FP Growth Algorithm Implementation

Shivam Sidhu
Department of Computer Science, Bharati Vidyapeeth's College of Engineering, New Delhi, India

Upendra Kumar Meena
Department of Computer Science, Bharati Vidyapeeth's College of Engineering, New Delhi, India

Aditya Nawani
Department of Computer Science, Bharati Vidyapeeth's College of Engineering, New Delhi, India

Himanshu Gupta
Department of Computer Science, Bharati Vidyapeeth's College of Engineering, New Delhi, India

Narina Thakur
Department of Computer Science, Bharati Vidyapeeth's College of Engineering, New Delhi, India

ABSTRACT
Data mining is to discover and assess significant patterns from data, followed by the validation of these identified patterns. Data mining is the process to evaluate the data from different perceptions and summarizing it into valuable information. This summarized information consequently can be used to design business strategies to upsurge revenue, occasionally drive down costs, or both. The Apriori association algorithm is based on pre-computed frequent item sets and it has to scan the entire transaction log / dataset or database which will become a problem with large item sets. With FP trees, there is no necessity for candidate generation, unlike in the Apriori algorithm, and the frequently occurring item sets are discovered by just traversing the FP tree. This paper discusses the FP Tree concept and implements it using Java for a general social survey dataset. We use this approach to determine association rules that occur in the dataset. In this manner, we can establish relevant rules and patterns in any set of records.

General Terms
Associaiton, FP Growth Algorithm

Keywords
Data mining, Frequent Pattern Tree, Apriori, Association

1. INTRODUCTION
In the data mining section we would discuss the different types of data mining techniques such as association. After a brief discussion on each type, the concept of FP Trees (which comes under Association type) would be discussed in detail. The main reason of the popularity of the widely used FP Tree concept is an interesting algorithm defined as the 'FP Tree growth' technique, given by Han.

The General Social Survey (GSS) – a statistical dataset was taken from MathCS.org website. The General Social Survey (GSS) conducts basic scientific research on the structure and development of American society with a data-collection program designed to monitor social changes within nations. The GSS data sets contain a standard 'core' of demographic and attitudinal questions[1], plus topics of special interest, representing the population of adults, 18 years of age or older[1].

1.1 Data Mining
Data mining is the practice of repeatedly searching huge chunks of data to determine patterns and trends that go beyond simple analysis. Data mining uses[2] sophisticated mathematical algorithms to segment the data and evaluate the probability of future events.

1.2 The need for data mining
Enormous data is being collected[3] and warehoused during:

- purchases at department/grocery[3] stores
- Bank/Credit Card transactions
- Web data, e-commerce[4]

Data is being collected and stored at enormous speeds (GB/hour), like in the cases of remote sensors on a satellite, telescopes scanning the skies, microarrays generating gene expression data, scientific simulations generating terabytes of data.

Data mining may help scientists

- in classifying and segmenting data
- in Hypothesis Formation

Moreover, it helps provide better, customized services for an *edge* (e.g. in Customer Relationship Management) in today's world where competitive pressure is strong.

1.3 Association
Association rules are if-then statements that help uncover relationships between seemingly unrelated data in a relational database. Association rules which are based on the concept of strong rules were introduced for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems. An association rule is theoretically divided into two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found together with the antecedent. Association also uses the criteria of Support and Confidence to identify the most important relationships where Support is an indication of how frequently the items appear in the database and Confidence indicates the number of times the if-then statements have been found to be true.

For example, the rule {cheese, bread}=>{eggs} found in the sales data of a supermarket would indicate that if a customer buys cheese and bread together, she is likely to also buy eggs.

Algorithms used in association rules are:

- Apriori algorithm
- FP Tree growth algorithm
- Eclat algorithm
- GUHA procedure ASSOC

1.4 FP Growth Algorithm

The Association technique gave way to the FP-Growth Algorithm, propounded by Han[5]. It is an efficient method wherein the mining is done by an extended prefix-tree structure on a complete set of frequent patterns by patterns fragment growth. The tree structure stores the compressed information about frequent patterns. In his study, Han proved that due to the Divide and Conquer method and other methods, this algorithm is more efficient than other popular methods for frequent pattern mining e.g. the Apriori Algorithm.[6]

This algorithm begins by compressing the input database, thereby developing an instance of a frequent pattern tree. The compressed database is then divided into a few conditional databases, where every database represents one unique frequent pattern. Finally, mining of every database is carried out discretely. Hence, the search costs are significantly lessened, offering good selectivity.[7]

The reasons of the FP Growth algorithm being more efficient than other algorithms are:

1. Divide and Conquer:
   The mining data is decomposed into sub-datasets according to the frequent patterns identified.
   It leads to more focused search of smaller databases.
2. There is no candidate generation. As a result no candidate test is required.
3. No repeated scans of the whole database.

1.5 FP Trees

- A frequent pattern tree consists of a root[8] labelled as null, a set of item-prefix subtrees as the children of the root, and a frequent item header table.
- Each node in the item-prefix subtree[8] consists of three fields: item-name, count and node-link, where item-name registers which item the node represents, count registers the number of transactions represented by the portion of the path reaching that node and node-link links to the next node in the FP Tree, that carries the same item-name or null if there is none.
- Each entry in the frequent–item- header table consists of two fields: an item-name and a head of the node-link.[9].

2. FP-TREE CONSTRUCTIVE ALGORITHM

Algorithm : FP-Growth

Input: DB Database, depicted by FP-tree built according to Algorithm 1, and a minimum support threshold ?.
Output: The entire group of frequently occurring rules.
Method: call FP-growth(FP-tree, null).

Procedure FP-growth(Tree, a) {
(01) if the Tree comprises a unique prefix path then // Mining single prefix-path FP-tree {
(02) let P be the unique prefix-path element of the Tree;
(03) Assuming Q to be the multipath element with the topmost branching node replaced by a null root;
(04) for each combination (denoted as ß) of the nodes in the path P do
(05) generate pattern ß ∪ a with support = minimum support of nodes in ß;
(06) letfreq pattern set(P) be the set of patterns so generated;
}
(07) else let Q be Tree;
(08) for each item ai in Q do { // Mining multipath FP-tree
(09) generate pattern ß = ai∪ a with support = ai .support;
(10) build ß's pattern-base (which is dependent on conditions) and then ß's conditional FP-tree Tree ß;
(11) if Tree ß ≠ Ø then
(12) call FP-growth(Tree ß , ß);
(13) letfreq pattern set(Q) be the set of patterns so generated;
}
(14) return(freq pattern set(P) ∪freq pattern set(Q) ∪ (freq pattern set(P) × freq pattern set(Q)))
}

3. IMPLEMENTATION

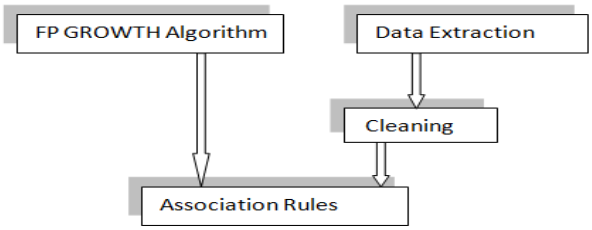

Fig 1: Flowchart of stages during implementation

Figure 1 shows the implementation process. The implementation starts with the user feedback dataset obtained online and comprised of a range of attributes. This dataset is then cleaned by rectifying and resolving the missing and incorrect values. Then the available FP Growth algorithm is applied on the clean dataset which results in formation of association rules required for analysis.

3.1 Dataset

The dataset was obtained online and comprised of a range of attributes: race, age, sex, marital status, number of siblings, number of children etc. It is a user feedback dataset. Different people from many backgrounds and societies were asked questions, and they provided information about themselves.

Now the main parameters that were taken into consideration are discussed. SEX is classified as Male or Female. Marital Status is divided into five groups- Married, Never married, Widowed, Divorced as well as Separated. HIGHEST DEGREE obtained by an individual can be- High School, Graduate,

Bachelor, Less than High School (Less than HS), Junior College. SPEAK LANGUAGE OTHER THAN ENGLISH can be Yes or No. Out of the many attributes, SPEAK LANGUAGE OTHER THAN ENGLISH, SEX, MARTIAL STATUS, HIGHEST DEGREE were the major ones taken into consideration.

3.2 Data Preprocessing

First, the data was cleaned. All the missing values were resolved, and wrong values were rectified. This is essential, as cleaner data would provide for a better analysis.

A numeric value was assigned to each of the input entries. This way, all the categories taken had a unique numeric identity. The excel file was converted to a .txt file and fed as input to the system. Since four broad categories were selected, four columns consisting of unique numeric assignments were used.

Table 1. Parameters of dataset and codes assigned to them

COLUMN	CODE ASSIGNED	ATTRIBUTES
SPEAK LANGUAGE OTHER THAN ENGLISH	1	Yes
	2	No
SEX	101	Male
	102	Female
MARITAL STATUS	201	Married
	202	Never Married
	203	Widowed
	204	Divorced
	205	Separated
HIGHEST DEGREE	300	Less than HS
	301	High School
	302	Junior College
	303	Bachelor
	304	Graduate

4. RESULT AND ANALYSIS

4.1 Finding association rules

Finally, after readying the dataset for input and usage, association rules in it are found. The default support and confidence levels are taken as 20% and 80% respectively. Ultimately, one rule is found, signifying the rules {102,301} -> {2}, which suggests that when the element fields 'Female' and 'High School' are found, they are accompanied by the field 'No', where 102 depicts the value 'Female' of SEX attribute, 301 depicts the value 'High School' of HIGHEST DEGREE attribute, and 2 depicts the 'No' value of SPEAK LANGUAGE OTHER THAN ENGLISH attribute. In a nutshell, when a person is female and her highest degree of qualification is High School, then 81.05% times, she cannot speak any language other than English. Figure 2 shows the association rules that were found for the dataset.

```
FP TREE CREATION....
File name is: tp14.txt
No. of records in input file is: 2023
No. of columns in input file is: 304
Min support is: 404 records (20%)
Confidence is: 80%
FP TREE
(1) 2:1468 (ref to null)
(1.1) 102:808 (ref to null)
(1.1.1) 301:445 (ref to 301:341)
(1.1.1.1) 201:210 (ref to 201:77)
(1.1.1.2) 202:91 (ref to null)
(1.1.2) 201:163 (ref to 201:210)
(1.1.3) 202:75 (ref to 202:47)
(1.2) 301:341 (ref to 301:113)
(1.2.1) 201:167 (ref to 201:185)
(1.2.1.1) 101:167 (ref to 101:185)
(1.2.2) 101:174 (ref to 101:167)
(1.2.2.1) 202:103 (ref to 202:51)
(1.3) 201:185 (ref to null)
(1.3.1) 101:185 (ref to 101:72)
(1.4) 101:134 (ref to 101:77)
(1.4.1) 202:75 (ref to 202:50)
(2) 102:286 (ref to 102:808)
(2.1) 301:104 (ref to 301:445)
(2.1.1) 201:40 (ref to 201:89)
(2.1.1.1) 1:40 (ref to 1:89)
(2.1.2) 1:64 (ref to 1:72)
(2.1.2.1) 202:39 (ref to 202:103)
(2.2) 201:89 (ref to 201:167)
(2.2.1) 1:89 (ref to 1:79)
(2.3) 1:93 (ref to 1:77)
(2.3.1) 202:51 (ref to 202:75)
(3) 301:113 (ref to null)
(3.1) 201:41 (ref to 201:163)
(3.1.1) 101:41 (ref to 101:134)
(3.1.1.1) 1:41 (ref to 1:93)
(3.2) 101:72 (ref to null)
(3.2.1) 1:72 (ref to null)
(3.2.1.1) 202:50 (ref to 202:75)
(4) 201:77 (ref to 201:40)
(4.1) 101:77 (ref to 101:79)
(4.1.1) 1:77 (ref to 1:40)
(5) 101:79 (ref to 101:174)
(5.1) 1:79 (ref to 1:64)
(5.1.1) 202:47 (ref to 202:91)
FP tree storage is: 914 bytes.
Association Rules obtained from FP tree:- a) {102
301} -> {2} 81.05%
```

Fig 2: The association rules in the dataset

4.2 Analysis of association rules

Finally, after readying the dataset for input and usage, association rules in it are found. The default support and confidence levels are taken as 20% and 80% respectively. Ultimately, one rule is found, signifying the rules {102,301} -> {2}, which suggests that when the element fields 'Female' and 'High School' are found, they are accompanied by the field 'No', where 102 depicts the value 'Female' of SEX attribute, 301 depicts the value 'High School' of HIGHEST DEGREE attribute, and 2 depicts the 'No' value of SPEAK LANGUAGE OTHER THAN ENGLISH attribute. In a nutshell, when a person is female and her highest degree of qualification is High School, then 81.05% times, she cannot speak any language other than English.

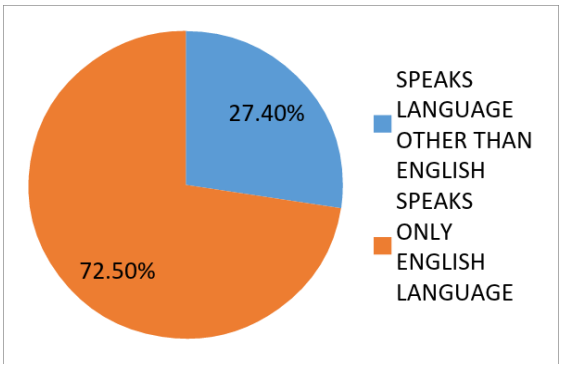

Fig 3: Frequency of occurrences for all records on the basis of 'SPEAK LANGUAGE OTHER THAN ENGLISH'

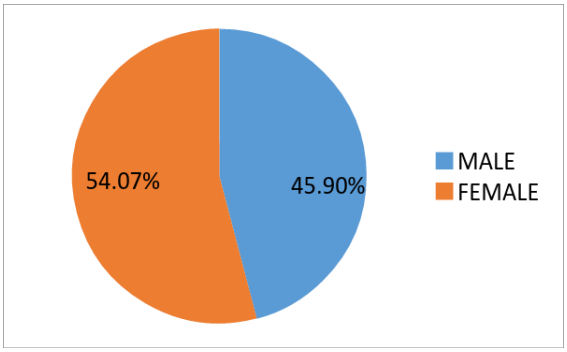

Fig 4: Frequency of occurrences for all records on the basis of 'SEX'

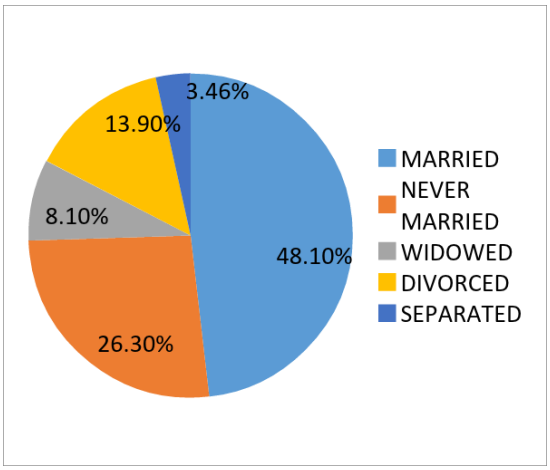

Fig 5: Frequency of occurrences for all records on the basis of 'MARITAL STATUS'

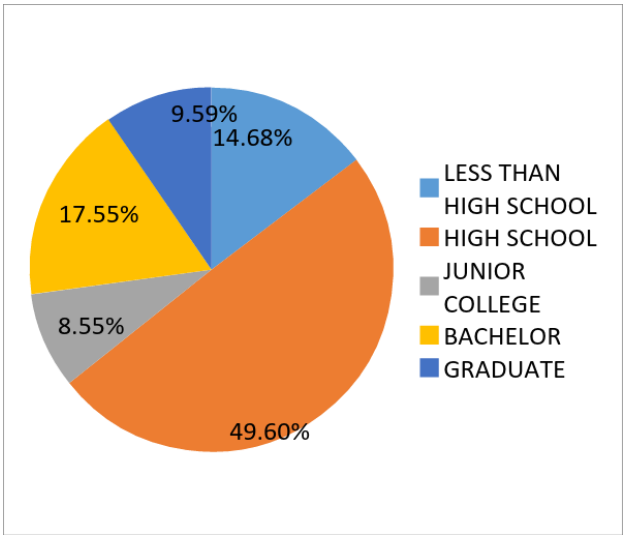

Fig 6: Frequency of occurrences for all records on the basis of 'HIGHEST DEGREE'

From Figure 3, we observed that people speaking only English language are more than those speaking other language. Figure 4 talks about frequency of occurrences on the basis of 'SEX'. Here, we can conclude that female population exceeds the male population and hence the sex ratio (Number of males/Number of females) is 1.17. In Figure 5, the frequency of occurrences is based on 'MARITAL STATUS'. Here it can be inferred that nearly half the population is married and more than a quarter are single. Similarly, from Figure 6 we can observe that the graduates constitute the least percent of population whereas high school passouts are in majority.

5. CONCLUSION

This paper presents the importance of using the FP Tree algorithm in order to obtain association rules between related data, which would help in targeting favourable association rules according to the requirements. This technique can prove to be extremely useful in market researches. One can find otherwise hidden information and relationships from the data, and take further decisions based on the acquired knowledge.

For long, many different algorithms like the Apriori algorithm have been used in the field of analysis of patterns. But it has been found that these algorithms possess some drawbacks such as repeated scans of the whole database, and candidate key generation, which further requires candidate tests. Hence, if the data is too large or complex, the time and complexity are increased.

The FP-growth algorithm uses the 'Divide and Conquer' strategy and does not require candidate key generation tests. Furthermore, it doesn't undergo repeated scans of the data. So, it can be safely concluded that the FP-growth algorithm has a vast future scope in the area of marketing in the organized sector. Hence, we could see greater involvement of the FP Tree concept in competitive global markets in the future.

6. REFERENCES

[1] General Social Survey (Subset) of 2008, 1 Oct 2009 (http://sda.berkeley.edu/archive.htm)

[2] J. Bhatia, Anu Gupta, "Mining of Quantitative Association Rules in Agricultural Data Warehouse: A Road Map", International Journal of Information Science and Intelligent System, 3(1): 187-198, 2014.

[3] D. PUGAZHENDI, "Apriori algorithm on Marine Fisheries Biological Data", International Journal of Computer Science & Engineering Technology, Dec 12, 2013.

[4] Santhosh Kumar, B.; Rukmani, K. V. "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms", International Journal of Advanced Networking & Applications . May/Jun2010, Vol. 1 Issue 6.

[5] Jiawei Han, MichelineKamber, "Data Mining:Concepts and Techniques", June 2011, Elsevier.

[6] Yong Qiu ;Yong-Jie Lan ; Qing-Song Xie, "An improved algorithm of mining from FP-tree", Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on (Volume:3)

[7] Yi Sui; FengJing Shao ; Rencheng Sun ; Jinlong Wang, "A Sequential Pattern Mining Algorithm Based on Improved FP-tree", Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD '08. Ninth ACIS International Conference

[8] M.H Nadimi-Shahraki, Norwati Mustapha, MdNasir B Sulaiman, Ali B Mamat, "Efficient Candidacy Reduction For Frequent Pattern Mining", International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

[9] Changjie Tang, Charles X. Ling, Xiaofang Zhou, Nick Cercone, Xue Li, " Advanced Data Mining and Applications", 4th International Conference, ADMA 2008, Chengdu, China, October 8-10, 2008, Proceedings, Springer 2008.