Survey on Mining in Semi-Structured Data

Rajashree Shettar[1] , Dr. Shobha G[2]

[1]Asst.Professor, Dept of C.S.E, R.V.C.E, Bangalore
[2]Professor, Dept of C.S.E, R.V.C.E, Bangalore

Summary

Emerging technologies of semi-structured data have attracted wide attention of networks, e-commerce, information retrieval and databases. In these applications, the data are modeled not as static collections but as transient data streams, where the data source is an unbounded stream of individual data items. It is becoming increasingly popular to send heterogeneous and ill-structured data through networks. Since traditional database technologies are not directly applicable to such data streams, it is important to study efficient information extraction methods for semi-structured data. Hence there has been increasing demand for automatic methods for extracting useful information, particularly, for discovering rules or patterns from large collection of semi-structured data, namely, semi-structured data mining.

In this survey paper we begin by reviewing popular data mining techniques like association rules, clustering and prediction for semi-structured data. We provide a brief description of each technique as well as efficient algorithms for implementing the technique. Then we talk about the applications of semi-structured data. Finally, we conclude by listing research challenges that need to be addressed in the area of semi-structured data mining.

Keywords:
semi-structured data mining, association, clustering, prediction, graph based data structure.

1. Introduction

Data Mining is concerned with the discovery of patterns and relations in large collection of data. Data Mining is referred to as Knowledge Discovery in Databases. It deals with issues such as representation schemes for the concept or pattern to be discovered, design of appropriate functions and algorithms to find patterns. Most of the data mining algorithms can handle data with a fixed structure, where data scheme is defined in advance. However data on the web and bioinformatics databases often lack such a regular structure. We call such data as semi-structured.

By rapid progress of network and storage technologies, a huge amount of electronic data such as Web pages and XML data has been available on intra and internet. These electronic data are heterogeneous collection of ill-structured data that have no rigid structure, and are often called semi-structured data [1][2].

Structured data is one that can be neatly modeled, organized, and formatted into ways that are easy for us to manipulate and manage. The most frequent examples include databases, spreadsheets, fixed-format files, log files, etc.

Unstructured data incorporates the mass of information that does not fit easily into a set of database tables. The most recognizable form of unstructured data is text in documents, such as articles, slide presentation or message components of e-mails.

Semi-structured data refers to set of data in which there is some implicit structure that is generally followed, but not enough of a regular structure to qualify for the kinds of management and automation usually applied to structured data. Examples include the World Wide Web, bioinformatics databases and data ware housing. Unlike unstructured raw data such as image and sound, semi-structured data has some structure: objects share (parts of) their structure.

Despite the structural irregularity, semi-structured data typically do possess some structure [3]. Such structures implicit in semi-structured data can serve the following purposes: optimizing query evaluation, obtaining general information contents, facilitating the integration of data from several information sources, improving storage, assisting in building indexes and views and making it possible for structure-based document clustering [4][5]. Most data mining algorithms are not designed for semi-structured data and should at least be adapted in order to deal with such data [1].

We start this survey with a discussion on mining in semi-structured data and explore the various challenges and look into the algorithms for mining semi-structured data.

2. Semi-structured data

The use of semi-structured data can be felt in the areas involving raw data which does not have any fixed format. Semi-structured data is convenient for data integration. Web-sites containing semi-structured data are ultimately graphs.

More and more data sets do not fit in the rigid relational model because the individual data items do not have the same structure completely. Rather, the data items share only

Manuscript received August 5, 2007.
Manuscript revised August 20, 2007.

partly the same structure. Such databases are called semi-structured databases [5]. In a semi-structured database, there is no fixed database schema: conceptually the data is stored in a graph-like structure (like XML) which contains both information about the data as well as the data itself. Prime examples of semi-structured databases are XML databases and many of the bioinformatics databases. These bioinformatics databases differ from standard databases in the sense that

1. The bioinformatics database 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.);
2. The databases are often distributed and contain many links to other databases;
3. Often, data is missing, and there is conflicting data.

Therefore, mining for patterns and models in the structure of the data, e.g., for frequent substructures, is an important aspect of mining semi-structured databases.

In that respect, semi-structured data mining is close to multi-relational data mining [13] and Inductive Logic Programming [14]: in both cases patterns have structure.

In order to get some idea of semi-structured databases we examine a simplified movie database (the example is inspired by [6]). Suppose that a movie object has a name, person and a company. Every person has a (not necessarily unique) name, usually a nationality and a home town. A company has a name and a home town. All objects can have more data elements, such as pictures. Other databases contain extensive information on the persons and so on.

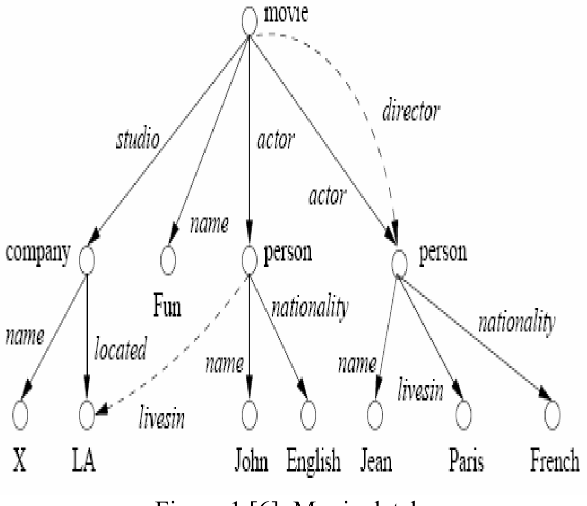

Figure 1 [6]: Movie database

The tree in figure 1 represents a sample movie. It is possible to consider graphs; the occurrence of cycles has to be considered. For example, in a movie database a director may serve as an actor in his/her own movie.

3. Semi-Structured Data Mining Techniques

In this section, we briefly describe key semi-structured data mining algorithms that have been developed to handle data which does not have rigid structure. A number of these algorithms are also applicable in the Web and Bioinformatics context.

3.1 Association Rules

Algorithms for mining association rules from relational data have been well developed. Several query languages have been proposed, to assist association rule mining such as [15].

Mining semi-structured data, for example XML data has received little attention, as the data mining community has focused on the development of the techniques for extracting common structure from heterogeneous XML data [16].

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. Hence not suitable for XML data streams [19].

XQuery [27], 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. This led to the use of XQuery 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. The authors in [16], the have implemented Apriori algorithm [17] by using XQuery.

The approach followed in [19], for XML rule mining is to use programs written in java to work directly with XML documents which offers more flexibility and performs well compared to other techniques. The authors plan to implement the FP-based algorithm [20] by using XQuery and compare its results with the java based algorithm.

In [18], the authors propose a method of discovery of association rules in semi-structured data, namely, in a set of conceptual graphs. The method is based on conceptual clustering of data and construction of a conceptual hierarchy. Conceptual clustering of graphs is considered as a kind of index of the collection, and to take advantage of this structure when searching for association. Given a set of conceptual graphs C= $\{G_i\}$, we define an association rule as an expression of the $g_i \Rightarrow g_j$ (α,β), where g_i is a generalization of g_j $(g_i < g_j)$; c is the confidence of the rule and s it supports. An association rule of this kind indicates that the conceptual graphs of the collection that contain the graph g , c% of the times also contains the more specialized graphs g_j ; also indicates that s% of the graphs of the collection contains the graph g_j .

3.2 Clustering

Clustering in data mining is, a useful technique for discovering interesting data distributions and patterns in the underlying data. Clustering is also helpful for categorizing www documents, grouping genes and proteins that have similar functions or the detection of seismic faults by grouping the entries in an earthquake catalog. All these examples have in common that the better the quality of the clustering algorithm, the higher the benefits realized.

In [20], the authors present an approach based on the use of two languages of description of classes for the automatic clustering of semi-structured data. The first language of classes guides the construction of a lattice of classes covering whole set of data. The second language of classes is the basis for the refinement of a part of the lattice that user wants to focus on.

Document Type definition (DTD) is an important concept of XML, the full advantage of this is not taken in current applications. In [21], the authors describe a cluster based algorithm which when given to a semi-structured data extracts a schema for that data. a new method for clustering DTDs is presented, and it can be used in XML document clustering. The two-level method clusters the elements in DTDs and clusters DTDs separately. Element clustering forms the first level and provides dement clusters, which are the generalization of relevant elements. DTD clustering utilizes the generalized information and forms the second level in the whole clustering process. The two-level method has the following advantages: 1) It takes into consideration both the content and the structure within DTDs; 2) The generalized information about elements is more useful than the separated words in the vector model; 3) The two-level method facilitates the searching of outliers. Specifically the authors show that this method is able to categorize the relevant DTDs effectively.

Knowledge discovery in text systems is done with the documents which are represented by keywords and knowledge discovery is performed by analyzing the co-occurrence frequencies of the various keywords representing the document. The clustering of documents is done by extracted knowledge, which can reduce the search space[22].

In [23], the XML document is viewed as an ordered labeled tree. Irrelevant tags are removed for clustering operation. Data mining steps are applied to each cluster. For each cluster the goal is to transform each XML documents into a sequence. Classifying ordered labeled trees can be done by characterizing each pre-defined cluster in terms of frequent structural patterns.

3.3 Prediction

Prediction using micro array technologies is an important application in bioinformatics. Based on patients' genotypic microarray data, predictions are made to estimate patients' survival time and their risk of tumor metastasis or recurrence. So, accurate prediction can potentially help to provide better treatment for patients. Patient outcome prediction using microarray technologies is an important application in bioinformatics. The authors in [24] present a new computational method for patient outcome prediction. Two types of two types of extreme patient samples: short-term *survivors* who got an unfavorable outcome within a short period and long-term *survivors* who were maintaining a favorable outcome after a long follow-up time. These extreme training samples yield a clear platform to identify relevant genes whose expression is closely related to the outcome. The selected extreme samples and the relevant genes are then integrated by a support vector machine to build a prediction model, by which each validation sample is assigned a risk score that falls into one of the special pre-defined risk groups. In [25], the authors describe a system called DISCOTEX, to discover prediction rules from natural language. Specific data is to be located in natural language text using Information Extraction system (IE). The data to be extracted is typically given by a template which specifies a list of slots to be filled with sub-strings taken from the document. After constructing an IE system that extracts the desired set of slots for a given application, a database is constructed. Rules are induced for predicting each piece of information. Rules can predict absence of a filler in a slot.

4. Applications

All applications of semi-structured data can be broadly classified as follows[12]:

[a] Web Content Mining

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 modeled as association patterns from a collection of hyperlinked Web pages that were accessed.

[b] E-shopping

Electronic shopping patterns can be discovered by modeling partially ordered services as semi-structured data and discovering typical partial orderings. The manager can use such patterns to organize service chains more effectively. Customers' interests and access patterns can also be discovered.

[c] Multimedia Data mining

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

[d] Classification of chemicals/proteins/living things

Chemical information systems organize chemical compound files into semi-structured trees in which further information about each fragment occurs at each successively lower level. Classification of chemical structures is based on typical fragments in such trees. Protein structure classification also depends on identifying structural similarity.

[e] Event and causality analysis

Applications such as job scheduling, dependency discovery, workflow and process management, resource management and discovery, and medical diagnosis, an object represents either an atomic or complex event. An association pattern is a regularity about how events are composed and how sub-events are dependent on each other.

[f] Intrusion detection

In Intrusion detection, data mining can be used separate the normal activity from alarm data so that analysts can focus on real attacks, to identify false alarm generators and to identify long ongoing patterns (different IP address, same activity). To accomplish this data summarization, visualization, clustering, association rule discover, anomaly detection and classification techniques can be applied[28].

[g] Bioinformatics data mining

Bioinformatics is the science of managing, mining and interpreting information from biological sequences and structures. It includes gene finding, protein function domain detection, disease diagnosis, disease treatment optimization. Based on the data provided by the patients, accurate prediction algorithm can help to provide better treatment. Clustering method can be applied to cluster genetic data.

Genomic sequencing and mapping efforts have produced a number of databases. In addition, there are also a wide variety of other on-line databases, including those containing information about diseases, cellular function, and drugs. Finding relationships between these data sources, which are largely unexplored, is another fundamental data mining challenge. Recently, scalable techniques have been developed for comparing whole genomes[30].

5. Research Challenges

There are a number of issues that are identified in this survey so as to make efficient use of semi-structured data. The primary objective is of course to develop algorithms that perform well and give optimized results. The research areas can be classified as follows:

[a] Frequent Trees and Graphs

There are a number of issues identified in this survey that can be implemented in the form of efficient algorithms to model semi-structured data. One of the primary objectives of the research challenges is to discover rules or patterns from large collections of semi-structured data. From figure 1, one of the interesting patterns can be search for movies with many local actors or it could be the search for the most often occurring actor. This might seem like answering SQL-queries, but it rather requires the discovery of the interesting ones for which efficient and effective algorithms are to be designed with a focus on tree- and graph-structured data.

Data stored in the form of tree structured XML documents, we can always discover frequent item sets, and it would be interesting to find frequent (sub-) trees. Frequent trees give a feeling for the general information content in the database: it is a way of summarizing the data. This information can then be used to formulate queries, as a guideline for building indexes, as basis for structure based clustering. These can be seen as generalization of the popular frequent item sets, which form the basis of association rules [7].

[b] Semi-structured Data Repository

Semi-structured data is claimed to be self-describing, making it important to define a schema for the data. The move from DTD (document type definition), a simple schema definition language, to XML Schema, a more expressive schema definition language, highlights the importance of a schema definition for semi-structured data applications [29]. The data models that have been proposed specifically for semi-structured data, (e.g., DOM (Document Object Model), Dataguides, YAT, and UnQL), capture limited semantics in the data, and do not model the semantics expressed in the schema. The research challenge is to define a data model that captures richer semantics. Using the semantically rich data model, we can investigate and experiment with more efficient storage mechanisms, identify valid views of the base data, define efficient view maintenance algorithms, and propose methods for efficiently querying a semi-structured data repository.

[c] Bioinformatics Data

Biological data is complex and semi-structured in nature. The flood of genomic data, their high variety and heterogeneity pose new challenges in computing. It has inherently deeply nested hierarchical structures (e.g., ontologies), or is best modeled as graph structures at the conceptual level (e.g., metabolic pathways, or signaling pathways). It is heterogeneous, in the sense that it involves a wide array of data types, including text, image, and sequence data, as well as streaming data (e.g., medical sensors data), temporal data, and incomplete and missing data. Novel high throughput techniques such as DNA microarrays are generating overwhelming amount of data. In addition to being large, diverse and distributed, biological data has three important characteristics that pose additional challenges for data management, such as 1) complexity, 2) Heterogeneity, and 3) Evolution of both data and the schema [30].

[d] Storage, Retrieval and Query Processing

Providing efficient storage, retrieval and query processing is one of the research challenges. Current approaches range from the use of files and native storage mechanisms to the use of traditional database management systems. Query processing and validity checking, for instance, are particularly harder to perform using current DBMSs. Constraint-based technology[8,9] can be used to query and reason about semi-structured data Constraint based mining can be used to cover frequent patterns in semi-structured data. The definitions of classes of useful constraints on the structures(trees and graphs) to be mined from semi-structured data. A categorization of those constraints similar to what has been done for mining frequent item sets Development of algorithms that can use such constraints to increase the efficiency of the search process. Condensed based representation [10,11] is a complementary method for reducing the large amount of frequent patterns. The major research issue is to develop condensed representations for patterns, i.e. trees and graphs, that are mined from semi-structured data and to develop algorithm for their discovery.

[e] XML Data mining

There is an increasing research efforts going on in XML data mining which is based on XML documents. These documents are rarely static. A novel research problem called *XML structural delta mining* [26] has been taken up. The objective of XML structural delta mining is to discover knowledge by analyzing structural evolution pattern (also called *structural delta*) of history of XML documents. Unlike existing approaches, XML structural delta mining focuses on the dynamic and temporal features of XML data. Furthermore, the data source for this novel mining technique is a sequence of historical versions of an XML document rather than a set of snapshot XML documents. Such mining technique can be useful in many applications such as change detection for very large XML documents, efficient XML indexing, XML search engine, etc. XML structural delta mining research issues are *identifying various interesting structures*, *discovering association rules from structural deltas*, and *structural change pattern-based classification*.

6. Summary

Semi-structured data arise in many application areas. The World Wide Web, XML further increases the availability of semi-structured data. In the semi-structured world, while many proposals exists on modeling, searching information exchanging and structure extracting, data mining is largely unexplored. Mining task has to deal with both data and schema.

In this survey paper we highlighted a wide range of application. As more and more data do not impose a rigid schema, as those on WWW or digital libraries, we believe that the data mining algorithms dealing with semi-structured information are of emerging importance. We talked about the requirement for efficient and effective mining algorithms as most data mining algorithms are not designed for semi-structured data and should at least be adapted in order to deal with such data.

We talked about the research issues with a focus on finding frequent sub-trees and sub-graphs considering tree and graph structured data. More research, however, is still needed in the areas of storage of semi-structured data i.e. XML documents, and integration and querying of XML documents originating from different sources.

References

[1] S. Abiteboul, P. Buneman , D. Suciu, *Data on the Web*, Morgan Kaufmann, 2000.

[2] W3C, Extensive Marrkup Language (XML) 1.0 (Second Edition), *W3C Recommentation,* 06 October 2000.

[3] P.Buneman and B.Pierce. Union types for semi-structured data. In technical report MS-CIS-99-09, Dept. of CIS, university of Pennsylvania.

[4] S. Abiteboul, P. Buneman , D. Suciu. Data on the Web: from relations to semi-structured data and XML. Morgan Kaufmann Publishers, 2000.

[5] K.Wang and H.Liu. Discovering Structural Association of Semistructured data. In IEEE Transactions on knowledge and data engineering, 12(3), pages 353-371, may/June, 2000.

[6] G.Grahne, X. Wang, and L.V.S. Lakshmanan. Efficient mining of constraint correlated sets. In proceedings of the 16th International Conference on Data Engineering (ICDEE'00), pages 512-521, 2000.

[7] L.Lakshmanan, R. Ng, J. Han, and A. Pang. Optimization of constrained frequent set queries with 2-variable constraints. In Proceedings of the 1999 ACM-SIGMOD International Conference on Management of Data (SIGMOD'99), pages 157-168, 1999.

[8] J. Pei and J. Han. Constraint frequent pattern mining: A pattern-growth view, SIGKDD *Explorations,* 4:31-39, 2002.

[9] R. Potharst and A. Feelders. Classification trees for problems with monotonicity constraints. SIGKDD *Explorations,* 4:1-10, 2002.

[10] R.J. Bayardo Jr. Efficiently mining long patterns from databases. In L.M.Haas and A. Tiwary, editors, Proceedings of the ACM SIGMOD *Conference on Management of DATA,* pages 85-93. ACM Press,1998.

[11] B.Jeudy and J-F. Boulicat. Using condensed representations for interactive association rule mining. In T.Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of the sixth European Conference on Principles of Data Mining and Knowledge Discovery(PKDD 2002), volume 2431 of Lecture notes in Artificial Intelligence,* Pages 225-236.

[12] Ke Wang, Huiqing Liu. *Mining is-part-of association Patterns from Semistructured Data*. World Scientific 2001.

[13] A.J.Knobbe, A. Siebes and B. Marseille. Involving aggregrate functions in multi-relational search. In T.Elomaa, H.Manilla, and H.Toivonen, editors, *Proceedings of the Sixth European Conference on Principles of Data Mining and Knowledge Disciovery (PKDD 2002),* volume 2431 of lecture notes in Artificial Intelligence, pages 287-289. Springer-Verlag, 2002.

[14] James Stuart Aitken *"Learning Information Extraction rules : An Inductive Logic Programming Approach" .*

[15] T. Imielinski and a. Virmani, *A query language for database mining, 1999.*

[16] Jacky W.W. Wan , Gillian Dobbie, Mining association rules from XML data using XQUERY.

[17] D. Braga, A. Campi, M. Klemettinen and P.L Lanzi. Mining association rules from XML data. In Proceedings of the 4th International conference on data warehousing and knowledge discovery. France 2002.

[18] M. Montesy-y-Gomez, A. Gelbukh, A Lopez-lopez. Discovering association Rules in Semi-structured data sets.

[19] Qin Ding and Gnanasekaran Sundarraj. Association rule mining from XML data.

[20] Nathalie Pernelle, Marie-Christine Rousset, Veronique Ventos. *Automatic Construction and refinement of a class hierarchy over semi-structured data.*

[21] Svetlozar Nestorov, Serge Abiteboul, Rajeev Motwani. *Extracting schema from semi-structured data.*

[22] Gurusamy,S.;Manjula,D.;Geetha,T.V. ***Text mining in apos; Request for Comments Document Seriesapos;*** Language Engineering Conference, 2002. Proceedings Volume , Issue , 13-15 Dec. 2002 Page(s): 147 – 155.

[23] Calin Garboni, Florent Masseglia and Brigitte Trousse. *Sequential Pattern Mining for Structure-Based XML Document Classification*. AXIS research team 2004.

[24] Huiqing Liu *, Jinyan Li and Limsoon Wong. *Use of extreme patient samples for outcome prediction from gene expression data.* Institute for Infocomm Research Heng Mui Keng Terrace, Singapore 119613. [25] Un Yong Nahm, Raymond J. Mooney. Using Information Extraction to aid the discovery of Prediction rules from text.

[26] Qiankun Zhao[a], Ling Chen[a], Sourav S. Bhowmick[a,] and Sanjay Madria[b] October 2005. *XML structural delta mining: Issues and challenges.*

[27] World Wide Web Consortium. XQuery 1.0: AnXMLQueryLanguage(W3CWorkingDraft). http://www.w3.org/TR/2002/WDxquery20020816, Aug. 2002.

[28] Eric Bloedorn, Alan D. Christiansen, William Hill, Clement Skorupka, Lisa M. Talbot, Jonathan Tivel. *Data Mining for Network Intrusion Detection: How to get Started.*

[29] Alin Deutsch, Mary Fernandez, Dan Suciu. *Storing semistructured data with STORED.*

[30] *Syed Ahsan, Abad Shah, University of Engineering and Technology, Lahore, Pakistan.* Biological Databanks: Distribution, Heterogeneity, Diversity and Provenance.

Rajashree Shettar, Asst.Professor in the Department of Computer-Science,R.V. College of Engineering, Bangalore. She is pursuing her doctoral degree in the area of Semi-Structured Data. She obtained her M.S degree in Software Systems from BITS, PILANI and B.E degree in Computer-Science from Karnatak University. Her research interests are Data Mining, Semi-Structured Data Mining. She has guided several undergraduate and post-graduate projects. She is teaching courses on DBMS, Data Mining, Computer Architecture, Operating Systems, and Distributed Systems.

Dr. Shobha G., Professor in the Department of Computer Science & Engg. She has been awarded Ph.D for her thesis titled "Knowledge Discovery in Transactional Database Systems" from Mangalore University, Mangalore. She obtained her M.S. degree in Software Systems. from BITS, Pilani and BE in Computer Science from Gulbarga University. Her research interests are Data Mining, DBMS, Operating Systems & Networking. She has guided more than 30 undergraduate and 09 post graduate projects. Currently she is teaching courses on DBMS, Data Mining, Networks & Operating System. She has presented and published papers at national and International journals / conference.