J. Basic. Appl. Sci. Res., 2(6)5908-5914, 2012
© 2012, TextRoad Publication
ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Association Rules Mining for Urdu Language Using Transaction Hash Tables based Apriori (THT-Apriori)

Nazish Asad, M. Younus Javed, Usman Qamar, Memoona Javeria Anwar

National University of Sciences and Technology (NUST), Islamabad, Pakistan.

ABSTRACT

This paper explains that how Association rules can play an important rule to automate Urdu language i.e. to create thesaurus, to mine Urdu text on web, to provide adaptive tools that can use in printing and publishing areas. To extract strong associations from Urdu text Apriori algorithm is used, but it is not worked well for Urdu text so a new Association Rules Mining (ARM) algorithm named as Transaction Hash Table Apriori (THT-Apriori) is proposed. Both algorithms are tested on different Urdu corpuses and results has shown that THT-Apriori is better than Apriori in both aspects i.e. time and number of association rules.
KEY WORDS: Association Rules Mining; Urdu Language; Urdu Mining Model.

INTRODUCTION

A lot of research in natural language processing shown that pattern recognition in natural language was always remained a hot cake for researchers. ARM mined association rules from transactional databases. Although the technique was introduced for market basket analysis i.e. to find the items purchased together but later it used successfully to extract patterns from natural languages. Extensive research in this area revealed that data mining techniques especially ARM was very effective to mine interesting patterns known as Association rules.

Basic concepts of association rules are:

- Let I = {i1,i2,……..., im } be a set of items.
- Let D be a set of transactions, where each transaction T contains a set of items.
- An association rule is an implication of the form X=>Y, where X $\subset$ I and Y $\subset$ I, and X $\cap$ Y=Φ.
- The association rule X=>Y holds in the database D with confidence c if c% of transactions in D that contain X also contain Y.
- The association rule X=>Y has support s if s% of transactions in D that contain X $\cup$ Y.

Association rules mining techniques extract all those rules from the data which satisfy user defined thresholds for support s% and confidence c%.

Association rules mining is a two step process: 1. is to find out the frequent itemsets which is also known as candidate items and 2. is to filter out important association rules from the candidate itemsets [1].

Identification of frequent itemsets is a resource and time consuming task and most of the research focuses that how to prune items to generate minimum valid frequent itemsets and maximum association rules.

Text mining can mine association rules between letters, words, sentences and even paragraphs, they can be used for building a statistical thesaurus, extracting phrases form text and enhancing search results.

The important considerations in text mining are:

- In text databases, distribution of words varies from the conventional transactional databases.

***Corresponding Author:** Nazish Asad, National University of Sciences and Technology (NUST), Islamabad, Pakistan
Email: nasad887@gmail.com

- Numbers of unique words are significantly larger than the number of unique items in a transactional database.
- Text data other than English language is based upon Unicode instead of ASCII code that increases the complexity of implementation.

Rest of the paper is organized in the following fashion: section 2 contains related work, section 3 is based upon association rules mining for Urdu language, section 4 shows results and section 5 is conclusions and future work.

Related Work

In [2], John and Soon have given the idea of Parallel Multipass Inverted Hashing and Pruning for text databases. Proposed algorithm used hash tables to avoid several passes on the database during the mining process. This algorithm adopted the pruning strategy to cut off the infrequent itemsets on the occurrence of items in transactions that were stored in hash tables. It divided the database among various partitions to improve the efficiency of algorithm as compare to Apriori and Count Distribution algorithms. Greater efficiency was achieved by using this technique.

In [3], Zhou targeted the engineering documents for association rules. The documents mining procedures distributed in two sub-processes: one was document structure generation and other was document content generation. Apriori algorithm was used for mining interesting patterns in engineering documents. This algorithm filtered out structure-structure association rules, structure-item association rules and item-item association rules.

In [4], Chao Tang and Chen Liu utilized Apriori algorithm to find out grammatical rules from Chinese text. A new model was proposed for grammatical rules mining which had three major steps: pre-processing, association rules mining and verification of association rules to get real rules. Four different corpuses had been chosen for testing purpose and results have indicated the interesting fact about the length of sentences and effectiveness of Apriori algorithm i.e. For small sentences the algorithm worked well as compare to large sentences because the large sentences have contained combined smaller rules.

In [5], Wu Gongxing devised a new distributed algorithm to extract association rules for the XML data. This algorithm created the DOM tree at the beginning and then had used this tree to extract association rules. The distributed algorithm had worked on multiple web sites. Each website had executed the FreqTree algorithm to compute local support count and sent it to global website. The global site then determined the global frequent items (FI) on the basis of sum of support counts which were gathered from all local sites. At the end, a verification process had applied to filter out the valid XML rules from the global frequent items.

In [6], Al-Zoghby, A., Eldin, A.S., Ismail, N.A. and Hamza, T. introduced a new system based upon Apriori and CHARM algorithm to determine soft-matching association rules for Arabic language. Frequent Closed Itemsets were tested along with Frequent Itemsets. Proposed system has converted the Arabic corpora in transactional databases, then performed cleaning and morphological analysis on the database and finally used Apriori and CHARM algorithms to find out association rules mining. Results has shown that Frequent Closed Itemsets worked well as compare to Frequent Itemsets because Frequent Closed Itemsets reduced redundancy up to a significant level which was present in Frequent Itemsets.

In [7], Yong-le Sun and Ke-liang Jia utilized the Association rules mining based upon Apriori algorithm to solve the Word Sense Disambiguation problem. Apriori successfully extracted association rules between the sense of the ambiguous words and contexts and generated very precise association rules.

In [8], Doug Won Choi and Young Jun Hyun improved Apriori algorithm to discover the candidate itemsets and as well as the frequent itemsets. This algorithm cut off candidate itemsets on the basis of two support values ‘minimum support’ and ‘minimum relative support’ in order to find the transitive relations. This method gave a second chance to items so that more items could be generated in second attempt.

Existing Approach

As shown in figure 1, Urdu mining model consists three major steps: Pre-processing—remove punctuation marks and numeric text from the Urdu text files, Creating transactional database from the cleaned files and finally applying Apriori algorithm to get association rules from Urdu text.

Pre-processing is a very important step because numeric data and punctuation marks increase the number of 1-itemsets that results in more higher order invalid frequent itemsets and mining of these frequent itemsets is a wastage of precious resources.

Conversion of text files to transactional database is a fundamental requirement of Apriori algorithm.

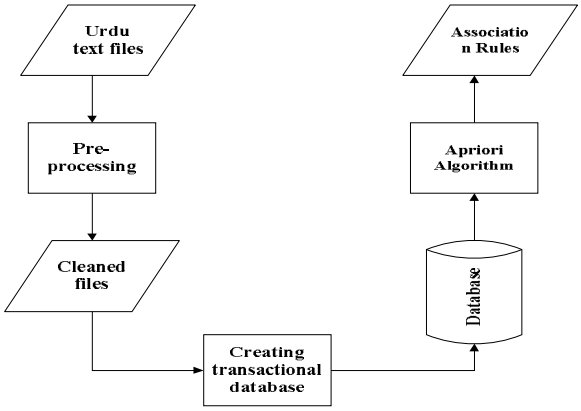

Figure 1 Urdu Text Mining Model [9]

Proposed Approach

Above mining model is used to mine Urdu language but a new algorithm is proposed to extract association rules along with Apriori algorithm. This is because the Aprori algorithm used natural join to find frequent items i.e. a time consuming task. To overcome this problem a new algorithm is proposed. THT-Apriori algorithm is a combination of Multipass with Inverted Hashing and Pruning (MIHP) and Apriori algorithms. It uses hashtables to store the frequent items and their frequencies as key value pairs as MIHP does and utilizes minimum support to filter out the strong association rules from the frequent items. Figure 2 describes the pseudo code of the algorithm.

Algorithm: THT-Apriori

Input: Transactional Database

Output: Association Rules

```
1.  Comment: Read the transactions, count the occurrences of each item formulate all possible frequent items from
    1 to n length.
2.  foreach transaction t ∈ Database do begin
3.      form all possible frequent item                    from 1 to n    length.
4.      hash frequency of frequent items in respective hashtables.
5.  end
6.  Comment: Generate association rules which satisfy minimum support.
7.  foreach FI in hashtable do begin
8.      if (Support of FI >= minimum support)
9.          Add to strong rules
10. end
```

Figure 2 THT-Apriori Algorithm

This algorithm limits the number of frequent itemsets because it only generates those items which exist into the words.

EXPERIMENTS AND RESULTS

Initially, Urdu text files have contained 126, 396 words, after preprocessing the number of words is reduced to 97,922 words. Each word has maximum length of 10 characters. A transactional database is created on the basis of these text files by assigning each word a unique transaction ID. THT-Apriori and Apriori algorithm are tested on this pre-processed transactional database. Testing is done by using different number of words and minimum supports.

Figures 3-6 show that THT-Apriori algorithm performs better than Apriori algorithm at every threshold. On average the THT-Apriori algorithm's efficiency is 43.60% better than Apriori algorithm.

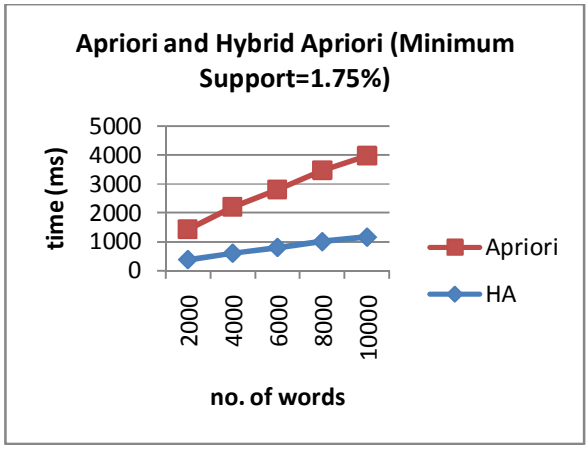

Figure 3 Apriori and THT-Apriori (Minimum Support=1.75%)

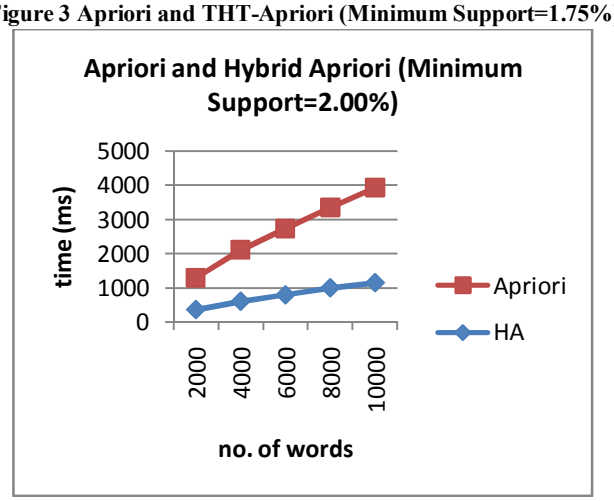

Figure 4 Apriori and THT-Apriori (Minimum Support=2.00%)

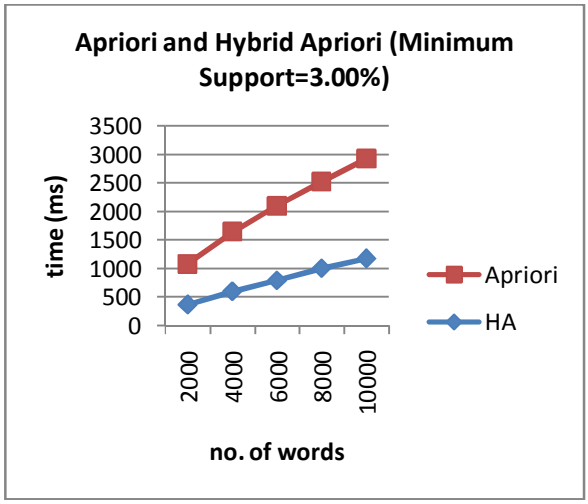

Figure 5 Apriori and THT-Apriori (Minimum Support=3.00%)

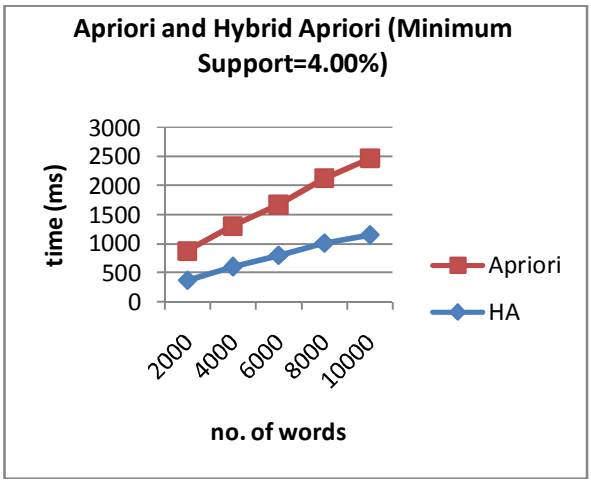

Figure 2 Apriori and THT-Apriori (Minimum Support=4.00%)

Figures 7-10 support the THT-Apriori algorithm results and reveal that number of association rules produced by THT-Apriori is greater than number of association rules generated by Apriori algorithm at every threshold.

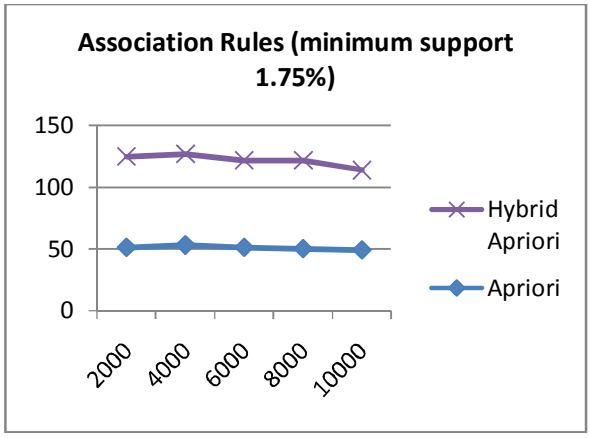

Figure 3 Association Rules (minimum support 1.75%)

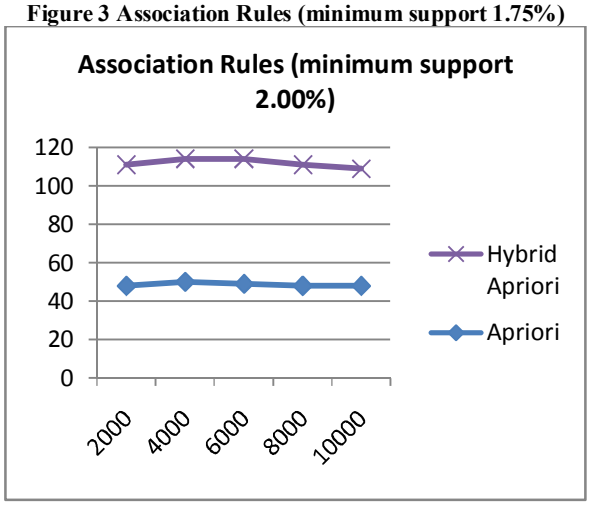

Figure 4 Association Rules (minimum support 2.00%)

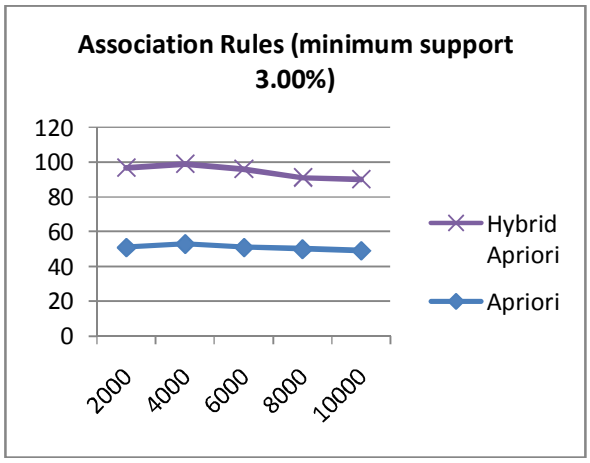

Figure 5 Association Rules (minimum support 3.00%)

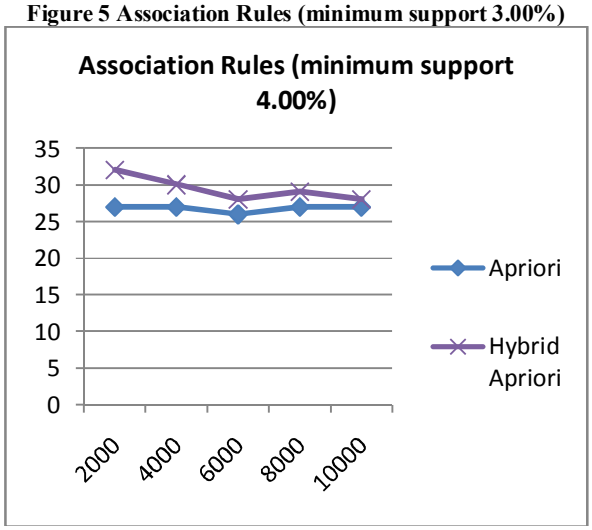

Figure 6 Association Rules (minimum support 4.00%)

THT-Apriori algorithm beats Apriori in both terms that is time and number of association rules.

Conclusions and Future Work

Data mining techniques can effectively apply for extracting association rules from Urdu text. In this context Apriori algorithm is not sufficient to find out interesting patterns. Therefore it is necessary to propose more algorithms to mine Urdu language. THT-Apriori algorithms produce better results for Urdu language as compare to Apriori algorithm in terms of efficiency and accuracy.

Corpus plays a vital rule for association rules mining as the association rules vary from corpus to corpus and are affected by the number of words, sentences, paragraphs and even text files. To extract more accurate and logical rules, it is necessary that corpus is significantly large and contains logically related data. Another important consideration for Urdu language is Unicode processing i.e. file pre-processing, conversion from text to transactional database and implementation are needed to be set according to Unicode coding scheme that varies among programming languages and database management systems.

Urdu language requires more research work on association rules not only among letters but on a broader spectrum i.e. among words, sentences and grammar. Urdu text is needed to be digitized and more Urdu databases

are required for this purpose. So Urdu scholars and users will have more automated tools such as grammar rules extractors, thesaurus and efficient web search in future.

REFERENCES

[1] Data Mining Concepts and Techniques by Jiawei Han and Micheline Kamber.

[2] Holt, J.D., and Chung, S.M. Parallel Mining of Association Rules from Text Databases on a Cluster of Workstations. Proceedings of the 2004 18th international Parallel and Distributes Processing Symposium, (Digital Object Identifier: 10.1109/IPDPS.2004.1303027).

[3] Zhou, J. Discovering Association Rules in Engineering Documents. Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on (Digital Object Identifier: 10.1109/NLPKE.2003.1275927, Publication Year: 2003), pp. 339 – 344.

[4] Chao Tang and Chen Liu. Method of Chinese Grammar Rules Automatically Access Based on Association Rules. Computer Science and Computational Technology, 2008 (ISCSCT '08) International Symposium on Volume: 1 (Digital Object Identifier: 10.1109/ISCSCT.2008.68) Publication Year: 2008, pp. 265 – 268.

[5] Wu Gongxing. A Study on the Mining Algorithm of Fast Association Rules for the XML Data. Computer Science and Information Technology, 2008. (Digital Object Identifier: 10.1109/ICCSIT.2008.89 Publication Year: 2008), pp. 204 – 207.

[6] Al-Zoghby, A., Eldin, A.S., Ismail, N.A. and Hamza, T. Mining Arabic Text Using Soft-Matching Association Rules. Computer Engineering & Systems, 2007. (ICCES '07, Digital Object Identifier: 10.1109/ICCES.2007.4447080, Publication Year: 2007) 2007, pp. 421 – 426.

[7] Yong-le Sun and Ke-liang Jia. Research of Word Sense Disambiguation Based on Mining Association Rules. Intelligent Information Technology Application Workshops, 2009. Third International Symposium on (Digital Object Identifier: 10.1109/IITAW.2009.85, Publication Year: 2009), pp. 86-88.

[8] Doug Won Choi and Young Jun Hyun. Transitive Association Rule Discovery by Considering Strategic Importance Computer and Information Technology (CIT). 2010 IEEE 10th International Conference on (Digital Object Identifier: 10.1109/CIT.2010.292, Publication Year:2010), pp. 1654-1659.

[9] Nazish Asad, M. Younus Javed and Usman Qamar, "Association Rules Mining for Urdu Language", International Journal of Computer and Communication Engineering (IJCEE), ISSN 2010-3743, Vol 1, No. 1, May 2012.