Association Rules Mining for Urdu Language

Nazish Asad, M. Younus Javed, and Usman Qamar

***Abstract*—This paper explains the importance of Association Rules for several domains of Urdu language (e.g. education, publication and web development). Although various association rules mining techniques have successfully used for market basket analysis but no one has applied on Urdu text. A new mining model i.e. Urdu Mining Model (UMM) is proposed based upon Apriori algorithm and used to extract unique words and phrases from Urdu language. UMM is tested on an Urdu corpus and results have shown that it has worked well on Urdu text i.e. Apriori is really effective to mine text databases as well. Results are communicated for further research and enhancement of existing tools and systems which deal Urdu text in one way or the other.**

***Index Terms*—Urdu language, association rules mining, apriori algorithm.**

I. Introduction

Human history has shown that man is always interested in finding patterns from data since the beginning of life. With the progress of technology, most of the organizations and countries automate their systems and more advance methods are introduced to find out important patterns from these systems. These automated systems have produced and dealt with the variety of data i.e. images, texts, transactions, sounds, videos etc.

Association rules mining is an effective data mining technique to extract interesting patterns from transactional databases. This technique is usually used for market basket analysis i.e. to find out that which items are purchased together, so that management will be able to make effective decisions.

Association rule mining can also be used for mining association rules from textual data with few changes. Association rules mining for textual data can use to create statistical thesaurus, to extract grammatical rules and to search large online data efficiently.

Basic concepts of association rules are:

Let $I = \{i1,i2,\ldots\ldots..., \text{im}\}$ 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 $X \cap Y=\Phi$.

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 rule 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 is very important task as automation generates a lot of text data. 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.

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.

II. Related Work

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

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

Manuscript received February 28, 2012; revised April 29, 2012.

Authors are with the Department of Computer Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan (e-mail: nasad887@gmail.com, myjaved@ceme.nust.edu.pk, usman_zaman@yahoo.com).

association rules and item-item association rules.

In [4], Chao Tang has utilized Apriori algorithm to find out grammatical rules from Chinese text. A new model has proposed for grammatical rules mining which has three major steps: pre-processing, association rules mining and verification of association rules to get real rules. Four different corpuses have 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 has devised a new distributed algorithm to extract association rules for the XML data. This algorithm has created the DOM tree at the beginning and then has used this tree to extract association rules. The distributed algorithm has worked on multiple web sites. Each website has executed the FreqTree algorithm to compute local support count and sent it to global website. The global site then determined the global frequent items on the basis of sum of support counts which were gathered from all local sites. At the end, a verification process has applied to filter out the valid XML rules from the global frequent items.

In [6], Aya Al-Zoghby has devised 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 has utilized the Association rules mining based upon Apriori algorithm to solve the Word Sense Disambiguation problem. Apriori has successfully extract association rules between the sense of the ambiguous words and contexts and has generated very precise association rules.

In [8], Doug Won Choi has improved Apriori algorithm to discover the candidate itemsets and as well as the frequent itemsets. This algorithm has 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 has given a second chance to items so that more items can be generated in second attempt.

III. Association Rules Mining for Urdu Language

Association rules mining from Urdu language is a complex task because:

1) It is mixture of Turkish, Persian and Arabic languages,
2) It is written from right to left, opposite of many other international languages,
3) It has a larger itemset: contains total 52 characters (39 basic characters and 13 extra characters) and punctuation marks as well as that increases the total number of itemsets [9], and
4) It is necessary to perform pre-processing on Urdu text files to convert them in a transactional database.
5) Technique to tackle all above mentioned challenges are discussed in next section.

A. Mining Model for Urdu Text

As shown in fig. 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.

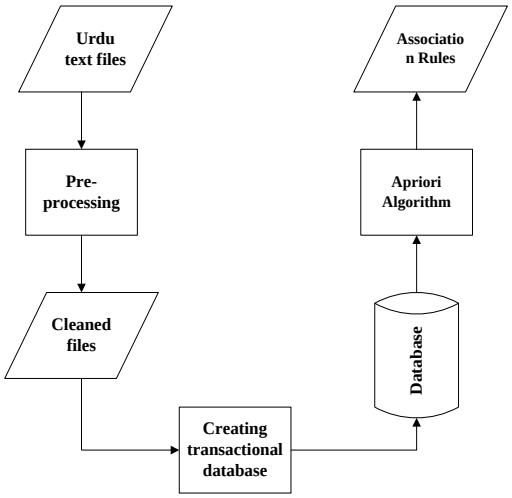

Fig. 1. Urdu text mining model.

IV. Experiments and Results

Implementation of Apriori algorithm on Urdu text has extracted interesting association rules from the database.

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. Apriori algorithm is tested on this pre-processed transactional database. Testing is done by using different number of words and minimum supports at average of 40%, 60% and 80% confidence.

Results in tab. 1 has shown that Apriori algorithms treated Urdu language in the same fashion as it did English language or any other language i.e. the number of association rules are decreased as the number of words are increased, validity of Urdu rules are greatly affected by the number of words being processed and increment of minimum support has become the reason of decrement in association rules.

Tab. 2 has indicated that time consumed by Apriori algorithm on Urdu database is inversely proportional to the minimum support i.e. increase in minimum support results as a decrease in time.

TABLE I: ASSOCIATION RULES AT DIFFERENT MINIMUM SUPPORTS.

Number of Words	Minimum Support				
	1.75%	**2.00%**	**3.00%**	**4.00%**	**5.00%**
20000	49	47	32	23	19
40000	52	48	32	23	19
60000	48	43	26	20	18
80000	43	42	26	20	18

TABLE II: TIME CONSUMPTION AT DIFFERENT MINIMUM SUPPORTS

Number of Words	Minimum Support				
	1.75%	**2.00%**	**3.00%**	**4.00%**	**5.00%**
20000	4.24	3.37	2.05	1.74	1.02
40000	8.19	6.70	4.07	2.42	2.21
60000	10.14	8.20	4.53	3.47	2.73
80000	12.89	10.20	7.34	3.87	3.54

V. CONCLUSION AND FUTURE WORK

Data mining techniques can effectively apply for extracting association rules from Urdu text. 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] J. Han and M. Kamber, *Data Mining Concepts and Techniques*.

[2] J. D. Holt and S. M. Chung, "Parallel Mining of Association Rules from Text Databases on a Cluster of Workstations," in *Proceedings of the 2004 18th international Parallel and Distributes Processing Symposium*, 2004.

[3] J. Zhou, "Discovering Association Rules in Engineering Documents," in *proc. of International Conference on Natural Language Processing and Knowledge Engineering*, 2003, pp. 339 – 344.

[4] C. Tang and C. Liu, "Method of Chinese Grammar Rules Automatically Access Based on Association Rules," in *Proc. Computer Science and Computational Technology(ISCSCT 2008)* vol. 1, 2008, pp. 265 – 268.

[5] G. Wu, "A Study on the Mining Algorithm of Fast Association Rules for the XML Data," *Computer Science and Information Technology*, 2008, pp. 204 – 207.

[6] A. Al-Zoghby, A. S. Eldin, N. A. Ismail, and T. Hamza, "Mining Arabic Text Using Soft-Matching Association Rules," *Computer Engineering and Systems*, 2007, pp. 421 – 426.

[7] Y. Sun and K. Jia, "Research of Word Sense Disambiguation Based on Mining Association Rules," *Intelligent Information Technology Application Workshops*, 2009, pp. 86-88.

[8] D. W. Choi and Y. J. Hyun, "Transitive Association Rule Discovery by Considering Strategic Importance Computer and Information Technology (CIT)," in *proc. 2010 IEEE 10th International Conference* , 2010, pp. 1654-1659.

[9] BBC. [Online]. Available: www.bbc.co.uk/language/other/guide/urdu/history.shtml