Result Analysis Using Various Pattern Mining Techniques:-

A Recommendation to Strengthen the Standard of Technical Education

Dr. Kanak Saxena
Professor & Head,
Computer Application
SATI, Vidisha,
kanak.saxena@gmail.com

D.S. Rajpoot
Ph.D Research Scholar,
RGPV, Bhopal
dsrphd@yahoo.com

***Abstract:* We are trying to use Data mining techniques in such a manner so that we will be able to learn pattern of large database in such a way to understand effect of changing or modification in Present database as per there pattern learning into the future. It is impossible to supposed future effect of any changes in present Data. However, there are important new issues which arise because of the sheer size of the data. One of the important problems in data mining is the Classification-rule learning which involves finding rules that partition given data into predefined classes. In the data mining domain where millions of records and a large number of attributes are involved, the execution time of existing algorithms can become prohibitive, particularly in interactive applications we are trying to learn pattern of different Result data.**

I. Introduction

In General we are unable to decide how we should justify a change in syllabus of any course until unless we have not seen the effect, which is also difficult because of not able to seen in future. Our work support to think in future about positive effect of change. In classification/clustering we analyze a set of data and generate a set of grouping rules which can be used to classify future data. For example, one may classify diseases and provide the symptoms which describe each class or subclass. We can also use Data mining technique to understand crime pattern and its future repetitions [8-11]

In *sequential Analysis*, we seek to discover patterns that occur in sequence. This deals with data that appear in separate transactions (as opposed to data that appear in the same transaction in the case of association). For e.g.: If a shopper buys item A in the first week of the month, then she/he buys item B in the second week etc.

There are many algorithms proposed that try to address the above aspects of data mining. Compiling a list of all algorithms suggested/used for these problems is an arduous task. I have thus limited the focus of this report to list only some of the algorithms that have had better success than the others. Algorithm which are developed having complexity. We will consider the platform and coding so that the earlier complexity of the algorithm will be reduced.

II. Previous Work has done:

We now return to the sequential pattern mining framework of Agrawal & Srikant [13] which basically extends the frequent item sets idea described above to the case of patterns with temporal order in them. The database *D* that we now consider is no longer just some unordered collection of transactions. Now, each transaction in *D* carries a time-stamp as well as a customer ID. Each transaction (as earlier) is simply a collection of items. The transactions associated with a single customer can be regarded as a sequence of itemsets (ordered by time), and *D* would have one such transaction sequence corresponding to each customer. In effect, we have a database of transaction sequences, where each sequence is a list of transactions [15] ordered by transaction-time.

The temporal patterns of interest are also essentially some (time ordered) sequences of itemsets. A sequence s of itemsets is denoted by $s_1,s_2, \cdots s_n$, where s_j is an itemset..

While we described the framework using an example of mining a database of customer transaction sequences for temporal buying patterns, this concept of sequential patterns is quite general and can be used in many other situations as well. Indeed, the problem of motif [14] discovery in a database of protein sequences that was discussed earlier can also be easily addressed in this framework. Another example is web navigation mining. Here the database contains a sequence of websites that a user navigates through in each browsing session. Sequential pattern mining can be used to discover those sequences of websites that are frequently visited one after another.

We next discuss the mechanism of sequential pattern discovery. The search for sequential patterns begins with the discovery of all possible itemsets with sufficient support. The

ISSN : 0975-3397

Apriori algorithm described earlier can be used here, except that there is a small difference in the definition of support. Earlier, the support of an itemset was defined as the fraction of *all* transactions that contained the itemset. But here, the support of an itemset is the fraction of *customer transaction sequences* in which at least one transaction contains the itemset. Thus, a frequent itemset is essentially the same as a large 1-sequence (and so is referred to as a *large* itemset or *litemset*) . Once all litemsets in the data are found, a transformed database is obtained where, within each customer transaction sequence, each transaction is replaced by the litemsets contained in that transaction.

The next step is called the sequence phase, where again, multiple passes are made over the data. Before each pass, a set of new potentially large sequences called candidate sequences are generated. Two families of algorithms are presented by Agrawal & Srikant (1995) [13] and are referred to as count-all and count-some algorithms. The count-all algorithm first counts all the large sequences and then prunes out the non-maximal sequences in a post-processing step. This algorithm is again based on the general idea of the Apriori algorithm of Agrawal & Srikant (1994) [13] for counting frequent itemsets. In the first pass through the data the large 1-sequences (same as the litemsets) are obtained. Then candidate 2-sequences are constructed by combining large 1-sequences with litemsets in all possible ways. The next pass identifies the large 2-sequences. Then large 3-sequences are obtained from large 2-sequences, and so on.

The count-some algorithms by Agrawal & Srikant (1995) intelligently exploit the maximality constraint. Since the search is only for maximal sequences, we can avoid counting sequences which would anyways be contained in longer sequences. For this we must count longer sequences first. Thus, the count-some algorithms [16] have a forward phase, in which all frequent sequences of certain lengths are found, and then a backward phase, in which all the remaining frequent sequences are discovered. It must be noted however, that if we count a lot of longer sequences [17] that do not have minimum support, the efficiency gained by exploiting the maximalist constraint, may be offset [18] by the time lost in counting sequences without minimum support (which of course, the count-all algorithm would never have counted because their subsequences were not large). These sequential pattern discovery [21] algorithms are quite efficient and are used in many temporal data mining applications and are also extended in many directions.

The last decade has seen many sequential pattern mining methods being proposed from the point of view of improving upon the performance of the algorithm by Agrawal & Srikant (1995) [1]. Parallel algorithms for efficient sequential pattern discovery are proposed by Shintani & Kitsuregawa (1998) [7]. The algorithms by Agrawal & Srikant (1995) need as many database passes as the length of the longest sequential pattern. Zaki (1998) [20]proposes a lattice-theoretic approach to decompose the original search space into smaller pieces (each of which can be independently processed in main-memory) using which the number of passes needed is reduced considerably. Lin & Lee (2003) propose a system for interactive sequential pattern discovery, where the user queries with several minimum support thresholds iteratively and discovers the desired set of patterns corresponding to the last threshold.

Another class of variants of the sequential pattern mining framework seek to provide extra user-controlled focus [19] to the mining process. For example, Srikanth & Agrawal (1996) generalize the sequential patterns framework to incorporate some user-defined taxonomy of items as well as minimum and maximum time-interval constraints between elements in a sequence. *Constrained association queries* are proposed (Ng *et al* 1998) where the user may specify some domain, class and aggregate constraints on the rule antecedents and consequents. Recently, a family of algorithms called *SPIRIT* (Sequential Pattern mining with Regular expression constraints) is proposed [20] in order to mine frequent sequential patterns that also belong to the language specified by the user-defined regular expressions (Garofalakis *et al* 2002).

The performance of most sequential pattern mining algorithms suffers when the data has long sequences with sufficient support, or when using very low-support thresholds. One-way [22] to address this issue is to search, not just for large sequences (i.e. those with sufficient support), but for sequences that are *closed* as well. A large sequence is said to be *closed* if it is not properly contained in any other sequence which has the same support. The idea of mining data sets for frequent *closed* item sets [23] was introduced by Pasquier *et al* (1999). Techniques for mining sequential closed patterns are proposed by Yan *et al* (2003); Wang & Han (2004). The algorithm by Wang & Han (2004) is particularly interesting in that it presents an efficient method for mining sequential closed patterns without an explicit iterative candidate generation step.

III. IDENTIFIED WORK IN THIS FIELD :

An example of such a pattern is that customers typically rent ``Star Wars", then ``Empire Strikes Back", and then ``Return of the Jedi". Note that these rentals need not be consecutive. Customers who rent some other videos in between also support this sequential pattern. Elements of a sequential pattern need not be simple items. ``Fitted Sheet and flat sheet and pillow cases", followed by ``comforter", followed by ``drapes and ruffles" is an example of a sequential pattern in which the elements are sets of items. This problem was initially motivated by applications in the retailing industry, including attached mailing, add-on sales, and customer satisfaction. But the results apply to many scientific and business domains. For instance, in the medical domain, a data-sequence may correspond to the symptoms or diseases of a patient, with a transaction corresponding to the symptoms exhibited or diseases diagnosed during a visit to the doctor. The patterns discovered using this data could be used in disease research to help identify symptoms/diseases that precede certain diseases.

The task of sequential patterns in knowledge discovery and data mining is to identify the item that frequently precedes another item. Generally a sequential pattern can be described as a finite series of elements such as $A \rightarrow B \rightarrow C \rightarrow D$ where A, B, C, and D are elements of the same domain. Each sequential pattern in data mining comes with a minimum support value, which indicates the percentage of

ISSN : 0975-3397

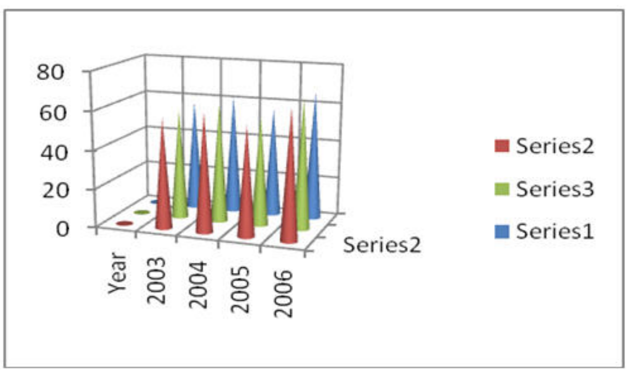
total records that contain the pattern. An arbitrary example of a sequential pattern is 90% of the die-hard fans who saw the movie Titanic went on to buy the movie sound track CD, followed by the video-tape when it was released. The primary goal of sequential pattern discovery is to assess the evolution of events against a measured time-line and detect changes that might occur coincidentally. This information has been used to detect medical fraud in insurance claims, evaluate drug performances in pharmaceutical industry, and determine risk factors in military operations.

IV. RESULT ANALYSIS OF USED DATA FOR LAST 5 YEARS :

We have taken result data of 5 years and have study about their pattern, we find that there is a lacunae during redesigning of syllabus because mistakes are repeated and we have not modified syllabus in correct directions

2.1 Year wise Data for Subject Code BE-201

Year	percent
2003	66.9
2004	69.9
2005	76.4
2006	69.8
2007	38.7

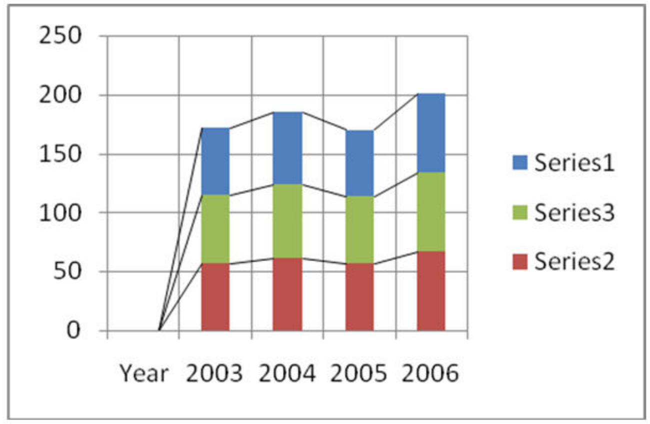


RResult Graph for BE-201

We have seen in fig 2.1 data table that in 2007 we have got reduced result, which is only 38.7 percentage , so clearly we can say that we had to revised syllabus of Subject code BE-201 in year 007 to improve result in 2008-09.

2.2 **Year** wise Data for Subject Code BE-203

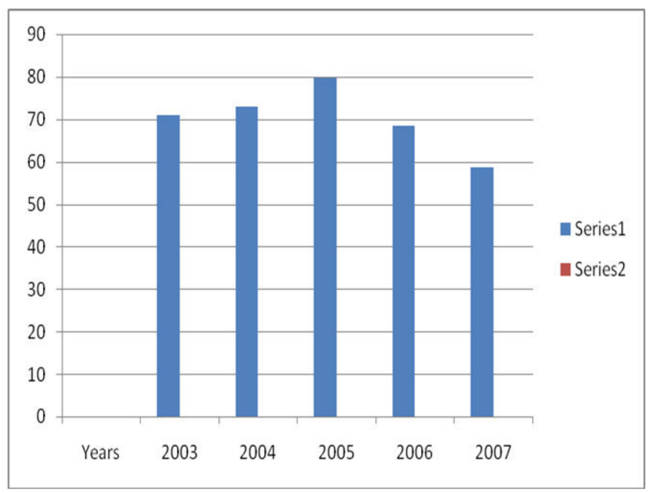
Year	percent
2003	57.29
2004	61.71
2005	56.73
2006	66.9
2007	54.16



2.2 Result Graph for BE-203

2.3 Year wise Data for Subject Code BE-204

Years	Percentage
2003	71.03
2004	73.04
2005	79.8
2006	68.52
2007	58.76



2.3 Result Graph for BE-204

We have seen that in graph 2.2 that in BE-203 we have note down that we have got satisfactory result in 2004 and then it degraded in 2005 and after some modifications in syllabus we have a better result in 2006 but again in 2007 we have a failure in result as a percentage degradations.

Similarly we can see the pattern of BE-204 for year 2003-07. After year 2005 we have a degraded result in 2006 and 2007 we have a continuous degraded result so it is essential advise to strengthen the Institution result to use a technique based on data mining.

ISSN : 0975-3397

V. CONCLUSIONS & FUTURE WORK:

We have seen that as per our study about patterns of data of result displayed by University we can better understand how to Redesigned and modified syllabus of any university to enhance result percentage, we have tyied to study of subjects of a Single Semester subjects having different Codes and we find that it's a big issue for present scenario to enhance result to get better ranking and recognition within the world of Well Recognise Universities. In future we are also going to learn and study of more data with Well define Software and then we will put our crisp thoughts about modifying the Syllabus.

VI. REFERENCES

[1] Hsieh, Chia-Ying; Yang, Don-Lin; Wu, Jungpin;" An Efficient Sequential Pattern Mining Algorithm Based on the 2-Sequence Matrix" ICDMW '08 . IEEE International Conference on 15-19 Dec.2008 Page(s):583-591.

[2] Xue, Anrong; Hong, Shijie; Ju, Shiguang; Chen, Weihe;" Application of sequential patterns based on user's interest in intrusion detection" IT in Medicine and Education, 2008. ITME 2008. IEEE International Symposium on 12-14 Dec. 2008 Page(s):1089 - 1093

[3] Chuancong Gao, Jianyong Wang, Yukai He, Lizhu Zhou; " Efficient Mining of Frequent Sequence Generators " ACM, Beijing, China, April 21-25, 2008.

[4] Y.Hirate, H.Yamana; "Sequential Pattern Mining with Time Interval," 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006), Singapore, April 9-12, 2006.

[5] Mendes, L.F. Bolin Ding Jiawei Han "Stream Sequential Pattern Mining with Precise Error Bounds" ICDM '08. Eighth IEEE International Conference 15-19 Dec. 2008.

[6] Lei Chang Tengjiao Wang Dongqing Yang Hua Luan; "SeqStream: Mining Closed Sequential Patterns over Stream Sliding Windows", Data Mining, 2008. ICDM '08. Eighth IEEE International Conference 15-19 Dec. 2008 page(s): 83-92

[7] Srivatsan Laxman & P S Sastry; " A Survey of Temporal Data Mining" Vol.31 Part 2, April 2006.

[8] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, "Crime Data Mining: An Overview and CaseStudies", AI Lab, University of Arizona, proceedings National Conference on Digital Government Research, 2003, available at: http://ai.bpa.arizona.edu/.

[9] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, "Crime Data Mining: A General Framework and Some Examples", IEEE Computer Society April 2004.

[10] C McCue, "Using Data Mining to Predict and Prevent Violent Crimes", available at: http://www.spss.com/dirvideo/richmond.htm?source=dmpage&zone=rtsidebar.

[11] Cesar Vialardi, Javie Bravo, Leila Shafti, Alvaro Otiose, "Recommendation in Higher Education Using Data Mining Techniques" Educatinal Data Mining Report, 2009.

[12] R. Agrawal, R. Srikant, ``Mining Sequential Patterns", Proc. of the Int'l Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995.

[13] Ayres J, Gehrke J, Yu T and Flannick J: "Sequential Pattern Mining using a Bitmap Representation" in Int'l Conf Knowledge Discovery and Data Mining, (2002) 429-435

[14] Garofalakis M, Rastogi R and Shim k, "Mining Sequential Patterns with Regular Expression Constraints", in IEEE Transactions on Knowledge and Data Engineering,(2002), vol. 14, nr. 3, pp. 530-552

[15] Pei J, Han J. et al: "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth" in Int'l Conf Data Engineering, (2001) 215-226

[16] Pei J and Han J: "Constrained frequent pattern mining: a pattern-growth view" in SIGKDD Explorations, (2002) vol. 4, nr. 1, pp. 31-39

[17] Antunes C and Oliveira A.L: "Generalization of Pattern-Growth Methods for Sequential Pattern Mining with Gap Constraints" in Int'l Conf Machine Learning and Data Mining, (2003) 239-251

[18] R. Srikant, R. Agrawal: ``Mining Sequential Patterns: Generalizations and Performance Improvements", Proc. of the Fifth Int'l Conference on Extending Database Technology (EDBT), Avignon, France, March 1996.

[19] Zaki M, "Efficient Enumeration of Frequent Sequences", in ACM Conf. on InformationKnowledge Management, (1998) 68-75

[20] R. Agrawal, A. Arning, T. Bollinger, M. Mehta, J. Shafer, R. Srikant: "The Quest Data Mining System", Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August, 1996.

[21] Eui-Hong (Sam) Han, Anurag Srivastava and Vipin Kumar: "Parallel Formulations of Inductive Classification Learning Algorithm" (1996).

[22] Agrawal, R. Srikant: ``Fast Algorithms for Mining Association Rules", Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, Sept. 1994.

[23] J. Han, J. Chiang, S. Chee, J. Chen, Q. Chen, S. Cheng, W. Gong, M. Kamber, K. Koperski, G. Liu, Y. Lu, N. Stefanovic, L. Winstone, B. Xia, O. R. Zaiane, S. Zhang, H. Zhu, `DBMiner: A System for Data Mining in Relational Databases and Data Warehouses", Proc. CASCON'97: Meeting of Minds, Toronto, Canada, November 1997.

[24] Cheung, J. Han, V. T. Ng, A. W. Fu an Y. Fu, `` A Fast Distributed Algorithm for Mining Association Rules", Proc. of 1996 Int'l Conf. on Parallel and Distributed Information Systems (PDIS'96), Miami Beach, Florida, USA, Dec. 1996.

[25] Ron Kohavi, Dan Sommerfield, James Dougherty, "Data Mining using MLC++ : A Machine Learning Library in C++", Tools with AI, 1996

ISSN : 0975-3397