Eleventh Pacific Rim in International Conference on Artificial Intelligence (PRICA1 2010)

A Statistical Interestingness Measures For XML Based Association Rules

En. Izwan Nizal Mohd Shaharanee

A Statistical Interestingness Measures for XML Based Association Rules

Izwan Nizal Mohd Shaharanee, Fedja Hadzic, and Tharam S. Dillon

Digital Ecosystem and Business Intelligence Institute, Curtin University of Technology, Perth 6102, Australia
izwan.mohdshaharanee@postgrad.curtin.edu.au,
{f.hadzic,tharam.dillon}@cbs.curtin.edu.au

Abstract. Recently mining frequent substructures from XML data has gained a considerable amount of interest. Different methods have been proposed and examined for mining frequent patterns from XML documents efficiently and effectively. While many frequent XML patterns generated are useful and interesting, it is common that a large portion of them is not considered as interesting or significant for the application at hand. In this paper, we present a systematic approach to ascertain whether the discovered XML patterns are significant and not just coincidental associations, and provide a precise statistical approach to support this framework. The proposed strategy combines data mining and statistical measurement techniques to discard the non significant patterns. In this paper we considered the "Prions" database that describes the protein instances stored for Human Prions Protein. The proposed unified framework is applied on this dataset to demonstrate its effectiveness in assessing interestingness of discovered XML patterns by statistical means. When the dataset is used for classification/prediction purposes, the proposed approach will discard non significant XML patterns, without the cost of a reduction in the accuracy of the pattern set as a whole.

Keywords: data mining, interesting rules, statistical analysis, semi-structured data.

1 Introduction

Data mining or knowledge discovery from data (KDD) is known for its capabilities in extracting knowledge that is comprehensible, valid on tests and new data with some degree of certainty, potentially useful, actionable, and novel [1]. With the fast growth in the amount of electronic data such as Web pages and XML data, this offers a new dimension in pattern recognition and rules discovery. These electronic data are heterogeneous collection of ill-structured data that have no rigid structures, and often referred to as semi-structured data [2]. A well known data mining technique, namely association rule mining is widely used for discovering interesting associations and correlations between data elements in a diverse range of applications. While there are great achievements in discovering the association rules within the well-structured

B.-T. Zhang and M.A. Orgun (Eds.): PRICAI 2010, LNAI 6230, pp. 194–205, 2010.

(relational) data, still a number of works remain in preliminary stages for semi-structured data [3]. Since the introduction of the association rule mining problem by [4], substantial work has gone into various trends, including the development of efficient algorithms in finding the association [5-7] and measuring the interestingness of the association rules in structured data [8-14]. As the increase in data captured in semi structured format such XML begins to permeate many applications, association rule mining from the semi-structured data has become a new and interesting research area [15]. The general problems of association rule mining include the extraction of all the frequent itemsets from which association rules are formed. A rule is said to be interesting if they meet certain minimum support and confidence criteria [3]. The same holds for mining the frequent substructures in semi-structured data which comprise candidate substructure enumeration and frequency counting.

Works such as [2, 16-18] focus on developing algorithms to enable efficient and effective association rule mining from semi-structured data. While these frequent substructure mining techniques may discover an interesting association from a given dataset, the problem that remains is that they may only reflect aspects of the database being observed. As such, the patterns may not reflect the "real" significant associations between the underlying structures. This problem arises because some association rules are discovered due to pure coincidence resulting from certain randomness in the particular dataset being analyzed. Since the nature of data mining techniques is data driven, the patterns generated by these techniques must be validated by a statistical methodology for them to be useful in practice [19]. Statistics has previously addressed the issues of how to separate out the random effects to determine if the measured association (or difference in other areas) is significant [20]. Thus additional measures based on statistical independence and correlation analysis are needed to ensure that the results have a sound statistical basis and are not purely random coincidence.

Therefore, the motivation behind our proposed method is to investigate how data mining and statistical measurement techniques can be combined to arrive at more reliable and interesting set of rules. The focus of the work presented in this paper is to evaluate the frequent substructures extracted from XML documents and verify their significance using statistical analysis. In this paper we apply the IMB3 algorithm [21] to the Prions database in order to extract the frequently occurring substructures, while statistical analysis, namely Chi-Squared and Log-Linear have been utilized to ascertain the discovered substructures. In the next section, we explain the problem of discovering and ascertaining association rules from semi structured data. In Section III, we describe some related works in the area of frequent substructure mining and finding of significant patterns. We show experimental findings of significant substructures in Prions dataset in Section IV. Section V concludes the paper and explains our ongoing work in this field of study.

2 Problem Definition

This section starts by describing some necessary aspects of association rule mining in the context of XML document mining which will lay the ground work to define the problem of ascertaining patterns/association rules from semi-structured data. XML

document has a hierarchical document structure, where an XML element may contain further embedded elements, and these can be attached with a number of attributes. Elements that form sibling relationships may have ordering imposed on them. Each element of an XML document has *name* and *value*. Given such parallelisms, an XML document can therefore be modeled as a rooted labeled ordered tree, where a node in the tree corresponds to an XML element [15, 17]. If only structure is to be considered, then a node in the tree will only correspond to an element name. However, in the case of the current study we are interested in attribute names and the attribute values from a particular domain, and hence a node will correspond to an element name and value.

A tree can be denoted as $T(v0,V,L,E)$, where:

(1) $v0 \in V$ is the *root* vertex;
(2) V is the set of *vertices* or *nodes*;
(3) L is the set of *labels* of vertices, for any vertex $v \in V$, $L(v)$ denotes the label of v; and
(4) $E = \{(x,y) | x,y \in V\}$ is the set of *edges* in the tree.

The main problem in association mining from semi-structured documents such as XML, is that of frequent pattern discovery, where a pattern corresponds to a subtree in this case, and a transaction to a fragment of the database tree whereby an independent instance is described. This problem is more complex than in traditional frequent pattern mining from relational data because structural relationships need to be taken into account. It is known as the ***frequent subtree mining*** problem, and can be generally stated as: given a tree database T and minimum support threshold (σ), find all subtrees that occur at least σ times in T.

Furthermore, depending on the domain of interest and the task that is to be accomplished in a particular application, different types of subtrees can be mined using different support definitions. For an overview of existing subtree types and support definitions and their usage implications for general knowledge analysis tasks please refer to [22]. Many frequent subtree mining algorithms have been developed to date, and for an extensive overview of the current state-of-the-art in the field, including comparisons of different approaches highlighting their advantages/ disadvantages, we refer the interested reader to [3, 15].

Due to the nature of the domain considered and the data used in this paper we focus on ordered induced subtrees and the transaction based support definition is used. These can be formally defined as follows:

Definition 1. Given a tree $S = (v0_S, V_S, L_S, E_S)$ and tree $T = (v0_T, V_T, L_T, E_T)$, S is an ***ordered induced subtree*** of T, iff (1) $V_S \subseteq V_T$;(2) $L_S \subseteq L_T$, and $L_S(v)=L_T(v)$; (3) $E_S \subseteq E_T$; and (4) the left to right ordering of sibling nodes in the original tree is preserved.

When using the ***transaction-based support*** *(TS)* definition, the transactional support *(*σ*)* of a subtree t, denoted as $\sigma_{tr}(t)$ in a tree database T_{db} is equal to the number of transactions in T_{db} that contain at least one occurrence of subtree t.

Definition 2. Let the notation $t \prec k$, denote the support of subtree t by transaction k, then for *TS*, $t \prec k = 1$ whenever k contains at least one occurrence of t, and 0 otherwise. Suppose that there are N transactions k_1 to k_N of tree in *Tdb*, the $\sigma_{tr}(t)$ in T_{db} is defined as:

$$\sum_{i=1}^{N} t \prec k_i \tag{1}$$

Hence, in our current work we focus on ascertaining the interestingness of discovered ordered induced subtree patterns that have been extracted from a tree-structured database (XML), and that satisfy the minimum transaction-based support threshold.

Let us denote the set of these frequent subtree patterns as *SF*. Please note that the patterns from *SF* have not been assigned a particular class label to be used for a prediction/classification task, and as such simply reflect the frequently occurring associations, that may not necessarily have a sound statistical basis. Hence, in the first problem setting our aim is to reduce the *SF*, by filtering out the patterns that are not statistically significant with respect to the statistical measures used.

In the second problem setting, one of the attributes from the data is considered as a class to be predicted for classification task purposes. Hence, we only consider those patterns from *SF*, that contain this class attribute, as they will represent the set of values that frequently occur together when a particular class value is present. Hence, as such these patterns can be seen to have predictive power and can be evaluated for their accuracy on correctly predicting the class value from the trained data and unseen data. In addition to predictive accuracy, simple rules are preferred as they are easier to comprehend and are expected to perform better on unseen data since they are more general. Hence, when in the process of optimizing a rule set, a trade-off needs to be made between several factors and the common ones are:

- *Misclassification rate* (*MR*) – number of incorrectly classified instances
- *Coverage rate* (*CR*) - number of captured instances
- *Generalization power* (*GP*) – capability of correctly classifying future instances

When optimizing the rule set, the *MR* should be minimized while the *CR* should be maximized. *GP* is achieved by simplifying the rules in terms of overall rule set size and the number of attribute constraints in the rule. The trade-off occurs especially when the data set is characterized by continuous attributes where a valid attribute range constraint needs to be determined for a particular rule. Increasing the range constraint usually leads to the increase in *CR* of that rule but at the cost of an increase in *MR* of that rule. Similarly, if the rules are too general, they may lack the specificity to distinguish some domain characteristics and hence the *MR* would increase. Generally speaking, an optimized rule set should be either more accurate than the original rule set and/or the balance between the trade-off factors should be much greater. For example, if there are many rules with small *CR* but very low *MR*, a rule set with a significantly smaller number of rules may be preferred even at the cost of an increase in *MR*.

Since the number of patterns/association rules generated through frequent subtree mining can be quite large, their usefulness for classification/prediction task may be limited unless they are significantly reduced in size and number. While their *MR* may be small, their *GP* is likely to be poor as all frequent patterns are considered, that can be insignificant, redundant and unnecessarily complex. Hence in the second problem considered in this paper, we aim to apply a variety of statistical/heuristic methods to reduce the pattern/rule set size and simplify individual rules.

Let us denote the subtree patterns from the frequent subtree set *SF* that have a class label (value), as *SFC*. The problem considered in the second setting can be stated as. Given *SFC* with accuracy *ac*, reduce *SFC* into *SFC'* such that *SFC'* has accuracy ≥ $(ac - \varepsilon)$, such that ε is an arbitrary user defined small value (ε is used to reflect the noise that is often present in real world data).

3 Related Works

Our work in this paper focused on ascertaining the XML rules discovered from an XML-enabled association rule framework. [17] have initiated this framework which resulted in a more flexible and powerful representations of both simple and complex structured association relationships inherent in XML documents. There has been an active development of frequent subtree mining algorithms [16, 18, 21, 23-25]. For a more detailed description of the existing approaches and latest development on these algorithms please refer to [3, 15]. Currently there has been limited works in rule evaluation phase of semi-structured rules. Many of the well developed rule interestingness measures are in structured data and they have had great success in evaluating rule interestingness as discussed in [12]. Initial work on evaluating the discovered patterns based on statistical significant are [13, 26-28] but these are limited to structural data. The existence of vast well developed measuring techniques to evaluate interestingness of rules from relational data, offers great opportunities in adapting these techniques for verifying significant substructures from semi-structure data. The applicability of these interestingness measures needs to be explored in context of frequent substructure mining, where necessary adjustments and extensions need to take place to ascertain the validity of the methods in presence of more complex structural aspects in the data, which often need to be preserved in the rules.

One line of work in focusing on more interesting substructure patterns is in reducing the patterns and the application of plausible constraints techniques. The problem of mining mutually dependent ordered subtrees has been addressed in [29]. The proposed algorithm utilizes the hyperclique method [30] in the tree mining context so that all the components of a subtree are highly correlated together. These hyperclique subtree patterns are discovered using a *h-confidence* measure which is the minimum probability of an item from a pattern in one transaction implying the presence of all other items in the same transaction. Hence, the extracted hyperclique subtree patterns will satisfy the minimum *h-confidence* threshold. The work done in [31] uses the method proposed for database compression in regards to item set mining in [32] to demonstrate how the same minimum description length principle can yield good results for sequential and tree-structured data. Another notable work presented in [33] extends the idea of the item constraint [34] to that of node-inclusion constraint in subtrees. In addition to that, [35] proposed the application of monotone constrain namely anti-monotone, monotone convertible and succinct in frequent subtree mining. Such an opportunistic pruning strategy is used to mine frequent subtrees under the defined constraints. An approach for mining of frequent subtrees where the distance between the nodes is used as additional grouping criterion has been presented in [36]. [37] proposed and demonstrated an efficient ways to discover interesting association

rules from dynamic XML documents. The work done in [37] mainly motivated by the facts that the XML document's content and /or structure are always fluctuates.

Besides the aforementioned constraint-based techniques, to our knowledge we found limited works on verifying the significance of discovered frequent substructures The frequent occurring substructure discovered from frequent substructure mining algorithm commonly offers a complete pattern set and is too numerous to be utilized efficiently and effectively for the application at hand [9, 38]. [39] proposed and developed an application of statistical hypothesis testing to re-rank the significant frequent subtrees. This approach ranks the significant patterns according to *P-values* obtained from the *Fisher's Exact* test of significance. The significant patterns were then used for Glycan classifications problems. Recently [38], proposed a mining framework called LEAP (Descending Leap Mine) in checking and mining a significant frequent subgraph which will help in discarding redundant frequent subgraphs. For a predefined class label in XML documents, an efficient XRules classifier have been develop by [40]. This approach offers promising results in terms of the structural classifier for semi-structured data.

In this work we employed the IMB3 miner algorithm for mining ordered embedded subtrees. While these algorithms, offer some constraints in discovering strong patterns/rules, many misleading, uninteresting and insignificant rules in that domains may still be produced [1]. The problem arises because some association rules are discovered due to pure coincidence resulting from certain randomness in the particular dataset being analyzed. Statistics has previously addressed the issues of how to separate out the random effects to determine if the measured association (or difference in other areas) is significant [20]. Thus additional measures based on statistical independence and correlation analysis are needed to ensure that the results have a sound statistical basis and are not purely random coincidence.

A common multivariate statistical analysis is the association analysis problem [20]. For associations between categorical variables there are several inferential methods involved. Chi-Squared analysis is often used to measure the difference between observed and expected frequencies. The significance used of the Chi-Squared statistics is for hypothesis testing in tests of independence. In addition to that the Log-Linear analysis offers a unique feature in capturing interrelationship among data items [41].

4 Experimental Results

The evaluation of the unification framework is performed using the Prions database which is a type of infectious agent. Prions are abnormally structured forms of host protein, which are able to convert normal molecules of protein into abnormally structured form. Prions dataset describes Protein Ontology database for Human Prions proteins in XML format [42]. It consists of 17348 protein sequences. The XML tags and values are first mapped to integer indexes similar to the format used in [21] and [25]. Representing label as integer instead of a string label has considerable performance and space advantages [21]. In this section, we first show the generated patterns obtained from frequent subtree mining approach, namely IMB3 algorithm in Section 4.1. Then we apply the two prominent statistical measurement techniques

namely Chi-Squared analysis and Log-Linear analysis in measuring the significance of the discovered frequent patterns in Section 4.2. In Section 4.3, we consider the Prions protein database in a classification/prediction problem setting. We have labeled the protein instances as either referring to Human's or Animal's protein. We then verified the extracted patterns using the statistical analysis.

4.1 Extracted Frequent Patterns

The discovery of structural patterns by matching data representation structures is essential for analysis and understanding of data. If a structural pattern occurs frequently, it is ought to be important in some way. On the other hand, infrequent patterns may also provide meaningful information [42]. Thus to extract meaningful information from XML data we need to mine structural patterns. In discovering the frequent patterns from Prions dataset we apply the IMB3 algorithm. There are a total of 27 occurring patterns discovered by IMB3 algorithm. The minimum support value used was 10 % and we managed to discover subtree patterns with the largest ones consisting of 5 nodes. Table 1 shows several examples of patterns discovered.

Table 1. Examples of Several Patterns Discovered Based on Frequent Tree Mining Technique

Patterns #	Patterns	# of Occurrences
1	ATOMChain(A) Element(C)	3957
2	ATOMChain(A) ATOMResidual(TYR) Occupancy (1)	1743
3	ATOMChain(A) Occupancy(1) Temperature(0) Element(C)	3805

Pattern number 1 shows an association between *ATOMChain(A)* with *Element(C)* and this pattern was discovered 3957 times. Here the *ATOMChain* with value *A* associates to *Elements* with value *C*. The patterns discovered by the IMB3 algorithm can aid in discovering potentially useful pattern structures in Protein Ontology datasets, which makes it useful for comparison of protein datasets taken across protein families and species and helps in discovering interesting similarities and differences. However, the question still remains whether these patterns are discovered due to pure coincidence resulting from certain randomness in the particular dataset being analyzed. Furthermore, they are often quite large in number, which can degrade the analysis procedure, and hence in the next section we measure the statistical significance of the discovered patterns, in order to remove any non-significant patterns.

4.2 Frequent Patterns Significant Test

Statistical analysis approaches, namely Chi-Squared and Log-Linear analysis were employed in order to determine the usefulness of frequent rules obtained. The results from Chi-Squared analysis are discussed first.

Table 2. Patterns Verification Based on Chi-Squared Analysis

Node Name		Sig. Att. Value
ATOMResidual(TYR)	Occupancy(1)	Not Sig.
Occupancy(1)	Temperature(0)	Not Sig.
ATOMChain(A)	Occupancy(1)	Not Sig.
ATOMChain(A)	Element(C)	Sig.
ATOMChain(A)	Element(H)	Sig.
Temperature(0)	Element(C)	Sig.
Temperature(0)	Element(H)	Sig.
ATOMChain(A)	ATOMResidual(TYR)	Sig.
ProteinOntologyID(3)	Occupancy(1)	Not Sig.
ProteinOntologyID(3)	Element(C)	Sig.
ATOMChain(A)	Temperature(0)	Sig.
Occupancy(1)	Temperature(1)	Not Sig.
Occupancy(1)	Element(N)	Not Sig.
Occupancy(1)	Element(C)	Not Sig.
Occupancy(1)	Element(O)	Not Sig.
Occupancy(1)	Element(H)	Not Sig.

Table 2 shows that, there are 16 association relationships among structures-values items discovered using the IMB3 algorithm. Based on Chi-Squared analysis, 7 out of 16 relationships are significant. Table 3 shows 11 patterns with more than two nodes. We apply the Log-Linear analysis in examining the association between these nodes. Only one pattern out of 11 patterns is accepted as a significant pattern based on this analysis. Based on the Log-Linear analysis, we can conclude that, there is significant association between *ATOMChain(A)*, *Temperature(0)* and *Element(H)*.

Table 3. Patterns Verification Based on Log-Linear Analysis

Node Name				Sig. Att. Value
ATOMChain(A)	ATOMResidue(TYR)	Occupancy(1)		Not Sig.
ATOMChain(A)	Occupancy(1)	Temperature(0)		Not Sig.
ATOMChain(A)	Occupancy(1)	Element(C)		Not Sig.
ProteinOnto(3)	Occupancy(1)	Element(C)		Not Sig.
ATOMChain(A)	Occupancy(1)	Element(H)		Not Sig.
Occupancy(1)	Temperature(0)	Element(C)		Not Sig.
ATOMChain(A)	Temperature(0)	Element(C)		Not Sig.
Occupancy(1)	Temperature(0)	Element(H)		Not Sig.
ATOMChain(A)	Temperature(0)	Element(H)		Sig.
ATOMChain(A)	Occupancy(1)	Temperature(0)	Element(C)	Not Sig.
ATOMChain(A)	Occupancy(1)	Temperature(0)	Element(H)	Not Sig.

4.3 Prions as a Classification Problem

As in our previous work [11], the unification framework involves several steps in ascertaining the rules discovered from association rules mining process. For Prions dataset, the similar steps were followed. We defined a new variable (target variable) identified as Human Protein or Animal Protein class. This new variable was derived from ProteinOntologyID and SuperFamily variables. Hence, we have excluded the

ProteinOntologyID and SuperFamily variables from the dataset to be considered in this task. Thus in this classification problem we have chosen the target variable (i.e. Human or Animal's Protein) as the right hand side/consequence of the association rules.

In this experiment, we divided the Prions dataset into 60% of training set and 40% of testing set. Then we apply the preprocessing techniques including the missing values removal and discretization of attributes with continuous data. The equal depth binning approach method was selected as this approached offered a better result as discussed in [11]. The determination of relevant attributes with respect to being able to predict the target attributes is shown in Table 4. This is based on Symmetrical Tau [43] and Mutual Information [12] techniques. As discussed in [11], the Symmetrical Tau (ST) approach offers better output in discriminating criterions for class to be predicted in comparison to Mutual Information (MI), as it does not favor multi-valued attributes. The attributes with ST values that are respectively lower than other attribute's ST values, are considered as irrelevant for the task. The significant difference was considered to occur at the position where that attribute's ST value is less than half of the previous attribute's ST value in the ranking. Hence for this dataset, attributes '*Occupancy*' and '*Y*' were considered as irrelevant for the prediction task and were removed.

Table 4. Comparison between ST and MI for Prions Dataset

Variables	ST Values		Variables	MI Values
ATOMChain	0.2088		ATOMChain	0.2605
Temperature	0.1230		Z	0.1610
Z	0.0812		Temperature	0.1526
ATOMid	0.0407		ATOMResSeqNum	0.1053
ATOMResSeqNum	0.0280		ATOMid	0.0721
X	0.0256		X	0.0549
Element	0.0153		Atom	0.0238
Atom	0.0109		ATOMResidue	0.0187
ATOMResidue	0.0082		Element	0.0162
Y	0.0029		Y	0.0048
Occupancy	0.0001		Occupancy	0.0000

Table 5. Examples of Prions Rules

Set Size	Confidence	Support	Count	Rules
2	75.32	8.97	934	X(g) ==> Class (Animal)
4	61.71	6.66	693	X(d) & Z(b) & ATOMChain(A) ==> Class (Human)

Next, the rules are then generated based on the minimum support and confidence framework of 5% and 60% respectively. Table 5 shows examples of the generated rules. The discovered rules are then ascertained with statistical techniques namely Chi Squared [20] and Logistics Regression [20]. Based on these statistical analyses we found that only variables *ATOMChain*, *ATOMResidual*, *ATOMResSeqNum*, *X* and *Z* were significant contributors towards target variable of class Human or Animal.

Additional constraint measurement techniques were applied in order to discard the existence of redundant rules [11, 13]. The combination of these rule ascertaining strategies will facilitate the association rule mining framework to determine the right and high quality rules. These rules will have a sound statistical basis and we can be more confident that they reflect the real world situation.

In Table 6 we show the progressive difference in the number of rules generated as statistical analysis and redundancy checks are being utilized. We also show the respective classification (% of correctly classified instances from the training set) and predictive accuracy (% of correctly classified instances from the training set) of those rule sets. Upon a removal of 73% rules, we found that both classification and predictive accuracies have increased by more than 5%. This demonstrates the importance of ascertaining the association rules by statistical analysis and redundancy check, as in this particular scenario the simplified rule set is more general and performs better on unseen data.

The combination of statistical significance analysis and redundant analysis provided proper ways in discarding non significant rules, which is a significant reduction in the overall complexity of the rule set. From Table 6 we can also see that this great reduction of rules was not at a cost of a reduction in accuracy, as it in fact increased for the Prions dataset in classifying and predicting the protein classes.

Table 6. Rules Accuracy for Prions Data

*Dataset Description	Rule #	Type of Analysis	Accuracy	
			Classification	Prediction
Train : 10407 records Test : 6938 records	42 11	Initial rules Statistical Analysis / Redundancy Check	74.36% 79.97%	75.00% 80.37%

* Two records with missing values were discarded.

5 Conclusions and Future Works

This was our preliminary work towards the combination of data mining and statistical techniques in ascertaining the rules/patterns from semi-structured data. The combination of the approaches used in this method demonstrated a number of ways for ascertaining the significant patterns obtained using frequent subtree mining approaches. In this paper we employed statistical analysis that provides some control in lowering the risk of discovering a pattern that is false and spurious. In our future work we aim to test the approach using tree-structured data of various characteristics and complexities.

References

1. Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann, San Francisco (2001)
2. Zhang, J., Ling, T.W., Bruckner, R.M., Tjoa, A.M., Liu, H.: On Efficient and Effective Association Rule Mining from XML Data. In: Proceedings of the 15th Int. Conf. Database and Expert Systems Applications, Zaragoza, Spain, pp. 497–507 (2004)
3. Chi, Y., Muntz, R.R., Nijssen, S., Kok, J.N.: Frequent Subtree Mining - An Overview. Fundamenta Informaticae 661, 61–198 (2005)

4. Agrawal, R., Imieliski, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Rec. 22, 207–216 (1993)
5. Aggarwal, C.C., Yu, P.S.: A new framework for itemset generation. In: Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 18–24. ACM, Washington (1998)
6. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the 20th Int. Conf. on Very Large Data Bases, Santiago, Chile (1994)
7. Toivonen, H.: Sampling Large Databases for Association Rules. In: Proceedings of the 20th Int. Conf. on Very Large Data Bases, Mumbai, India, pp. 134–145 (1996)
8. Bayardo, R., Agrawal, R., Gunopulos, D.: Constraint-Based Rule Mining in Large, Dense Databases. J. Data Mining and Knowledge Discovery 4, 217–240 (2000)
9. Lavrač, N., Flach, P., Zupan, B.: Rule Evaluation Measures: A Unifying View. In: Džeroski, S., Flach, P.A. (eds.) ILP 1999. LNCS (LNAI), vol. 1634, pp. 174–185. Springer, Heidelberg (1999)
10. Lenca, P., Meyer, P., Vaillant, B., Lallich, S.: On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. European Journal of Operational Research 184, 610–626 (2008)
11. Shaharanee, I.N.M., Hadzic, F., Dillon, T.: Interestingness of Assocaition Rules Using Symmetrical Tau and Logistic Regression. In: Nicholson, A., Li, X. (eds.) AI 2009. LNCS (LNAI), vol. 5866, pp. 422–431. Springer, Heidelberg (2009)
12. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 32–41. ACM, Alberta (2002)
13. Webb, G.I.: Discovering Significant Patterns. In: Machine Learning, pp. 1–33. Springer, Heidelberg (2007)
14. Yun, H., Ha, D., Hwang, B., Ho Ryu, K.: Mining association rules on significant rare data using relative support. J. Systems and Software 67, 181–191 (2003)
15. Tan, H., Hadzic, F., Dillon, T.S., Chang, E.: State of the art of data mining of tree structured information. Int. Journal of Computer Systems Science and Engineering 23 (2008)
16. Asai, T., Abe, K., Kawasoe, S., Arimura, H., Sakamoto, H., Arikawa, S.: Efficient Substructure Discovery from Large Semi-structured Data. In: Proc. of the 2nd SIAM Int. Conf. on Data Mining (SIAM 2002), pp. 158–174 (2002)
17. Feng, L., Dillon, T., Weigand, H., Chang, E.: An XML-Enabled Association Rule Framework. Database and Expert Systems Applications, 88–97 (2003)
18. Tan, H., Hadzic, F., Dillon, T.S., Chang, E., Feng, L.: Tree model guided candidate generation for mining frequent subtrees from XML documents. ACM Trans. Knowl. Discov. Data 2, 1–43 (2008)
19. Goodman, A., Kamath, C., Kumar, V.: Data Analysis in the 21st Century. Stat. Anal. Data Mining 1, 1–3 (2008)
20. Agresti, A.: An Intro. to Categorical Data Analysis. Wiley Interscience, New Jersey (2007)
21. Tan, H., Dillon, T., Hadzic, F., Chang, E., Feng, L.: IMB3-Miner: Mining Induced/ Embedded Subtrees by Constraining the Level of Embedding. In: Proceedings of the 8th Pacific-Asia Conference on Knowl. Discovery and Data Mining, pp. 450–461 (2006)
22. Fedja, H., Tharam, S.D., Elizabeth, C.: Knowledge Analysis with Tree Patterns. In: Proceedings of the 41st Annual Hawai Int. Conf. on System Sciences. IEEE, Los Alamitos (2008)
23. Fedja, H., Henry, T., Tharam, S.D.: U3 - Mning Unordered Embedded Subtrees Using TMG Candidate Generation. In: The 1st ACM Int. Conf. on Web Search and Data Mining, California, USA (2008)

24. Tan, H., Dillon, T., Hadzic, F., Chang, E., Feng, L.: MB3-Miner: efficiently mining eMBedded subTREEs using Tree Model Guided candidate generation. In: Proceedings of the 1st Int. Workshop on Mining Complex Data 2005, Texas, USA (2005)
25. Zaki, M.J.: Efficiently Mining Frequent Trees in a Forest: Algorithms and Applications. IEEE Transactions on Knowledge and Data Engineering 17, 1021–1035 (2005)
26. Aumann, Y., Lindell, Y.: A Statistical Theory for Quantitative Association Rules. J. Intell. Inf. Syst. 20, 255–283 (2003)
27. Meggido, N., Srikant, R.: Discovering Predictive Association Rules. In: 4th International Conference on Knowledge Discovery in Databases and Data Mining, pp. 274–278 (1998)
28. Webb, G.I.: Preliminary investigations into statistically valid exploratory rule discovery. In: Simoff, S.J., Williams, G.J., Hegland, M. (eds.) AusDM 2003, Sydney, pp. 1–9 (2003)
29. Ozaki, T., Ohkawa, T.: Mining Mutually Dependent Ordered Subtrees in Tree Databases. In: New Frontiers in Applied Data Mining: PAKDD 2008 Int. Workshops, Japan, Revised Selected Papers, pp. 75–86. Springer, Heidelberg (2009)
30. Hui, X., Pang-Ning, T., Vipin, K.: Hyperclique pattern discovery. Data Min. Knowl. Discov. 13, 219–242 (2006)
31. Bathoorn, R., Koopman, A., Siebes, A.: Reducing the Frequent Pattern Set. In: Proceedings of the 6th IEEE International Conference on Data Mining – Workshops, pp. 55–59 (2006)
32. Siebes, A., Vreeken, J., Leeuwen, M.V.: Item Sets That Compress. In: Proceedings of the SIAM Conference on Data Mining, Maryland, USA, pp. 393–404 (2006)
33. Nakamura, A., Kudo, M.: Mining Frequent Trees with Node-Inclusion Constraints. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 850–860. Springer, Heidelberg (2005)
34. Srikant, R., Vu, Q., Agrawal, R.: Mining Association Rules with Item Constraints. In: 3rd Int. Conf. on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, pp. 67–73 (1997)
35. Knijf, J.D., Feelders, A.J.: Monotone Constraints in Frequent Tree Mining. In: Poel, M., Nijholt, A. (eds.) BENELEARN, Enschede, The Netherlands, pp. 13–20 (2005)
36. Fedja, H., Henry, T., Tharam, D.: Mining Unordered Distance-Constrained Embedded Subtrees. In: Boulicaut, J.-F., Berthold, M.R., Horváth, T. (eds.) DS 2008. LNCS (LNAI), vol. 5255, pp. 272–283. Springer, Heidelberg (2008)
37. Rusu, L.I., Rahayu, W., Taniar, D.: Extracting Variable Knowledge from Multiversioned XML Documents. In: 6th IEEE International Conference on Data Mining - Workshops (ICDMW 2006), pp. 70–74 (2006)
38. Yan, X., Cheng, H., Han, J., Yu, P.S.: Mining significant graph patterns by leap search. In: SIGMOD Conference, Canada, pp. 433–444 (2008)
39. Hashimoto, K., Takigawa, I., Shiga, M., Kanehisa, M., Mamitsuka, H.: Mining significant tree patterns in carbohydrate sugar chains. Bioinformatics 24, 167–173 (2008)
40. Zaki, M.J., Aggarwal, C.C.: XRules: an effective structural classifier for XML data. In: SIGKDD 2003, Washington, DC (2003)
41. Wu, X., Barbar, D., Ye, Y.: Screening and interpreting multi-item associations based on log-linear modeling. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, Washington (2003)
42. Fedja, H., Tharam, S.D., Amandeep, S.S., Elizabeth, C., Henry, T.: Mining Substructures in Protein Data. In: Proceedings of the 6th IEEE Int. Conf. on Data Mining-Workshops (2006)
43. Zhou, X.J., Dillon, T.S.: A statistical-heuristic feature selection criterion for decision tree induction. IEEE Transactions on Pattern Analysis and Machine Intelligence (1991)