Effective Methodology for XML Document Clustering using Lattice Cube Approach

T.Vinoth Kumar M.E., M.B.A. and K.Sampath Kumar M.E., (Ph.D).,

Abstract-*The semantic web is the widely focused research area in the recent era. Xml performs a major role in the Semi structured and structured xml documents are used in the semantic web environment. Document management is the complex task in the semantic web environment. Document management includes the Document Clustering and Indexing process. Different Clustering techniques like Bit cube clustering and Lattice cube clustering schemes are used for the document clustering task. The Lattice cube based clustering scheme is proposed for the system. The ontology is used to generate the concept clusters from the xml documents. The system is divided into four major modules. They are xml documents, path verification, Cube construction and Clustering Processes. The path verification module is designed to perform the xml document hierarchical path verification and concept extraction process. The cube construction module is designed to construct the bit cube and lattice cube. The clustering module is designed to cluster the documents using the structure and contents.. The lattice cube reduces the size of the cube in a considerable manner.*

Keywords— lattice cube, clustering, xml, hbsh, rdf,tf/idf, Bit map,cube construction.

1. INTRODUCTION

Hierarchical clustering techniques are used to cluster the XML documents based on its layout. Path elements are used in the structure based clustering method. Content based clustering is performing on the data values. The term value and its counts are used for the content clustering methods. Structure based clustering didn't focus on content values. Content based clustering didn't consider the layout. Clustering accuracy is very low. Processing time is high which degrades the overall performance of the system. The system is designed to perform XML document clustering using contents and structure. Path features used to analyze the XML document hierarchy. Content analysis is done with concept relationship. The system Performs bit cube clustering and lattice cube clustering operations to improve the clustering accuracy.

T.Vinoth Kumar M.E., M.B.A. is with Computer Science and Engineering RVS CET, Dindigul, Tamilnadu and K.Sampath Kumar M.E., (Ph.D). is with Computer Science and Engineering P.G.P.CET, Namakkal, Tamilnadu, Email: vinothmanikumar@gmail.com.

2. RELATED RESEARCH

The proposed algorithm for Massive Text Clustering by hash based technique. This study proposes a new fast hierarchical text-clustering algorithm (Hash-based Structure Hierarchical Clustering), which is suitable for massive text clustering. Yin luo and yan fu HBSH performs the text clustering process without setting clustering centre number and has minor space complexity in advance, which can achieve better performance. The experimental results illustrate that the average Time of HBSH is faster than that of traditional text clustering algorithms.

An Improved XML Document Clustering Using Path Feature to improve the performance. Extensible markup language (XML) documents Clustering is useful to xml application such as XML search engine. Based on this idea, Jin-sha yuan1, xin-ye li and li- na ma proposed we improved the path-based XML clustering algorithm. Experiments are described to demonstrate its efficiency. The proposed a system to manage the heterogeneous XML documents effectively and efficiently. The existing so-called semantic xml document clustering algorithms usually use a synonymous word library to calculate semantic similarities among xml documents a clustering algorithm DT^2K-means is also proposed to cluster xml documents. Chong Zhou and yansheng lu Empirical experiment results on real world data sets show DT^2K-means can group the semantic similar XML documents together correctly, which contain different tags but describe the same object. Dimitrios Katsaros, Alexandros Nanopoulos and Yannis manolopoulos proposed a system for theProblem of discovering frequently occurring structures in semi- structured objects using the notion of association rules. In the paper it states the problem of discovering frequently occurring structures in semi-structured objects using the notion of association rules.

The structured link vector model for xml document Clustering by Kernel Matrix. The rapid growth of XML adoption has urged for the need of a proper representation for semi-structured documents, Jianwu yang, william k. Cheung and xiaoou chen proposed where the document structural information has to be taken into account so as to support more precise document analysis In the paper, an xml document representation named "structured link vector model" is adopted, with a kernel matrix included for modelling the similarity between

XML elements. Three different rule-based systems, each designed to take xml as input and produce xml as output and manipulate intermediate facts as xml. They use very different methods of representing the XML during rule processing this paper explores rule-based methods that leverage RETE-based techniques for xml-to-xml transformations. Tuanjie Tong, GoEguchi, Jaehoon, John callahan and Laurence Leff proposed rule-based methods that leverage RETE-based techniques for XML-to-XML transformations. Rule-based systems and xml can be combined in two ways. First, rules can be used as a way of transforming and manipulating the contents of xml documents. Please check all m.baggi proposed system on ontology-based approximate filtering of xml data. In the paper, it describes a system, written in Haskell, for the ontology-based approximate filtering of xml data the system can be used through a Web application which is endowed with a user-friendly graphical interface. Finally, we provide some meaningful examples which show the usefulness of the implemented filtering methodology.

3. LATTICE CUBE APPORACH FOR XML DOCUMENT CLUSTERING.

The document indexing tool is divided into three major modules. They are the xml documents, document analysis and cube construction. The xml documents module is designed to maintain the xml documents. Document parsing and analysis are the major tasks performed in this module. The document analysis module is designed to extract the path and concept details from the documents. The cube construction module is designed to construct the bit cube and lattice cubes. The lattice cube is also constructed with two weight estimation methods.

3.1 XML Documents

The XML documents module is the initial phase in the system. This module is designed to handle the xml documents in the database. The system uses a collection of 1000 xml documents for the implementation. The documents are collected from the leading international journal IEEE. Recently published paper in the domain of data mining is collected for the system. The documents and its details are collected and stored as a xml file.
The documents details include the year of publish, volume number, issue of the article, author, pages, index terms and abstract. The xml documents are also provides the details about the tables and the images the article. The documents are parsed to extract the elements.

The system displays the documents in two ways. They are XML view and the parsed content view. The xml document view is designed to display the content of the xml documents without any process. The parsed content view displays the parsed content of the XML documents. All the elements are separated from the documents and assigned in the text fields. The ontology view shows the concepts and its relationship. The concepts, synonyms, metonyms and hyponyms are displayed in the ontology view. This system uses the ontology for the data mining domain.

3.2 Document Analysis

The document analysis module is designed to perform the concept extraction and pre-processing tasks. In the pre- processing each document is analyzed separately. The special character elimination is taken first. Then the stop word elementsare removed from the documents. After the stop word elimination process the stemming process is carried out. The path analysis is performed to eliminate the unpopular path columns. The columns that are not frequently appeared are removed from the path matrix.

The document analysis module is divided into four sub modules. They are bit map, refined bit map, lattice map and refined lattice map. The bit map is created with the column elements and its occurrence. The bit map is formed with binary values 0 and 1. The refined bit map is the simplified form of the bit map. The unpopular column elements are removed from the bit map and produces the refined bit map is produced. The bit map displays the sixteen path columns and the refined bit map displays 10 columns only. The author3, image and table columns are removed from the bit map. The lattice map is a digit form of the bit map. The author, tables and image columns are grouped into single columns. In the refined lattice map also the tables and image column elements are removed.

3.3 Cube construction

The cube construction module is designed to construction two types of cube for the indexing process. They are bit cube and lattice cube. The bit cube is formed with the support of the refined bit map. The lattice cube is formed with the help of the refined lattice map. The bit represents three dimensions. They are the document name, term name and term weight. In the same way the lattice cube represent three dimensions. They are document name, concept and terms and semantic weight. The lattice cube also shows the TF/IDF values. The lattice cube shows the semantic weight and TF/IDF value in the same cube.

The document indexing scheme is implemented and tested with a collection of 1000 xml documents. The XML documents are formed with the IEEE journal details. Each document represents the details about a single article. The document title, published year, volume number, pages, author, index terms, images, abstract and table details are maintained in the document. The documents are named with a unique name. All the document analysis and path analysis activities are carried out with the documents. The concept extraction process is performed with the ontology. In this system the ontology for the data mining domain is used.

The bit map and lattice map are formed with the terms and the concepts. The refined bit map and refined lattice map are formed after the unpopular path column elimination process. The bit map and the bit cube are constructed with the terms. The terms are extracted from the documents after completion of the stop words elimination process and stemming process. The bit map represent the terms that are identified in the documents. The figure 6.1.and table 6.1.shows the documents and terms ratio for

a set of intervals. 1000 documents contains 99603 terms. The terms are constantly increased with reference to the document count. The figure 6.2.and table 6.2 Shows the documents and concept ratio for the same document collection. 2552 concepts are extracted from the 1000 documents. The comparison between the figure 6.1 and 6.2.shows that the concept based scheme reduces the matrix size in a considerable way.

Documents	Terms
200	13450
400	42792
600	57239
800	75240
1000	99603

Table 6.1 Analysis of Document VS Terms

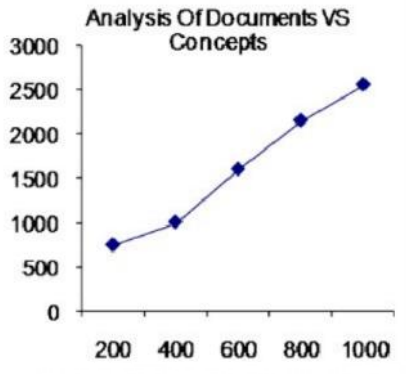
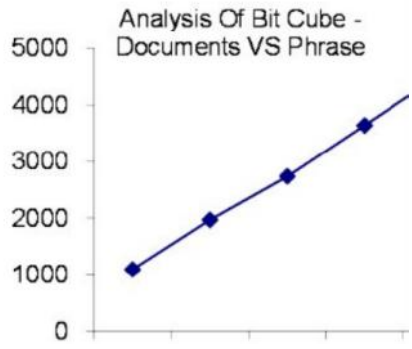


Table 6.2 Analysis of Bit Cube - Documents VS Phrase



Documents	Concepts
200	54
400	71
600	89
800	113
1000	134

Table 6.3 Analysis of Lattice Cube - Documents VS Concepts

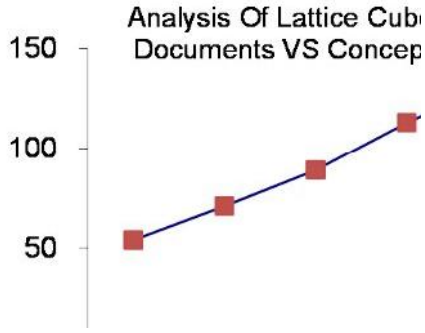
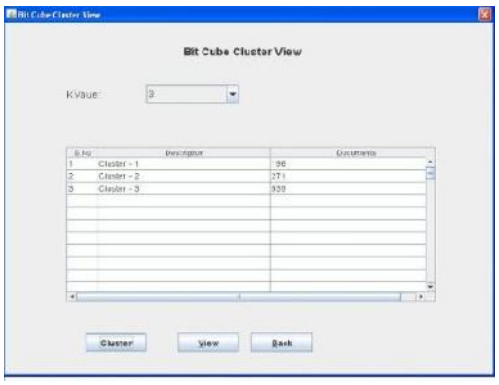


Figure 6.3 Analysis of Lattice Cube Documents VS Concept

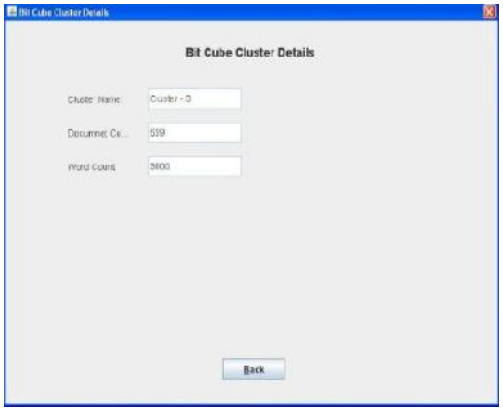
The cube analysis is performed with the documents for the bit cube and lattice cube process. The bit cube is constructed with the distinct terms and the lattice cube is formed with the distinct terms and concepts. The bit cube reduces the term count at a level of 4554 terms. The lattice cube reduces the distinct concept at the level of 134. The bit cube values are represented in figure
6.3. and table 6.3. The lattice cube values are shown in figure 6.4 and table 6.4. . The comparison shows that the lattice cube reduces the cube size very well.

IV. SYSTEM IMPLEMENTATION

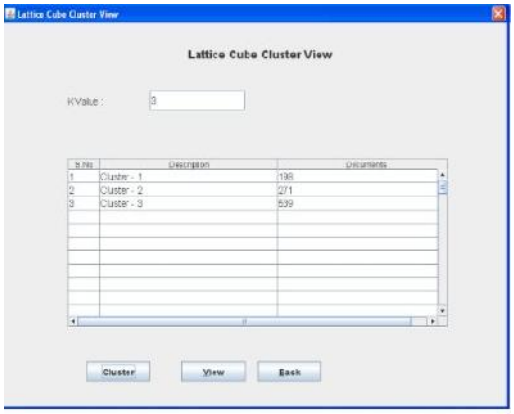
The concept cluster based indexing scheme for XML document is designed to perform the document management requirements. The XML documents are indexed with the semantic weight and/IDF values. The semantic weight is the suitable weight estimation mechanism for the XML documents. The system also applies the pre-processing and unpopular path column elimination techniques in the document indexing process. The technique improves performance of the document indexing process. The ontology is used to support the concept extraction process. The system is implemented as a tool to carry out the indexing process. The XML documents are parsed and performed a preprocessing operation. The stop word elimination and stemming process is performed during the preprocessing task. The pre-processed data is used for the index operation. 1000XML documents are used as a test bed for this system. The data mining based ontology is used for the system. All documents are collected in the data mining area from the IEEE journal. The system is tested and analyzed with the sample data set. The analysis results are given in the graphical form also. The results shows, that the concept cluster based indexing scheme produces better performance than the concept based and term frequency basedtechniquesThe system is implemented as a graphical user interface tool using the Java language and the Microsoft Access back end. The screens as follows.



Bit Cube Cluster View



Bit Cube Cluster Details



Lattice Cube Cluster View

Lattice Cube Cluster Details

V.CONCLUSION AND FUTURE

The XML document index system is developed to indexing and Analysis process on xml documents. The indexing process is performed with the ontology support. The system is developed as a graphical user interface based one. The future development of the system can include the following features. The current system performs the indexing operations using the lattice cube with the semantic weight. The future development can include the clustering process. The system can be enhanced with the document classification feature. The ontology used in the system is prepared in the xml format. The future development can be updated to use the RDF ontology. The indexing technique can be used with other document formats like text documents, rich text format documents and portable document formats. The system is implemented for the data mining domain based xml document collection. The future development can include other domain based documents like health care and scientific researches.

VI. REFERENCES

[1] Alexander Maedche, "Ontology Learning for the Semantic Web"Springer,2000.

[2] Christian Noon , Ruqin Zhang , Eliot Winer, James Oliver, Brian Gilmore, Jerry Duncan A system for rapid creation and assessment of conceptual large vehicle designs using immersive virtual reality, ELSEVIER.2012.

[3] Davood Rafiei, Daniel L. Moise and Dabo Sun, "Finding Syntactic Similarities between XML Documents" 17th International Conference on Database and Expert Systems Applications 2009.

[4] David A. Grossman and Ophir Frieder **"**Information Retrieval: Algorithms and Heuristics", Springer; 2 edition 2006.

[5] Gerardo Canfora, Luigi Cerulo and Rita Scognamiglio, "Measuring XML document similarity: a case study for evaluating Information Extraction Systems" 10th International Symposium on Software Metrics 2005.

[6] Ilhwan Choi, Bongki Moon, Hyoung-Joo Kim, and Science Direct A clustering method based on path similarities of XML data q, 2012.

[7] Joe Tekli, Richard Chbeir, Science Direct, A novel XML document structure comparison framework based-on sub- tree commonalities and label semantics, 2011.

[8] Jianwu Yang, William K. Cheung and Xiaoou Chen, "Integrating Element and Term Semantics for Similarity-Based XML " International Conference on Web Intelligence 2005.

[9] James Bean "XML for Data Architects: Designing Reuse and Integration" MorganKaufmann2003.

[10] Jan L.G. Dietz, "Enterprise Ontology: Theory and Methodology" Springer1edition2006.

[11] Krunal Patel and Kajal T. Claypool "SUSAX: Context- Specific Searching in XML Documents Using Sequence Alignment Techniques" 22nd International Conference on Data Engineering

Workshops 2006.
[12] Sun Wei and Liu Da-xin, "A Hybird Method for Efficient Indexing of XML Documents" International Workshop on Data Engineering Issues in E-Commerce 2005. Barbara Catania and Anna Maddalena, "XML Document Indexes: A Classification" the IEEE Computer Society 2005.
[13] W Scott Means Michael a Bodie **"**The Simple API for XML**"** No Starch Press, 1edition 2002.

Author[1]

T.Vinoth Kumar M.E., M.B.A., has put in
4 years of teaching experience. He is an active member in many technical bodies like ISTE, IAENG, IACSIT, He has published 4 articles in reputed journals and presented 2papers in national conferences. At present working as Assistant Professor in the department of MSc Software engineering Ratnavel Subramanian College of Engineering and technology Dindigul.Tamilnadu.

Author[2]

K.Sampath Kumar M.E., (Ph.D)., has put in
10 years of teaching experience. He is an active member in many technical bodies. And presenting various conferences and international conferences, seminars. At present working as Professor and head of the department of Compouter science and engineering P.G.P College of Engineering and technology Namakkal,Tamilnadu.