Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 4, No. 3, August 2013), Pages: 25-32
www.jacr.iausari.ac.ir

Discovering Users` Access Patterns for Web Usage Mining from Web Log Files

Maryam Jafari[1 *], Shahram Jamali[2], Farzad Soleymani Sabzchi[1]

(1) Department of Computer Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran
(2) Computer Engineering Department, University of MohagheghArdabili, Ardabil, Iran

m123_jafari@yahoo.com; jamali@iust.ac.ir; f_soleymani63@yahoo.com

Received: 2012/12/19; Accepted: 2013/02/24

Abstract

Web Usage Mining (WUM) is the automatic discovering of hidden information of user access pattern from the web log data. Frequent pattern discovery is one of the main techniques in WUM that can be used to implement recommender systems, forecast user`s navigational behaviour, and personalize web sites. Many algorithms have been suggested on obtaining frequent user navigation patterns. This paper presents PD-FARM (Pattern Discovery based on Fuzzy Association Rule Mining) algorithm to extract the web usage patterns, based on Fuzzy Association Rule Mining (FARM). Fuzzy Frequent Pattern-Growth (FFP-Growth) algorithm is used to FARM. Finally, an illustrated example is given for a complete description of the proposed algorithm.

Keywords: *Web usage mining, Pattern discovery, Navigational pattern, Fuzzy FP-Growth algorithm*

1. Introduction

Web mining is extracting valuable knowledge from Web data that can be broadly defined as the discovery and analysis of useful information from the World Wide Web [1]. In [2,3,4] web mining is divided into three distinct domains: Web content mining, Web structure mining and Web usage mining. Web content mining involves efficient extracting knowledge from the content of documents from a large number of Web sites and databases. Web structure mining is the process of gathering knowledge from the web pages organization and studying the Web pages schema of a collection of hyper-links. Web usage mining, also known as Web Log Mining, involves the analysis and discovery of user's interesting patterns from web browsing data that are stored in web server logs, proxy server logs or browser logs. It focuses on the techniques that could predict user`s navigational behavior. In [5], there are three main tasks for performing Web Usage Mining: Preprocessing, Pattern Discovery, and Pattern Analysis. Many papers focus on describing Pattern Discovery phase and the techniques of this phase. One of these techniques is association rule mining. In the past, many algorithms for extracting association rules from users` interactions were proposed such as Apriori algorithm and Frequent Pattern-tree (FP-tree) [6,7].

In this paper, a novel algorithm called PD-FARM is proposed for FP-tree mining process. This algorithm uses fuzzy FP-tree to find fuzzy association rules and obtain

desired access patterns from a database that contains users' sessions. A fuzzy FP-tree is a data structure keeping frequent fuzzy regions.

The remainder of this paper is organized as follows. In the next section motivation of this research is expressed. An overview over the related works is given in section 3. The proposed algorithm is described in section 4. An example to illustrate the proposed algorithm is given in section 5. Finally the conclusion of the work is presented in section 6.

2. Motivation of the research

Pattern discovery is an important phase in web usage mining which extracts the user's behavioral patterns from the formatted data. If this phase generates more interested patterns, the next phase (i.e. Pattern Analysis) will be more efficient. Web personalization, generating recommendation in recommender systems for active user and improving structure of the web sites are heavily depended on the extracted and analyzed patterns. By applying fuzzy partitioning method on the user`s sessions and extracting fuzzy association rules, the degree of user's interest visited pages will be calculated more accurately. Consequently, the patterns which are discovered by using this method would be more interesting.

3. Related Works

Etzioni [8] proposed a new concept in 1996 that was called "Web mining". He used data mining techniques to automatic discovery and extract information from abundant data on the World Wide Web. Web Usage Mining (WUM) was first proposed by Chen et al. [9,10], Mannila and Toivonen [11], and Yan et al. [12]. Baraglia and Palmerini presented a WUM system called SUGGEST which optimizes the web server performance by providing useful information. This system provides an objective behavior for user navigation. Jianhan Zhu et al., [13] used the Markov chains to model user`s navigational behavior. They proposed a method for building a Markov model of a web site based on previous users` behavior. Then the Markov model was used to make link predictions that help new users to navigate the web site. Jalali et al., [14] presented a system for extracting user`s navigational behavior using a graph partitioning model. An undirected graph based on connectivity between each pair of web pages was considered and also proposed a new formula for allocating weights to edges of the graph.

In [15] the prediction of user`s navigation patterns, using clustering and classification from web log data is proposed. First phase of this method focuses on separating the users in web log data, and in the second phase clustering process is used to group the users with similar preferences and in the third phase the results of classification and clustering are used to predict the users` next requests. Emine Tug et al., [16] found sequential accesses from web log files, using Genetic Algorithm (GA) that was called Automatic Log Mining via Genetic (ALMG). In their work, GA based on evolutionary approach for pattern extraction was used to found the best solutions for time consuming problem to discover sequential accesses from web log data.

4. Proposed Algorithm: PD-FARM

In this section, PD-FARM algorithm is presented for Pattern Discovery (PD) based on Fuzzy Association Rules Mining (FARM). This method uses Frequent Pattern-Growth (FP-Growth) algorithm.Before that, general concepts of the proposed algorithm are described.

The new concept of Frequent Pattern-tree (FP-tree) structure and FP-growth algorithm were proposed by Han et al. [17] for the efficient mining association rules which do not require candidate generations. The mining algorithm consisted of two phases. The first phase focused on constructing the FP-tree from a database, and the second phase focused on deriving frequent patterns from the FP-tree.

In this paper, fuzzy FP-tree mining algorithm finds fuzzy association rules from quantitative transaction data. The fuzzy FP-tree construction algorithm is designed to generate the tree structure for frequent fuzzy regions (terms). It first transforms the quantitative values of attributes in transactions into linguistic terms. For each term, only the linguistic term with the maximum cardinality is used. The frequent fuzzy items represented by linguistic terms, are then derived from the fuzzy FP-tree. However, while extending the crisp FP tree to the fuzzy one, the processing one becomes much more complex than the original, because fuzzy intersection in each transaction has to be handled. The paper utilizes fuzzy partition method using CURE clustering algorithm [6] in which each user`s session was treated as a linguistic variable that is a suitable number to partition. Next, applies FP-Growth in the first scan to find out regions with maximum count for each visited page. In second scan, a FFP-tree is established. Further, condition pattern base and conditional FFP-tree are established, respectively. Finally, it uses a recursive method to extract all fuzzy association rules according to min FC. You can find PD-FARM`s details as follows:

***PD-FARM Algorithm**:*

Input:

- A data base of n users sessions;
- Predefined minimum fuzzy support (min FS) for extracting fuzzy association rules;
- Predefined minimum fuzzy confidence (min FC) for extracting fuzzy association rules;
- A membership function for each linguistic value θ;

Output:

- A set of fuzzy association rules.

Steps of Algorithm:

Step 1: First, apply fuzzy partition method to partition items into linguistic values.

Step 2: After partitioning, sum the scalar cardinalities of every fuzzy region as the count value.

Step 3: Select a region with maximum count value from fuzzy regions of each item.

Step 4: In the first scan, if Maximum Count Value ≥ (min FS × n) then insert this value in the Header table.

Step 5: Sort the items of the Header table.

Step 6: According to the sorted Header table, build a new fuzzy set of transaction table.

Step 7: In second scan, establish a Fuzzy FP-tree based on the new table.

Step 8: Starting from the last item of the Header table, create a conditional FFP-tree for each node of FFP-tree of conditional pattern base. Next, repeat FFP-tree process mining, and mine frequent pattern of conditional FFP-tree. Finally, if the conditional FFP-tree is contained one path, it will list all patterns.

Step 9: Using the following substeps, mine fuzzy association rules:

Substep 9.1: List all frequent itemsets.

Substep 9.2: Calculate the supports and confidences for all frequent itemsets.
Substep 9.3: For each frequent itemset if (support ≥ min FS && confidence ≥ min FC) then it should be accepted as a fuzzy association rule.

5. Illustrative Example

In this section, an example is given to illustrate how to extract navigation patterns from a database of users` sessions. This database shown in Table 1 that consists of ten users` sessions and six web pages denoted P_1, P_2, P_3, P_4, P_5 and P_6. Each item is represented by a tuple (page name: visit time). The membership function used in this example is drawn in Fig. 1. In this example, values are represented by four fuzzy regions: *Short, Middle, Long and Too Long*. Thus, four fuzzy membership values are produced for each page in a session according to the predefined membership function. Here, assume that the predefined min FS and min FC are 0.20 and 0.60, respectively. By adopting membership function of Fig. 1, we only proceed with explanation and produce the result. The algorithm can be described as follows:

Step 1: Membership function in Fig. 1 of fuzzy partition method is applied which item attributes proceeded with partition, and the results of fuzzy set are gotten after partition. Take page P_1 in session 1 as an example. The value ‘‘83” of P_1 is converted into the fuzzy set P_1 (0.59, 0.41, 0, 0). This step is repeated for the other pages, and the results are shown in Table 2.

Step 2: The scalar cardinality of each fuzzy region in the session is calculated as the count value. Take the fuzzy region P_1.Short as an example. Its scalar cardinality is (0.59+0.71+0.67+0.4+0.53+0.57+0.31) = 3.78. This step is repeated for the other regions, and the results are shown in Table 3 where the notation page.term is called a fuzzy region.

Step 3: The fuzzy region with the maximum count value among the four possible regions for each page is found. Take page P_1 as an example. Its count is 3.78 for Short, 3.23 for Middle, 0.0 for Long and 0.0 for Too Long. Since the count for Short is the maximum among the four counts, the region Short is thus used to represent page P_1 in the later mining process. This step is repeated for the other pages. Thus, region Middle is chosen for P_2, P_3, and P_4, and region Long is chosen for P_5 and P_6. Note that Region Too Long has been selected for no page.

Table 1. The ten sessions used in the example

Session ID	Page Access Sequence
1	(P1: 83), (P3: 355), (P4: 191), (P6: 378)
2	(P2: 207), (P3: 267), (P4: 455), (P5: 407), (P6: 411)
3	(P1: 59), (P2: 181), (P5: 233), (P6: 256)
4	(P1: 67), (P2: 391), (P3: 271), (P4: 241), (P5: 291), (P6: 501)
5	(P2: 173), (P5: 302), (P6: 598)
6	(P1: 120), (P3: 107), (P4: 219), (P5: 391)
7	(P2: 132), (P3: 62), (P4: 141), (P5: 358), (P6: 311)
8	(P1: 95), (P2: 283), (P3: 211), (P6: 279)
9	(P1: 87), (P3: 93), (P4: 257), (P6: 351)
10	(P1: 139), (P2: 98), (P3: 148), (P5: 183), (P6: 393)

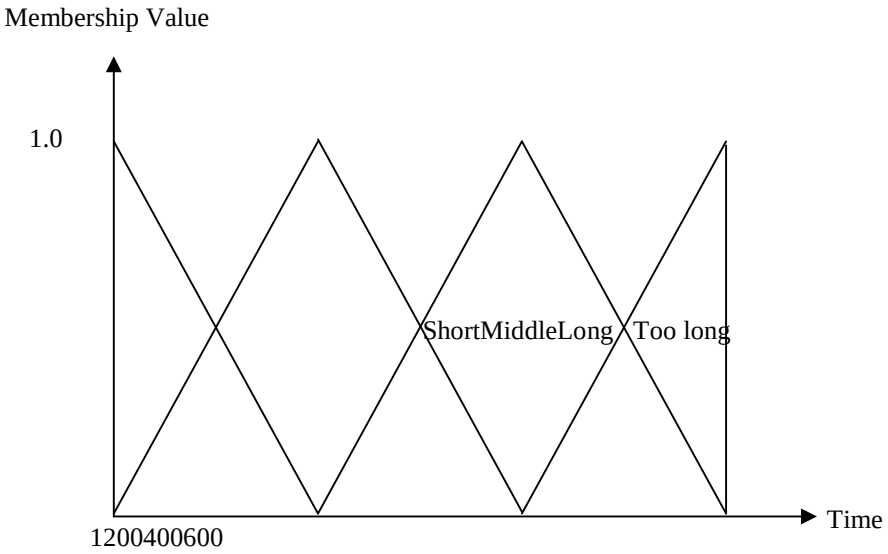

Figure 1. The membership function used in the example

Table 2. The fuzzy sets after partition: Pi (Short ,Middle ,Long ,Too Long)

SID	Page Access Sequence
1	P_1 (0.59,0.41,0,0), P_3 (0,0.23,0.77,0), P_4 (0.05,0.95,0,0), P_6 (0,0.11,0.89,0)
2	P_2 (0,0.96,0.04,0), P_3 (0,0.66,0.34,0), P_4 (0,0,0.72,0.28), P_5 (0,0,0.96,0.04), P_6 (0,0,0.94,0.06)
3	P_1 (0.71,0.29,0,0), P_2 (0.09,0.91,0,0), P_5 (0,0.83,0.17,0), P_6 (0,0.72,0.28,0)
4	P_1 (0.67,0.33,0,0), P_2 (0,0.05,0.95,0), P_3 (0,0.65,0.35,0), P_4 (0,0.79,0.21,0), P_5 (0,0.54,0.46,0), P_6 (0,0,0.5,0.5)
5	P_2 (0.14,0.86,0,0), P_5 (0,0.5,0.5,0), P_6 (0,0,0.01,0.99)
6	P_1 (0.4,0.6,0,0), P_3 (0.47,0.53,0,0), P_4 (0,0.91,0.09,0), P_5 (0,0.05,0.95,0)
7	P_2 (0.34,0.66,0,0), P_3 (0.7,0.3,0,0), P_4 (0.3,0.7,0,0), P_5 (0,0.21,0.79,0), P_6 (0,0.45,0.55,0)
8	P_1 (0.53,0.47,0,0), P_2 (0,0.58,0.42,0), P_3 (0,0.94,0.06,0), P_6 (0,0.61,0.39,0)
9	P_1 (0.57,0.43,0,0), P_3 (0.54,0.46,0,0,), P_4 (0,0.72,0.28,0), P_6 (0,0.25,0.75,0)
10	P_1 (0.31,0.69,0,0), P_2 (0.51,0.49,0,0), P_3 (0.26,0.74,0,0), P_5 (0.08,0.92,0,0), P_6 (0,0.04,0.96,0)

Step 4: The maximum count value of any region is checked against the predefined min FS. In this example, min FS is 20%. Since the count values of P_1.Short, P_2.Mid, P_3.Mid, P_4.Mid, P_5.Long and P_6.Long are larger than $10 \times 20\% = 2$, these fuzzy regions are put in the Header table shown in Fig. 2.

Step 5: The frequent fuzzy regions of the pages in Header table are sorted in the descending order of their counts which will be used to construct the fuzzy FP- tree later.

Step 6: According to the Header table, fuzzy set of the transaction table is rebuild from Table 2. For example, the new fuzzy set of session 1 is: P_6.Long:0.89, P_3.Mid:0.23, P_4.Mid:0.95, P_1.Short:0.59.

Step 7: In the second process of scan, a fuzzy FP-tree is established, shown as Fig. 2.

Step 8: The pages of Header table are mined sequence in descending order, and a conditional FFP-tree (Part of the tree that contains the path leading to the examined node) is established for each node of Header table.

Table 3. The counts of fuzzy regions

Page	Count	Page	Count	Page	Count
P_1.Short	3.78	P_3.Short	1.96	P_5.Short	0.09
P_1.Middle	3.23	P_3.Middle	4.52	P_5.Middle	3.04
P_1.Long	0.0	P_3.Long	1.5	P_5.Long	3.84
P_1.Too Long	0.0	P_3.Too Long	0.0	P_5.Too Long	0.04
P_2.Short	1.09	P_4.Short	0.34	P_6.Short	0.0
P_2.Middle	4.51	P_4.Middle	4.07	P_6.Middle	2.16
P_2.Long	1.41	P_4.Long	1.31	P_6.Long	5.29
P_2.Too Long	0.0	P_4.Too Long	0.27	P_6.Too Long	1.55

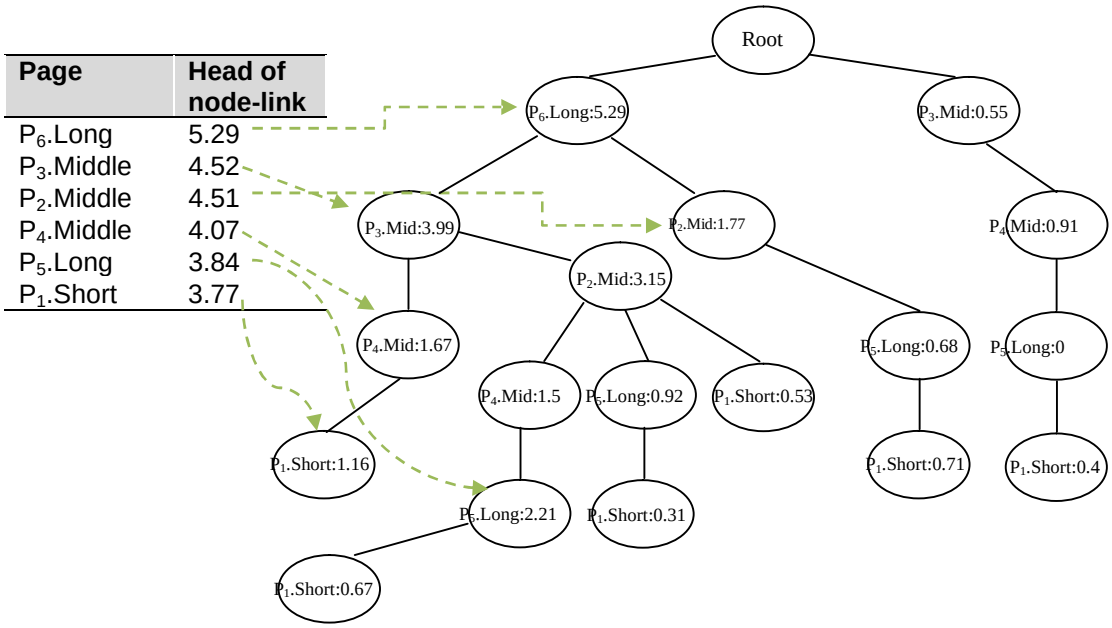

Fig. 2. The fuzzy FP tree constructed in the example

Next, the mining of FFP-tree is repeated which is led to include a frequent pattern of conditional FFP-tree. Finally, if the conditional FFP-tree is contained one particular path, all patterns are listed. According to this step, conditional patternsare shown in Table 4.

Step 9: The substeps are used to discover fuzzy association rules as bellow:

Substeps 9.1 and 9.2: All frequent itemsets are listed with their confidences.

P_5.Long → P_6.Long (support = 0.25, confidence = 0.65)
P_5.Long → P_3.Middle (support = 0.24, confidence = 0.63)
P_5.Long → P_2.Middle (support = 0.27, confidence = 0.70)
P_1.Short → P_6.Long (support = 0.20, confidence = 0.53)
P_4.Middle → P_6.Long (support = 0.22, confidence = 0.54)

P_2.Middle → P_6.Long (support = 0.25, confidence = 0.55)
P_2.Middle → P_3.Middle (support = 0.21, confidence = 0.45)
P_3.Middle → P_6.Long (support = 0.28, confidence = 0.62)

Table 4. Conditional pattern base

Conditional patterns	Fuzzy values	Conditional patterns	Fuzzy values
P1.Short	3.77	P6.Long,P3.Mid	3.99
P5.Long	3.84	P4.Mid,P5.Long	2.21
P4.Mid	4.07	P3.Mid,P1.Low	3.06
P2.Mid	4.51	P6.Long,P4.Mid	3.17
P3.Mid	4.52	P3.Mid,P2.Mid	3.15
P6.Long	5.3	P6.Long,P2.Mid,P5.Long	3.84
P6.Long, P1.Low	3.36	P6.Long,P3.Mid,P5.Long	3.13
P6.Long,P5.Long	3.8	P3.Mid,P2.Mid,P5.Long	3.13
P3.Mid,P5.Long	3.125	P3.Mid,P4.Mid,P5.Long	2.21
P2.Mid,P5.Long	3.84	P6.Long,P3.Mid,P4.Mid	3.17
P3.Mid,P4.Mid	4.1	P6.Long,P3.Mid,P2.Mid	3.15
P6.Long,P2.Mid	4.5	P6.Long,P3.Mid,P2.Mid,P5.Long	3.13

Substep 9.3: Since min FS = 0.20 and min FC = 0.60 and concerning the given sessions, the three rules can then serve as meta-knowledge:

{P_5.Long → P_6.Long},

{P_5.Long → P_3.Middle},

{P_5.Long → P_2.Middle},

{P_3.Middle → P_6.Long}.

In order to clarify the proposed approach, consider P_5.Long → P_6.Long rule as an example. Its confidence is 0.65. If the access time for page P_5 is long, then duration on the page P_6 is long. The rule reflects not only the association between web page P_5 and page P_6, but also it reflects their dependency from browsing duration point of view.

6. Conclusion

As the rapid growth of data on the web, discovery and analysis of useful information from the web have become increasingly important tasks. In this paper, the fuzzy FP-tree construction algorithm was proposed for mining frequent fuzzy item sets from the users` sessions and then finding fuzzy association rules by using the fuzzy partition method and FP-growth algorithm. Whether a web page is visited or not and visit time of the page are considered as the two important factors to show users' interest. Thus, fuzzy association rule mining not only reflects the relationship between web pages, but also relates to the time duration on these web pages. The fuzzy FP-tree structure is used to handle the page visit time efficiently and effectively for a deeper understanding of user navigational behavior.

When extending the FP-tree to handle fuzzy data, the processing becomes much more complex than the original. However the features of the proposed algorithm prove that this method does not need to generate candidate item sets and improves the efficiency of repetitious database scanning in Apriori algorithm. As the direction of

future works in this research area, this algorithm needs to be improved from storage space point of view.

7. References

[1] Cooley R, Mobasher B, Srivastava J. Web Mining: Information and Pattern Discovery on the World Wide Web. IEEE International conference on Tools with Artifical Intelligence 1997;9:558 – 567.

[2] Kosala R, Blockeel H. Web mining research: A survey. Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining 2000;2:1-15.

[3] Madria S, Bhowmick S, Ng W, Lim E. Research issues in web data mining. Data Warehousing and Knowledge Discovery 1999;1:303-312.

[4] Borges J, Levene M. Data mining of user navigation patterns. Workshop on Web Usage Analysis and User Profiling 2000;92-111.

[5] Srivastava J, Cooley R, Deshpande M, Tan P. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations 2000;2:12-23.

[6] Kaya M, Alhajj R, Polat F, Arslan A. Efficient Automated Mining of Fuzzy Association Rules. International Conference on Database and Expert Systems Applications 2002;13:133-142.

[7] Han J, Pei J, Yin Y, Mao R. Mining frequent patterns without candidate generation. ACM SIGMOD international conference on management of data 2000;29:1–12.

[8] Etzioni O. The world-wide web: Quagmire or gold mine. Communications of the ACM 1996;39:65–68.

[9] Chen M, Park J, Yu P. Data mining for path traversal patterns in a Web environment. International Conference on Distributed Computing Systems 1996;19:385–392.

[10] Chen M, Park J, Yu P. Efficient data mining for path traversal patterns. IEEE Transactions on Knowledge and Data Engineering 1998;10:209–221.

[11] Mannila H, Toivonen H. Discovering generalized episodes using minimal occurrences. International Conference on Knowledge and Data Mining 1996;2:146–151.

[12] Yan T, Jacobsen M, Garcia-Molina H, Dayal U. From user access patterns to dynamic hypertext linking. International World Wide Web conference on Computer networks and ISDN systems 1996;5:1007-1014.

[13] Zhu J, Hong J, Hughes J. Using Markov Chains for Link Prediction in Adaptive Web Sites. Lecture Notes in Computer Science 2002; 60–73.

[14] Jalali M, Mustapha M, Mamat A, Sulaiman M. A new clustering approach based on graph partitioning for navigation patterns mining. International Conference on Pattern Recognition 2008;9:1-4.

[15] Sujatha V, Punithavalli. Improved User Navigation Pattern Prediction Technique From Web Log Data. International Conference on Communication Technology and System Design 2011;92– 99.

[16] Tug E, Sakiroglu M, Arslan A. Automatic discovery of the sequential accesses from web log data files via a genetic algorithm. Knowledge-Based Systems 2006;19:180–186.

[17] Han J, Pei J, Yin Y, Mao R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery 2004;8:53–87.