International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-1 Issue-8 March-2013

Efficient Frequent Pattern Mining Techniques of Semi Structured data: a Survey

Leena A Deshpande[1], R.S. Prasad[2]
Assistant professor, VIIT, Pune[1], Professor, Dnyanganga College of Engineering & Research[2]

Abstract

Semi-structured data are a huge amount of complex and heterogeneous data sets. Such models capture data that are not intentionally structured, but are structured heterogeneously. These databases evolve so quickly like run time report generated by ERPs, World-Wide Web with its HTML pages, text files, bibliographies, various logs generated etc. These huge and varied become difficult to retrieve relevant information User is often interested in integrating various formats (like in biomedical data text, image or structured) that are generally realized as files, and also wants to access them in an integrated fashion. Users not only query the data to find a particular piece of information, but he is also keen in knowing better understanding of the query. Because of this variety, semi-structured DBs do not come with a conceptual schema. To make these databases more accessible to users a rich conceptual model is needed. Traditional retrieving techniques are not directly applied on these databases. Unfortunately the tools and methodologies used for RDBMS do not give efficient results and so fail to bridge the gap. Hence efficient and scalable methods for mining the semi-structured data is needed, via discovering rule or patterns from the huge semi-structured databases. These databases are modelled by trees and graphs.

Keywords

Semi structured database, XML, Association rule, Classification, rule based association.

1. Introduction

1.1XML as semi structured data

XML stands for eXtensible Markup Language, which is a markup language for documents containing structured information. It is mainly used for exchanging wide variety of data on the web. The increasing popularity of XML is partly due to the limitations of the other two technologies: Hypertext Markup Language (HTML) and Standard Generalized Markup Language (SGML)for representing structured and semi-structured documents. And hence XML has been widely adopted for its flexible and self-describing nature. Various approaches have been discussed and sproposed for managing, storing, querying, and representing XML data generated from diverse and heterogeneous sources. The data may be structured, but the structure is not known to the user; many times user ignores the structure for browsing purposes. Sometimes structure may be implicit, like formatted text, as opposite to traditional databases which are rigid and regular. Thus XML database research has become increasingly popular with the emergence of the World Wide Web and the concept of ubiquitous computing. Due to such structure uncertainty is a critical issue to be handled. Uncertainties in XML documents are structured naturally assigned, and interpreted. It can be interpreted with the help of probabilities and reliabilities. Hence managing uncertain data in XML raises many challenging issues.

1.2 Data Mining

Data mining is the discovery from large databases and to extract the hidden information. Data mining makes use of statistical and visualization techniques [18] to discover and present information in a form that is easily comprehensible. It is applied for decision support, forecasting, estimation. It identifies and label the data and form their relationships among data elements. It is designed to help organizations achieve business, operational, and scientific goals by revealing and analysing hidden patterns in their data — existing data from operational systems that may consume gigabytes or terabytes of data on a variety of operating system platforms.

2. Motivation

The recent success of XML as a standard to represent semi-structured data and the increasing amount of data available in XML pose new challenges to the data mining community [6]. However XML is still limited in data mining community.

Web Data mining is the large amounts of data from a hidden regularity found in the contents of the application to resolve data quality issues. Most of the data mining algorithms can handle data with a fixed structure [6][15], where data scheme is defined in advance. However data on the web, bioinformatics databases often lack such a regular structure. Due to availability of huge data on the Internet, or on the private Intranets of many companies, this data is structured in a multitude of ways. The structure generated in a traditional relational or object-oriented databases, is completely known. At another extreme we have data which is fully unstructured, such as images, sounds, and raw text [5]. But most of the data falls somewhere in between these two extremes, for a variety of reasons: the data may be structured, but the structure is not known to the user; many times user ignores the structure for browsing purposes. Sometimes structure may be implicit, like formatted text, as opposite to traditional databases which are rigid and regular .The data may be in non-traditional formats, such as the ASN.1 exchange format. The schema of the data is huge and changes often, so that we may prefer to ignore it. Due to such diverse structure of the data Data Mining as a young and promising field, still faces many challenges which pose new research issues for further study like efficient methods for mining multiple kinds of knowledge at multiple abstraction levels in varied databases like spatial databases, multimedia databases, web data etc. Accessing such data may be broadly classified in the following figure. The figure-1, shows that accessing huge semi structured will be carried out either by mining these databases using data mining techniques or by querying these databases using Keyword Based IR ranking methods. The Data mining techniques further may be classified as mining the structure of semi structure databases (i.e. XML) or mining the content of semi structured databases.

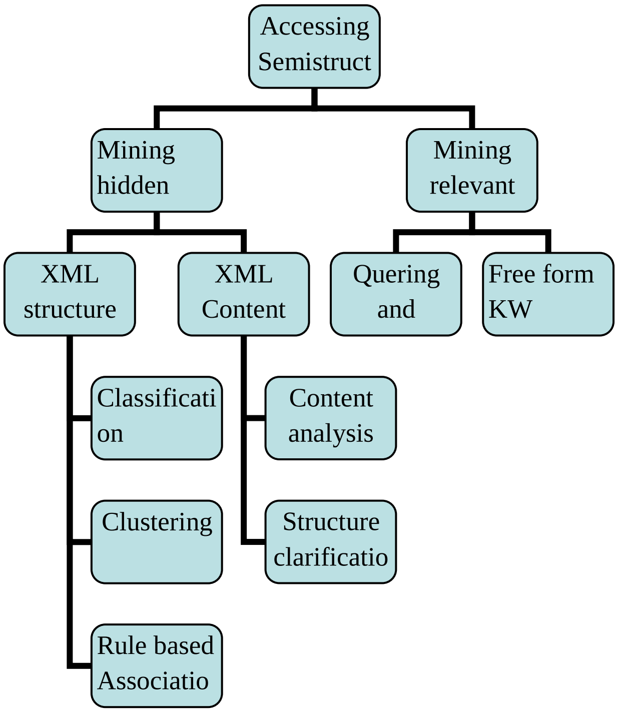

Fig 1: Classification of accessing semi structured data

The content mining further can be explored into two areas, Content analysis and structure classification. Content analysis mainly covers the usage pattern (e.g identifying the user behaviors through website usage). This analysis can further help website users for identifying the intended user.Another classification for accessing these database falls under the category of querying and processing the XML. IR-style approach utilizes the statistics of underlying XML data for retrieving most relevant document by extending efficient ranking function.

2.1 Data Mining applications

Semi structured data plays major role in the application of World Wide Web (WWW) Bioinformatics databases, Multimedia Data mining etc.

World Wide Web [16]: Web mining is the application of data mining techniques to the content, structure and usage of web resources; this can help discover global as well as local structure or patterns from the web pages. Unstructured data resides in web documents in the form of text, image, audio, video, metadata and hyper links. Detecting user's interests and browsing patterns on the web can help organize web pages and attract more businesses. This can be modelled as association patterns from a collection of hyperlinked Web pages that were accessed. Due to the sequential nature of the Web user's activity, Sequential Pattern Mining (SPM) [10] is particularly well adapted for the study of Web usage data. Traditional SPM techniques with very low support produce large number of SPs [6]. However they are unsuitable for extraction of knowledge about the minority users because of large diversified user's behaviours and difficult to locate.

Bioinformatics databases [7]: Domain specific sequence mining method such as bioinformatics, chemical, has a rich literature [8]. To interpret and mine the information from biological sequences and structures , data mining as applied in the field of bioinformatics for managing biological data. It contains many different kinds of data (e.g., genome sequences, pointers to journal articles, web pages, biochemical data, physical data, information about mutation experiments, etc.).It includes gene finding, protein function domain detection, disease diagnosis, disease treatment optimization etc. Often, in such cases data is missing, and there is conflicting data as the data is provided by the patients. So accurate prediction algorithm can help to provide better diagnosis and treatment. Genomic sequencing and mapping efforts have produced a number of biomedical databases. In addition, there are also wide variety of other on-line databases, including those containing information about diseases, cellular function, and drugs. One of the major data mining challenge is to find relationships between these data sources, which are largely unexplored. Many scalable techniques were used towards finding consecutive sequence whereas most bio sequence method allows base symbols as sequence. Mehmet Koyutürk, Ananth Grama [8] proposed an algorithm for detecting frequent graph in biological network. It extract frequently occurring patterns in metabolic pathways (KEGG database) modelled using directed hypergraphs, with nodes representing compounds (substrates and products), and hyperedges representing enzymes (reactions). Qingfeng Chen and Yi-Ping Phoebe[9] proposed a probabilistic graphic model . Author develops a Baysian Netwok based for structural learning from a discrete data set of protein kinases. It uses a Monte Carlo Markov chain algorithm to search the space of all directed acyclic graphics (DAGs) Jong Cheol Jeong, Xiaotong Lin, and Xue-Wen Chen[10] proposed a machine learning based methods for frequent based sequence prediction.

Multimedia data Mining [6] : It is a part of content mining where high-level information and knowledge explored from large online multimedia sources. A frequent pattern discovery can be done to answer such queries on request based on portion of the data received.

3. Literature Review

To deal with above mention application one of fundamental data mining task of sequential mining is applied .Sequence mining helps to discover frequent sequence pattern across time or position in a given data set. Therefore, mining for patterns and its models in the complex structure of the data, e.g., for frequent substructures, is an important aspect of mining semi-structured databases. Our Survey is characterized on the following key Challenges:

[1] Design an efficient algorithm to identify patterns from varied and huge semi structured data .
[2] Retrieving frequent graph pattern from a given set of graph for Mining

R Agarwal and R Shrikant has given a wide contribution in the field of sequential pattern mining. Since then many other approaches have followed. [4][5].Sequence mining is essentially an enumeration problem over the subsequence partial order where frequent pattern sequences are identified. Sequence mining in semi-structured data, for example XML data has received little attention [13][14] as the data mining community has focused on the development of the techniques to extract common structure from heterogeneous XML data [14]. The straight forward approach for association rule mining from XML data is to map the XML documents to relational data model and to store them in a relational database. This allows us to apply the standard tools that are in use to perform the rule mining from relational databases. This approach is time consuming and involves manual intervention because of mapping process [11][12]. Also this mapping to the structured form limits the flexible use of XML in domain specific application. XQuery[6] an XML query language addresses the need for the ability to intelligently query XML data sources. XQuery is flexible enough to query a broad spectrum of XML information sources, including both databases and documents. XQuery is used to perform the association rule mining directly from XML documents. Since XQuery is designed to be a general purpose XML query language, it is often difficult to implement Complicated algorithms. Authors [20]17][19] have implemented an Apriori algorithm by using Query. It described an approach based on Tree bases Association Rules for retrieving intentional information on both the structure and the contents of XML document. Many authors studied extensively in the web environment .O.R. Zaiane[3] proposed a method to find frequent behavioural patterns based on adapted navigational schemas. Bamshad Mobasher, Honghua Dai, Tao Luo Miki Nakagawa[21] described a framework for Web Personalization based on sequential and non-sequential pattern discovery from usage data. New approximate mining algorithm to mine approximate frequent itemsets over online data streams is proposed. Yongyan Wang, Kun Li and Hongan Wang[11] proposed a new approximate mining algorithm to mine approximate frequent itemsets over online data streams. Using an approximate frequent itemset tree (AFI-tree), called as AFI algorithm. Based on prefix tree , AFI tree based algorithms maintains the small no of nodes. All the infrequent child nodes of any frequent node are pruned and the maximal support of the pruned nodes is estimated to detect new frequent itemsets. Many researchers previously proposed the enumeration techniques of frequent trees .Various BFS enumeration techniques have been used by authors[11][20][22] .Yun Chi, presented[23] a CMTree miner that discovers only closed and maximal frequent subtrees in a database of labelled rooted trees, where the rooted trees can be either ordered or unordered. Wang and Liu presented the solution (maximal subtree) for difficulty in gaining the insight of the frequent tree due to huge number of subtrees produced . But author has extended their work to closed and Maximal patter mning . As the pruning takes place at early stage, the xpensive computation is eliminated. The enumeration *DAG* data structure and *Blanket* concept is introduced for pruning like left and right blanket pruning, occurrence matching and transaction matching. Their work is experimented on NASA dataset and is used for experimenting IP Multicast tree for efficiently sending message group of users.

4. Objective

The following are the primary objectives classified as follows:

[1] to design an efficient algorithm to identify patterns from varied and huge semistructured database

[2] Retrieving frequent graph pattern from a given set of graph for Mining

[3] Handling the uncertainity from the XML databases

5. Conclusion

XML standard widely used for interchanging information has various challenges and motivation for mining due to its varied, huge structure. Thus representing and managing this varied ,huge and uncertain information raises many research challenges. The two broad issues in this context are modelling the uncertain data and adapting data management and mining applications to work with the uncertain data. Through this literature survey it can be stated that various approaches proposed by various researchers in this domain possess different focus, features, advantages, and limitations. Some approaches focused on modelling, the others addressed merging uncertain information whereas some leads to efficient querying of these databases. To make the efficient use of semi structured data the main objective is to develop an efficient algorithm for optimized result by discovering rules or patterns from large collections of semi-structured data. Thus gives frequent sub trees for the general information content in the database for the summarization of data. This formulates queries, as a guideline for building indexes, as basis for structure based clustering. Thus this paper gives a brief survey of various data mining techniques and recent research issues for representing semi structured databases especially XML. Hence mining effective or desire patterns play important role in providing the optimized solution.

Acknowledgment

We are thankful to the reviewers for their valuable comments and suggestions.

References

[1] Agrawal, R., Imielinski, T, and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data.

[2] Chi, Yun, et al. "Mining closed and maximal frequent subtrees from databases of labeled

rooted trees." Knowledge and Data Engineering, IEEE Transactions on 17.2 (2005): 190-202.

[3] O. R. Zaiane" Web usage mining for a better web-based learning environment,in": Proceeding of Conference on Advanced Technology for Education, 2001,pp.450-455.

[4] Mohammed J. Zaki VLDB Elsvier 2009 "closed item set mining and non-redundant association mining ".

[5] Ke Wang and Huiqing Liu "Discovering Structural Association of Semistructured Data" IEEE Transaction on Knowledge and Data Engineering VOL. 12, NO. 3, MAY/JUNE 2000.

[6] Pardede, Eric, ed. Open and Novel Issues in XML Database Applications: Future Directions and Advanced Technologies: Future Directions and Advanced Technologies. IGI Global, 2009.

[7] Koyutürk, Ananth Grama and Wojciech Szpankowski An efficient algorithm for detecting frequent subgraphs in biological networks Mehmet Department of Computer Sciences, Purdue University, March 1, 2004.

[8] Mehmet Koyutürk, Ananth Grama and Wojciech Szpankowski "An efficient algorithm for detecting frequent subgraphs in biological networks" bioinformatics vol 20, 2004.

[9] Quingfen Chen, yi phig "Mining protein kinases regulation using graphical model" IEEE Transaction on Knowledge and Data Engineering Vol 10, March 2011 .

[10] Jong Cheol Jeong, Xiaotong Lin, and Xue-Wen Chen "On Position-Specific Scoring Matrix for Protein Function Prediction" IEEE/ACM Transactions on Computational Biology And Bioinformatics, VOL. 8, NO. 2, MARCH/APRIL 2011.

[11] Yongyan Wang, Kun Li and Hongan Wang "Maintaining Only Frequent Itemsets to Mine Approximate Frequent Itemsets over Online Data Streams"IEEE Transaction on Knowledge and Data Engineering 2009.

[12] Dr shobha, Meenakshi "Survey on Mining in Semi-Structured Data "International Journal of Computer Science and Network Security, VOL.7 No.8, August 2007.

[13] Ming syan chen "Data mining: an overview from Database perspective" IEEE Transaction on Knowledge and Data Engineering 1999.

[14] Sarawagi, Sunita, Shiby Thomas, and Rakesh Agrawal. Integrating association rule mining with relational database systems: Alternatives and implications. Vol. 27. No. 2. ACM, 1998.

[15] Abiteboul, Serge, et al. "Views for semistructured data." (1997).

[16] Srivastava, T., Prasanna Desikan, and Vipin Kumar. "Web mining–concepts, applications and research directions." Foundations and Advances in Data Mining. Springer Berlin Heidelberg, 2005. 275-307.

[17] Xinwei Wang and Chunjing Cao Mining Association Rules from Complex and Irregular XML Documents using XSLT and Xquery "IEEE computer society 2008.

[18] Chan, Carmen, and Bruce Lewis. "A basic primer on data mining." Information Systems Management 19.4 (2002): 56-60.

[19] An Improved Association Rule Mining Technique for Xml Data Using Xquery and Apriori Algorithm IEEE International Advance Computing Conference (IACC 2009).

[20] Mirjana Mazuran, Elisa Quintarelli, and Letizia Tanca "Data mining for XML query-answering support" IEEE Transaction on Knowledge and Data Engineering , Vol 24, No. 8, Aug 2012.

[21] Bamshad Mobasher, Honghua Dai, Tao Luo Miki Nakagawa,"Using sequential and Non-Sequential Patterns in Predictive Web Usage Mining Tasks",2002.

[22] Jiaheng Lu, Tok Wang Ling"Extended XML Tree Pattern Matching: Theories and Algorithms "IEEE Transaction on Knowledge and Data Engineering, March 2011.

[23] Yun chi, YirongYong "Mining closed and maximal frequent subtree from database of lebeled rooted trees" IEEE Transaction on Knowledge and Data Engineering Vol 17, March 2005.

Mrs. Leena deshpande has received M.E. Computer Engineering degree from Bharati Vidyapeeth ,Pune. Her research interest include Data Mining, XML ,association mining, text mining and distributed computing . She is a life member of ISTE chapter and CSI chapter of India.She is working as an assistant professor in VIIT, Pune.

Dr. R. S. Prasad graduated in Computer Science and Engineering from North Maharashtra University in 1996. He pursued M.B.A. in Marketing from North Maharashtra University in 1998 and then completed his M.E. in Computer Engineering from Pune University in 2004. He is currently working with Dnyanganga College of Engg. & Research (DCOER, Pune) since July 2012. He is a Life member of Professional Bodies like IEEE, ISTE, CSI and IAENG. His research interests are in the field of Artificial Intelligence, Soft Computing, Information Retrieval and Information systems. He has published more than 25 papers at National & International level journal and conferences. He has worked as a member of Program and Organizing committees in National and International Conferences in India and abroad. He is also working as reviewer of International journals.