IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-3 E-ISSN: 2347-2693

Frequent Itemset Mining: A Metadata Based Approach for Knowledge Discovery

Basavaraj A.[1*] Goudannavar [2], Prashant Bhat[3]

[1*,2]Department of Computer Science, Rani Channamma University, Belagavi-591156, Karnataka, India
[2]Department of Business Analytics/Data Science, Chris Institute of Management, Lavasa-412112, Pune, Maharashtra, India

Available online at: www.ijcseonline.org

Received: 17/Feb//2018, Revised: 26/Feb2018, Accepted: 19/Mar/2018, Published: 30/Mar/2018

***Abstract*—** Frequent sets play a crucial role in many Data Mining tasks that try to find interesting patterns from databases, such as correlations, association rules, classification and clustering. The Association Rules is one of the most used functions in data mining. The method is used both database researchers and data mining users. In this article, association rule mining algorithms are discussed and demonstrated. Mining Associate rule algorithm that search for approximate strong association rules from multimedia databases. The Apriori-like sequential pattern mining approach based on candidate generates-and test can also be explored by mapping a sequence multimedia database into vertical data format. This approach is useful to finding frequent itemsets, which probabilistic frequent itemsets based on possible datasets.

***Keywords*—** Web Multimedia Mining, Association rule, Frequent itemsets, Knowledge discovery

I. INTRODUCTION

Association rule mining is a data mining task that discovers relationships among items in a transactional database. Association rules have been extensively studied in the literature for their usefulness in many application domains such as diagnosis decisions support, telecommunication, multimedia database, etc. The efficient discovery of such rules has been a major focus in the data mining research community. From the original *apriori* algorithm [1] there have been a remarkable number of variants and improvements of association rule mining algorithms [2].

Association rules are required to satisfy both a minimum support and a minimum confidence constraint at the same time. At medium to low support values, often a great number of frequent itemsets are found in a database. However, since the definition of support enforces that all subsets of a frequent itemset have to be also frequent, it is sufficient to only mine all maximal frequent itemsets, defined as frequent itemsets which are not proper subsets of any other frequent itemset. Another approach to reduce the number of mined itemsets is to only mine frequent closed itemsets. An itemset is closed if no proper superset of the itemset is contained in each transaction in which the itemset is contained [3]. In this proposed work, using KNIME data mining tool is used [4], the web multimedia-video metadata are extracted and frequent itemsets and closed frequent itemsets on available metadata of web multimedia-videos using Apriori algorithm and FP-Growth algorithms. The frequent pattern mining results are analyzed.

The respite of the paper is ordered as follows: The part 2 represents related works on the association rule on web multimedia datasets, part 3 represents proposed methodology, part 4 represents evaluation analysis of association rule, and finally part 5 represents conclusion.

II. PRIOR WORKS

The authors [5] introduce an improvement to the mining Apriori association rule generates approximate association rules. The apriori algorithm takes into consideration missing values and noisy data. The experiment indicates that apriori effectively generates rules that approximate positive correlations in the input database.

In the proposed work, multiple database passes to generate the frequent item sets [6]. In this article a new Algorithm is proposed which can mine frequent patterns with a single scan of database. The time taken by this algorithm is compared with the other algorithms also and it proves that time taken by this algorithm is less than the partition algorithm and the Apriori but more than the time taken by FP-Tree Algorithms but FP tree has its own disadvantages such as if every transaction in database is different then we will have as many as 2n leaves in FP tree (where n is the no. of items) [7] [8]. In this article, proposed four parallel versions of a novel sequential mining algorithm for discovery of frequent itemsets are proposed. The parallel algorithms were compared analytically and experimentally, with respect to some factors, such as communication rate, response time, computation/communication ratio and load balancing. [9].

A novel adaptive classification method using random forests, which is a machine learning algorithm with proven good performance on many traditional classification problems [10]. Video classification is the first step toward

multimedia content understanding. When video is classified into conceptual categories, it is usually desirable to combine evidence from multiple modalities [11].

III. PROPOSED METHODOLOGY

In this part we describe an effectual methodology to extract the metadata from web multimedia files and categorize them based on the frequent itemsets of metadata by applying data mining techniques. In this proposed method, out of the total metadata dataset, 60% are used for training and remaining 40% are used for testing the classification model built using Decision Tree and SVM classification methods. The consequences are analyzed and the effectiveness of the proposed method has been demonstrated.

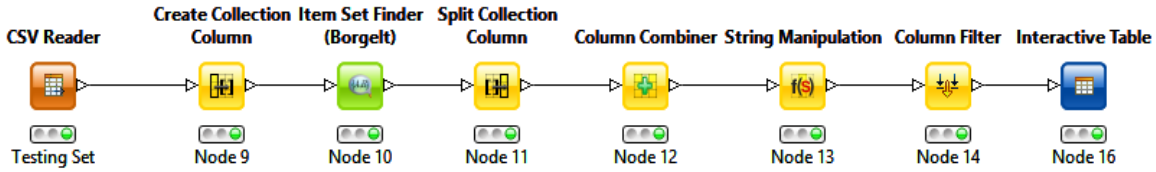


Figure 1: System model of the proposed methodology

The proposed system model is represented in Figure 1. It consists of the following components:

i) Web multimedia-video metadata extraction and Pre-processing
ii) Frequent pattern mining and association rules model
iii) Result analysis

The functionality of each element of the proposed system model is discussed in the following subsections.

3.1 Web Multimedia-Video Metadata Extraction and pre-processing

There are many tools available to decompose the web multimedia data meaning that, it is possible to retrieve individual components object from multimedia data. The various techniques to extract web multimedia data and pre-processing are discussed in our previous work [12].

3.2 Frequent pattern mining and association rules model

The proposed work, we adopt a model to mine frequent itemsets of web multimedia metadata. The mining precision and efficiency will depend on the constructed frequent itemset mining model. This section represents detailed procedure to construct frequent itemset model.

3.2.1 Association rules

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true. Given a set of transactions, each described by an unordered set of items, an association rule A ->B may be discovered in the data, where A and B are conjunctions of items. The intuitive meaning of such a rule is that transactions in the database which contain the items in A, tend to also contain the items in B. An example of such a rule might be that observed metadata multimedia datasets to find the frequent itemset mining also association rule generating. In this case, A = {Stream size Mib, Image Height, Audio duration} and B = {Video duration}. Two numbers are associated with each rule that indicate the support and confidence of the rule using multimedia data.

The problem of discovering association rules can be divided into two steps:

1. Find all *itemsets* (sets of items appearing together in a transaction) whose support is greater than the specified threshold.

2. Generate association rules from the frequent itemsets.

Confidence of a candidate rule is calculated as support of the multimedia metadata. All rules that meet the confidence threshold are reported as discoveries of the algorithm.

1. Scan the (entire) transaction database to get the support S of each 1-itemset, compare S with min_sup , and get a set of frequent 1-itemsets, L1

2. Use Lk-1 join Lk-1 to generate a set of candidate kitemsets.

3. Scan the transaction database to get the support S of each candidate k-itemset in the final set, compare S with m in_sup , and get a set of frequent kitemsets, L_k

4. Is the candidate set empty, if not goto 2

5. For each frequent itemset l, generate all nonempty subsets of l

6. For every nonempty subset s of l, output the rule S= (I-s) if its confidence C > min >conf

Algorithm 1: Apriori Algorithm

In this algorithm first calculates single item frequencies to determine the frequent 1-itemsets (Minimum support of Multimedia data). Once candidates are generated, itemsets are removed from consideration if any (k-1) subset of the candidate is not in Lk-1.

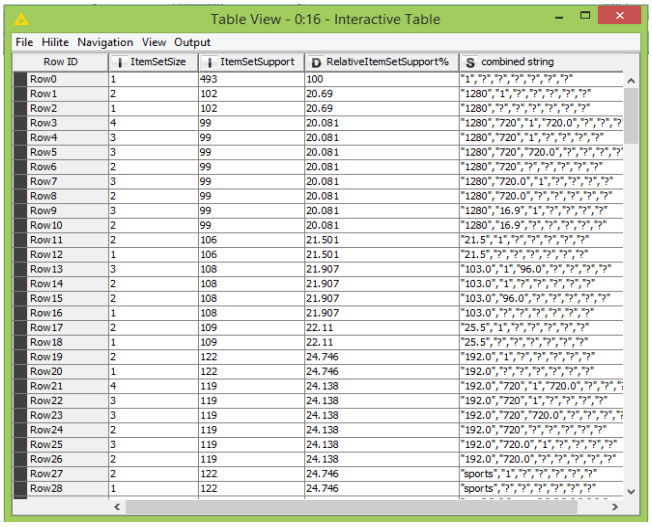
4. Experimental Results and Discussions

4.1 Association rule and frequent itemset analysis

The association rule algorithm used in this experiment is restricted to finding rule and frequent itemset between pairs of items only. In this experiment propose a novel approach for frequent item sets mining using multimedia metadata. Frequent itemsets is an item set that satisfies minimum support. The multimedia metadata information fields like video duration, video bit rate kbps, maximum bit rate kbps, width pixels, height pixels, display aspect ratio, bits/(pixel*frame), stream size mib, audio duration, audio bit rate kbps, maximum bit rate kbps, stream size mib, image resolution, image height, image width, text page, word count, character count, line count, paragraph count, size in kbps, class. The purpose of this experiment is to find out the result for generating frequent itemsets using Apriori algorithm, and added on the item that it took to find the association rules within the frequent itemsets.

The first step of Apriori is to count the frequencies in data set, called the supports, of each metadata item separately.

Table-1: Generation of frequent 1-itmeset



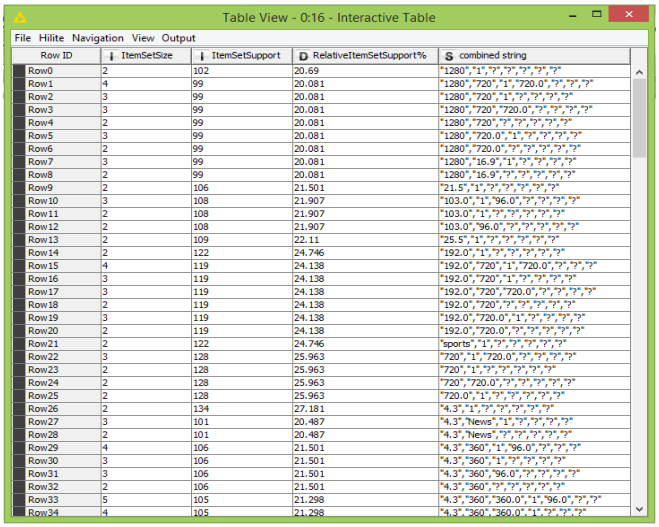
The support **supp(*A*)** of an itemset ***A*** is defined as the proportion of transactions in the dataset which contain the itemset i.e.,

supp(A) = no. of metadata which contain the itemset A / total no. of metadata

In the multimedia metadata, each item is a member of the set of candidate 1-itemsets, C_1. The algorithm simply scans all of the transactions in order to count the number of occurrences of each item. The minimum support count required is 99, that is, *min sup* = 99. (Here, we are referring to *absolute* support because we are using a support count. The corresponding relative support is 99/493= 20.08%). To be even more explicit the experiment multimedia metadata is 99 is the number of transactions from the medtadata which contain the itemset while 493 represents the total number of transactions. We can define a minimum support level to qualify as "frequent," which depends on the context. For this case, let minimum support 99 and relative support are 20.08. Therefore, all are frequent in 1-itemset.

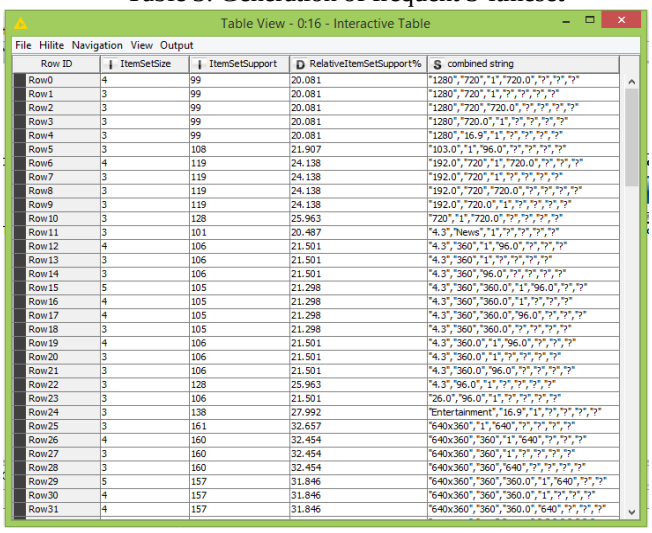
The next step is to generate a list of all 2-pairs of the frequent items. Had any of the above items not been frequent, they would not have been included as a possible member of possible 2-item pairs. In this way, Apriori *prunes* the tree of all possible data sets. Further we again select only these items (now 2- pairs are items). The web multimedia metadata 205 transaction database in 1-itemset, we found 186 which are frequent in 2-itemset, the remaining 19 items are rejected which are lower than the support value are shown in table-2.

Table 2: Generation of frequent 2-itmeset



Generation of all 3-triples of list the frequent items from the dataset. The above items not been frequent, they would not have been included as a possible member of possible 3-item pairs. Further we again select only these items (now 3-pairs are items) the 186 transaction database in 2-itemset, we found 136 transaction dataset which are frequent in 3-itemset, the remaining 50 items are rejected which are lower than the support value are shown in table-3.

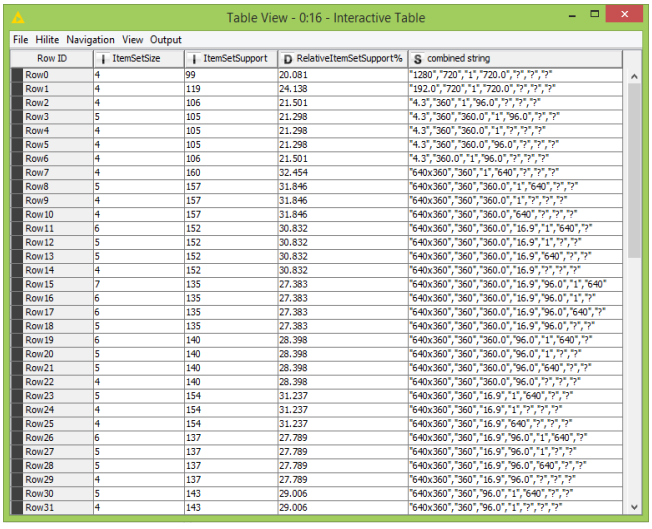
Table 3: Generation of frequent 3-itmeset



Row ID	ItemSetSize	ItemSetSupport	RelativeItemSetSupport%	combined string
Row0	4	99	20.081	"1280","720","1","720.0","?","?","?"
Row1	3	99	20.081	"1280","720","1","?","?","?","?"
Row2	3	99	20.081	"1280","720","720.0","?","?","?","?"
Row3	3	99	20.081	"1280","720.0","1","?","?","?","?"
Row4	3	99	20.081	"1280","16.9","1","?","?","?","?"
Row5	3	108	21.907	"103.0","1","96.0","?","?","?","?"
Row6	4	119	24.138	"192.0","720","1","720.0","?","?","?"
Row7	3	119	24.138	"192.0","720","1","?","?","?","?"
Row8	3	119	24.138	"192.0","720","720.0","?","?","?","?"
Row9	3	119	24.138	"192.0","720.0","1","?","?","?","?"
Row10	3	128	25.963	"720","1","720.0","?","?","?","?"
Row11	3	101	20.487	"4.3","News","1","?","?","?","?"
Row12	4	106	21.501	"4.3","360","1","96.0","?","?","?"
Row13	3	106	21.501	"4.3","360","1","?","?","?","?"
Row14	3	106	21.501	"4.3","360","96.0","?","?","?","?"
Row15	5	105	21.298	"4.3","360","360.0","1","96.0","?","?"
Row16	4	105	21.298	"4.3","360","360.0","1","?","?","?"
Row17	4	105	21.298	"4.3","360","360.0","96.0","?","?","?"
Row18	3	105	21.298	"4.3","360","360.0","?","?","?","?"
Row19	4	106	21.501	"4.3","360.0","1","96.0","?","?","?"
Row20	3	106	21.501	"4.3","360.0","1","?","?","?","?"
Row21	3	106	21.501	"4.3","360.0","96.0","?","?","?","?"
Row22	3	128	25.963	"4.3","96.0","1","?","?","?","?"
Row23	3	106	21.501	"26.0","96.0","1","?","?","?","?"
Row24	3	138	27.992	"Entertainment","16.9","1","?","?","?","?"
Row25	3	161	32.657	"640x360","1","640","?","?","?","?"
Row26	4	160	32.454	"640x360","360","1","640","?","?","?"
Row27	3	160	32.454	"640x360","360","1","?","?","?","?"
Row28	3	160	32.454	"640x360","360","640","?","?","?","?"
Row29	5	157	31.846	"640x360","360","360.0","1","640","?","?"
Row30	4	157	31.846	"640x360","360","360.0","1","?","?","?"
Row31	4	157	31.846	"640x360","360","360.0","640","?","?","?"

Here we generate list of all 4-triples of the frequent items from the dataset. The above items not been frequent, they would not have been included as a possible member of possible 4-item pairs. Next step we again select only these items (now 4- pairs are items) the 136 transaction database in 3-itemset, we found 76 transaction dataset which are frequent in 4-itemset, the remaining 60 items are rejected which are lower than the support value are shown in table-4.

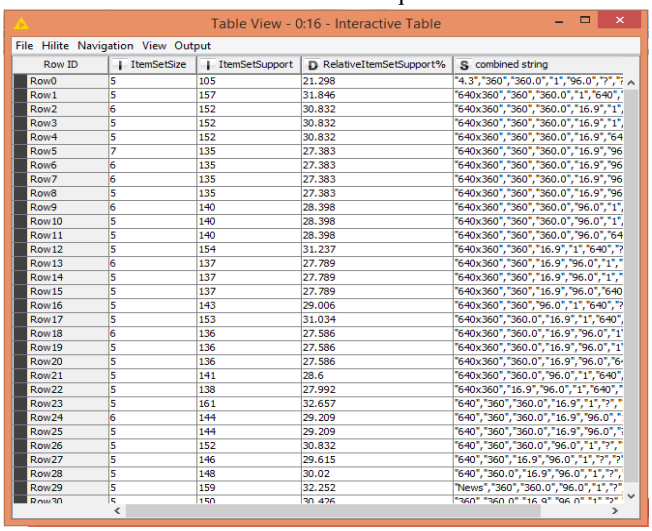
Table-4: Generation of frequent 4-itmeset



Row ID	ItemSetSize	ItemSetSupport	RelativeItemSetSupport%	combined string
Row0	4	99	20.081	"1280","720","1","720.0","?","?","?"
Row1	4	119	24.138	"192.0","720","1","720.0","?","?","?"
Row2	4	106	21.501	"4.3","360","1","96.0","?","?","?"
Row3	5	105	21.298	"4.3","360","360.0","1","96.0","?","?"
Row4	4	105	21.298	"4.3","360","360.0","1","?","?","?"
Row5	4	105	21.298	"4.3","360","360.0","96.0","?","?","?"
Row6	4	106	21.501	"4.3","360.0","1","96.0","?","?","?"
Row7	4	160	32.454	"640x360","360","1","640","?","?","?"
Row8	5	157	31.846	"640x360","360","360.0","1","640","?","?"
Row9	4	157	31.846	"640x360","360","360.0","1","?","?","?"
Row10	4	157	31.846	"640x360","360","360.0","640","?","?","?"
Row11	6	152	30.832	"640x360","360","360.0","16.9","1","640","?"
Row12	5	152	30.832	"640x360","360","360.0","16.9","1","?","?"
Row13	5	152	30.832	"640x360","360","360.0","16.9","640","?","?"
Row14	4	152	30.832	"640x360","360","360.0","16.9","?","?","?"
Row15	7	135	27.383	"640x360","360","360.0","16.9","96.0","1","640"
Row16	6	135	27.383	"640x360","360","360.0","16.9","96.0","1","?"
Row17	6	135	27.383	"640x360","360","360.0","16.9","96.0","640","?"
Row18	5	135	27.383	"640x360","360","360.0","16.9","96.0","?","?"
Row19	6	140	28.398	"640x360","360","360.0","96.0","1","640","?"
Row20	5	140	28.398	"640x360","360","360.0","96.0","1","?","?"
Row21	5	140	28.398	"640x360","360","360.0","96.0","640","?","?"
Row22	4	140	28.398	"640x360","360","360.0","96.0","?","?","?"
Row23	5	154	31.237	"640x360","360","16.9","1","640","?","?"
Row24	4	154	31.237	"640x360","360","16.9","1","?","?","?"
Row25	4	154	31.237	"640x360","360","16.9","640","?","?","?"
Row26	6	137	27.789	"640x360","360","16.9","96.0","1","640","?"
Row27	5	137	27.789	"640x360","360","16.9","96.0","1","?","?"
Row28	5	137	27.789	"640x360","360","16.9","96.0","640","?","?"
Row29	4	137	27.789	"640x360","360","16.9","96.0","?","?","?"
Row30	5	143	29.006	"640x360","360","96.0","1","640","?","?"
Row31	4	143	29.006	"640x360","360","96.0","1","?","?","?"

Similarly generate list of all 5-triples of the frequent items from the dataset. The above items not been frequent, they would not have been included as a possible member of possible 5-item pairs. Further we again select only these items (now 5- pairs are items) the 76 transaction database in 4-itemset we found 31 transaction dataset which are frequent in 5-itemset, the remaining 45 items are rejected which are lower than the support value are shown in table-5.

Table-5: Generation of frequent 5-itmeset



Row ID	ItemSetSize	ItemSetSupport	RelativeItemSetSupport%	combined string
Row0	5	105	21.298	"4.3","360","360.0","1","96.0","?","?
Row1	5	157	31.846	"640x360","360","360.0","1","640",
Row2	6	152	30.832	"640x360","360","360.0","16.9","1",
Row3	5	152	30.832	"640x360","360","360.0","16.9","1",
Row4	5	152	30.832	"640x360","360","360.0","16.9","64
Row5	7	135	27.383	"640x360","360","360.0","16.9","96
Row6	6	135	27.383	"640x360","360","360.0","16.9","96
Row7	6	135	27.383	"640x360","360","360.0","16.9","96
Row8	5	135	27.383	"640x360","360","360.0","16.9","96
Row9	6	140	28.398	"640x360","360","360.0","96.0","1",
Row10	5	140	28.398	"640x360","360","360.0","96.0","1",
Row11	5	140	28.398	"640x360","360","360.0","96.0","64
Row12	5	154	31.237	"640x360","360","16.9","1","640","?
Row13	6	137	27.789	"640x360","360","16.9","96.0","1",
Row14	5	137	27.789	"640x360","360","16.9","96.0","1",
Row15	5	137	27.789	"640x360","360","16.9","96.0","640
Row16	5	143	29.006	"640x360","360","96.0","1","640","?
Row17	5	153	31.034	"640x360","360.0","16.9","1","640",
Row18	6	136	27.586	"640x360","360.0","16.9","96.0","1
Row19	5	136	27.586	"640x360","360.0","16.9","96.0","1
Row20	5	136	27.586	"640x360","360.0","16.9","96.0","6
Row21	5	141	28.6	"640x360","360.0","96.0","1","640",
Row22	5	138	27.992	"640x360","16.9","96.0","1","640",
Row23	5	161	32.657	"640","360","360.0","16.9","1","?",
Row24	6	144	29.209	"640","360","360.0","16.9","96.0",
Row25	5	144	29.209	"640","360","360.0","16.9","96.0",
Row26	5	152	30.832	"640","360","360.0","96.0","1","?",
Row27	5	146	29.615	"640","360","16.9","96.0","1","?",
Row28	5	148	30.02	"640","360.0","16.9","96.0","1","?",
Row29	5	159	32.252	"News","360","360.0","96.0","1","?
Row30	5	150	30.426	"360","360.0","16.9","96.0","1","?

The algorithm generates a candidate set of 6-itemsets. Although the join results in {width pixels, height pixels, display aspect ratio, audio bit rate kbps, image resolution, text page}, this itemset is pruned because its subset {width pixels, height pixels, display aspect ratio, audio bit rate kbps, image resolution} is not frequent. Thus, $C5 = \Phi$, and the algorithm terminates, having found all of the frequent itemsets are shown in table-5.

4.2 Generating association rules from frequent itemsets

Once the frequent itemsets from transactions in a database *D* have been found, it is straightforward to generate strong association rules from them (where *strong* association rules satisfy both minimum support and minimum confidence). This can be done using Equation for confidence, which we show again here for completeness:

Confidence c, is the conditional probability that specify the c% of the transactions of database considered must specify $X \rightarrow Y$

Confidence ($X \rightarrow Y$) =$P(Y/X)$ = 0.20/0.99 = 20%

This means that for 20% of the transactions containing multimedia metadata rule is correct as shown below. Minimum Support and Minimum Confidence are needed to eliminate the unimportant association rules.

Confidence = 99/105 = 94%

Confidence = 99/157 = 63%

Confidence = 99/152 = 65%

Confidence = 99/152 = 65%

Confidence = 99/135 = 73%

Confidence = 99/135 = 73%

Confidence = 99/140 = 70%

Confidence = 99/154 = 64%

Confidence = 99/143 = 69%

If the minimum confidence threshold is, say, 20%, All of these association rules are considered strong because they meet the minimum confidence and support which was set at 20%.

5. Conclusion

In this paper, Apriori is one of the most well data mining approaches is to find frequent itemsets from a transaction dataset of multimedia metadata and to generate strong association rules. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Further we use apriori algorithms for finding informative patterns from the complex multimedia data sources.

REFERENCES

[1] Agrawal, R., Imielinski, T., Swami, "A.: Mining association rules between sets of items in large databases",

[2] Acm SIGKDD Explorations Newsletter, Volume 22, Issue 2, June 1993, ISBN: 0-89791-592-5.

[3] Bart Goethals, Mohammed J. Zaki, "Advances in Frequent Itemset Mining Implementations Report on FIMI'03", ACM SIGKDD Explorations Newsletter, Volume 6, Issue 1, June 2004, ISSN: 1931-0145.

[4] Michael Hahsler, Bettina Grun and Kurt Hornik, "arules – A Computational Environment for Mining Association Rules and Frequent Item Sets", Journal of Statistical Software, Volume 14, Issue 15,October 2005,.

[5] ChenGang "MediaInfo extractor – A Tool for Media Data Mining", 2011. http://mediaarea.net/en/MediaInfo.

[6] Jyothsna R. Nayak and Diane J. Cook, "Approximate Association Rule Mining", FLAIRS-01 Proceedings. 2001, AAAI (www.aaai.org).

[7] Syed Khairuzzaman Tanbeer , Chowdhury , Farhan Ahmed and Byeong-Soo Jeong , "Parallel and Distributed Algorithms for FP mining in large Databases", IETE Technical Review Vol 26, Issue 1 , pp 55-65, Jan 2009.

[8] Pradeep Chouksey* Juhi Singh, R.S. Thakur and R.C. Jain, "Frequent Pattern Mining using Candidate Generation approach with Single Scan of Database", Symposium on Progress in Information & Communication Technology 2009

[9] Ms. Manali Rajeev Raut, Ms. Hemlata Dakhore, "Association Rule Mining in Horizontally Distributed Databases", Manali Rajeev Raut et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7540-7544

[10] Zhi Liu,Tianhong Sunand Guoming Sang," An Algorithm of Association Rules Mining in Large Databases Based on Sampling ", International Journal of Database Theory and Application Vol.6, No.6 , 2013.

[11] Renáta Iváncsy, István Vajk', "Frequent Pattern Mining in Web Log Data", Acta Polytechnica Hungarica Vol. 3, No. 1, 2006.

[12] S. M. Fakhrahmad1 And Gh. Dastghaibyfard2, "An Efficient Frequent Pattern Mining Method and its Parallelization in Transactional Databases", Journal Of Information Science And Engineering 27, 511-525 (2011).

[13] Siddu P. Algur1, Basavaraj A. Goudannavar2*,"Web Multimedia Mining: Metadata Based Classification and Analysis of Web Multimedia", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE) November-2015, pp. 324-330 ISSN: 2277-128X

Authors

Dr. Basavaraj A. Goudannavar is working as Lecturer, Dept. of Computer Science, Rani Channamma University, Belagavi, Karnataka, India. He received BCA and MCA degrees from Karnatak University, Dharwad, Karnataka, India, in 2005 and 2008 respectively. His research interest includes Data Mining, Web Mining, Web multimedia mining, and Knowledge discovery techniques. He published 17 research papers in International Journals and International conferences. He has attended and participated in International and National Conferences and Workshops in his research field.

Dr. Prashant Bhat is working as Assistant Professor, Department of Business Analytics/Data Science, Chris Institute of Management, Lavasa- 412112, Pune, Maharashtra, India. He received B.Sc and M.Sc (Computer Science) degrees from Karnatak University, Dharwad, Karnataka, India, in 2010 and 2012 respectively. His research interest includes Data Mining, Web Mining, web multimedia mining and Information Retrieval from the web and Knowledge discovery techniques, and published 22 research papers in International Journals. Also he has attended and participated in International and National Conferences and Workshops in his research field.