November 2011. XML data mining, Part 3: Clustering XML documents for improved data mining.

[10] Lifei Chen, Qingshan Jiang, and Shengrui Wang, July 2012. Model-Based Method for Projective Clustering, IEEE, VOL. 24, NO. 7.

[11] Li-Xin Wang and Chen Wei, August 2000. Approximation Accuracy of Some Neuro-Fuzzy Approaches, IEEE, VOL. 8, NO. 4.

[12] Massimo Panella and Antonio Stanislao Gallo, February 2005. An Input–Output Clustering Approach to the Synthesis of ANFIS Networks, IEEE, VOL. 13, NO. 1.

[13] Mirjana Mazuran, Elisa Quintarelli, and Letizia Tanca, Data Mining for XML Query-Answering Support, IEEE, VOL. 24, NO. 8, A.

[14] Sanjay Kumar Shukla and Manoj Kumar Tiwari, February 2012. GA Guided Cluster Based Fuzzy Decision Tree for Reactive Ion Etching Modeling: A Data Mining Approach, IEEE, VOL. 25, NO. 1.

[15] Sergio Flesca, Giuseppe Manco, Elio Masciari, Luigi Pontieri, and Andrea Pugliese, February 2005. Fast Detection of XML Structural Similarity, IEEE Transaction on knowledge and data engineering VOL. 17, NO. 7.

[16] Timothy C. Havens, Senior Member, *IEEE*, James C. Bezdek, *IEEE*, Christopher Leckie, Lawrence O. Hall, *IEEE*, and Marimuthu Palaniswami, *IEEE*, December 2012. Fuzzy *c*-Means Algorithms for Very Large Data, IEEE Transaction on fuzzy system, VOL. 20, NO. 6.

[17] Wang Lian, David Wai-lok Cheung, Member, IEEE Computer Society, Nikos Mamoulis, and Siu-Ming Yiu, January 2004. An Efficient and Scalable Algorithm for Clustering XML Documents by Structure, IEEE Transaction on knowledge and data engineering,, VOL. 16, NO. 1.

[18] Wang Lian, Nikos Mamoulis, David Wai-lok Cheung and S.M. Yiu, July 2005. Indexing Useful Structural Patterns for XML Query Processing, IEEE Transaction on knowledge and data engineering, VOL. 17, NO. 7.

[19] Wen-Chen Lih, Satish T. S. Bukkapatnam, Prahalad Rao, Naga Chandrasekharan, and Ranga Komanduri, January 2008. Adaptive Neuro-Fuzzy Inference System Modeling of MRR and WIWNU in CMP Process with Sparse Experimental Data, IEEE, VOL. 5, NO. 1.

[20] Xiping Liu, Changxuan Wan, and Lei Chen, December 2011. Returning Clustered Results for Keyword Search on XML Documents, IEEE Transaction on knowledge and data engineering, VOL. 23, NO. 12.

[21] Yun Chi, Yi Xia, Yirong Yang, and Richard R. Muntz, February 2005. Mining Closed and Maximal Frequent Subtrees from Databases of Labeled Rooted Trees, IEEE Transaction on knowledge and data engineering, VOL. 17, NO. 2.