INTERNATIONAL JOURNAL OF RESEARCH
In Advanced Engineering Technologies

Volume 4, Issue 2 JULY 2015

# Avoiding Low-Frequency & Misinterpretation Problems for Text Mining using Effective Pattern Discovery Technique

**[1] PANTAMALLA SHARON PRADEEP, [2] K.SREERAM MURTHY**

[1]M.Tech Student, Department of IT, Sreenidhi institute of science & technology, Yamnampet (v), Ghatkesar(m), Ranga Reddy(d), Telangana state, India.

[2] Assistant Professor, Department of IT, Sreenidhi institute of science & technology, Yamnampet (v), Ghatkesar(m), Ranga Reddy(d), Telangana state, India.

***Abstract*— *Data mining is that the method of analyzing information from totally different views and summarizing it into helpful data - data that may be accustomed increase revenue, cuts costs, or both. Technically, data mining is that the process of finding correlations or patterns among dozens of fields in massive relative databases. several techniques are investigated on mining situation from documents together with the texts for needed patterns severally. there's a retardant on managing this explicit task for inventory patterns that area unit correct. analysis has done on this strategy, results have established that this strategy is facing 2 issues, i.e., 1) semantic relation and 2) lexical ambiguity (coexistence of the many doable meanings for a word). So, within the text mining, we will use the techniques of pattern mining to search out totally different text patterns, like co-occurring terms, frequent itemsets. thus currently this gift technique i.e., inventory pattern plays an important role within the investigation of the patterns. we tend to 1st conduct a group of large-scale measurements with a group of over totally different information sets into a information. we tend to transfer this information consisting of information sets for constructing the pattern taxonomy model, partial conflict tree and Chart. supported the measuring results, we've established that this system works expeditiously and effectively. It conjointly provides smart results for the implementation of task.***

***Key words: Patterns, Associations, or Relationships, Sequence patterns, Classification of text.***

## I. INTRODUCTION

Today, we've much more data than we will handle: from business transactions and scientific information, to satellite photos, text reports and military intelligence. data retrieval is just not enough any longer for decision-making. Confronted with immense collections of information, we've currently created new must facilitate United States build higher social control selections. These desires area unit automatic account of information, extraction of the "essence" of data hold on, and also the discovery of patterns in data. With this, data processing with inventory pattern came into existence and got popularized. data processing finds these patterns and relationships exploitation information analysis tools and techniques to create models. There area unit 2 main styles of models in data processing. One is prognosticative models, that use information with acknowledged results to develop a model which will be accustomed expressly predict values. Another is descriptive models, that describe patterns in existing information. All the models area unit abstract representations of reality, and may be guides to understanding business and recommend actions. These information provides several advantages and plays a significant role within the entire society in terms of social control business and analyzing the market by the actual extraction severally. methoding is a vital a part of data discovery process that we will analyze a vast set of information and acquire hidden and helpful knowledge. data processing is applied effectively not solely in business surroundings however conjointly in alternative fields like weather outlook, medicine, transportation, healthcare, insurance, government…etc.

For the extraction of the information and knowledge, an outsized range of investigations had been taken place within

the mining of the information antecedently. There area unit some existing techniques follows mining by the rule of association, successive patterns, Mining by the assistance of item set phenomena, Mining by agglomeration, Mining by the classification, Mining by the Prediction. There area unit some limitations within the this bestowed techniques. they're enforced beneath one explicit frame of your time, i.e., beneath the restricted areas. even if when generating the patterns there's a tangle in mining, they lend themselves discovering frequent item sets and also the order they seem. therefore to beat of these issues a selected methodology has been enforced that plays a key and crucial role within the field of mining of text within the kind of extraction of the information accurately followed by a decent change method within the entire phenomena. Mining of text is that the detection of information supported the interest within the documents consisting of text severally. currently there's major difficult task concerned in it wherever it should give service for the user within the kind of correct extraction of the data within the mining of text within the kind of generated patterns. There area unit sizable amount of experimental analysis takes place within the assortment of the information and conjointly the retrieval of the text accurately and exactly counting on the user's alternative. thus our gift methodology overcomes all the on top of issues accurately and carries the success.

**Architecture:**

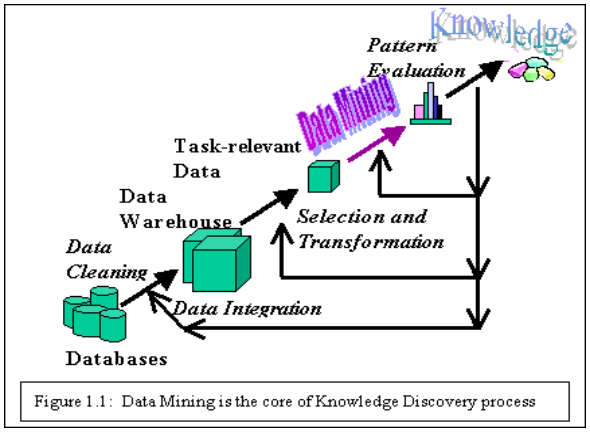



Figure 1.1: Data Mining is the core of Knowledge Discovery process

## II. RELATED WORK

**(a) Models of baseline:**

The models supported the baseline square measure differentiated in to a few phases for the effective execution of the performance of the system behavior and therefore the three phases square measure as follows strategies supported the mining of the data , coming up with of the model supported the abstract approach and eventually coming up with of the supported terms severally.

**(b) Measure:**

within the time of implementation of the projected methodology several factors are taken into the thought that\'s the speed of recall and conjointly the speed of exactitude respectively. Here the speed of exactitude is termed as knowledge retrieval that square measure analogous to the subject and within the alternative hand rate of recall is that the destruction of similarity within the retrieved documents.

**(c) Presumption:**

Here the foremost task involvement takes place wherever it's to indicate the differentiation between the projected technique and therefore the varied existing strategies and the way come back it overcomes the issues of the prevailing strategies and the way correct it and the way effective and economical respectively. this can be a primary concern at the time of analysis of the performance of the projected enforced approach.

## III. FRAME WORK

To improve the performance within the mining of a text based mostly aspects within the sort of the patterns associated with the closed strategy an additional info is employed within the system for this purpose we tend to square measure speculated to hold that individual phenomena by the name as D pattern severally that is principally used for the burden evaluation. thus this D patterns plays a significant role within the sort of the technique by the name term within which there isn't any word for the bigger price. thus these specific strategy is totally differed from the generalized state of affairs basis wherever it's completely repose passionate about the terms concerned within the documentation in it. more up the

performance of the system within the sort of affectivity within the mining pattern of taxonomy wherever the differentiation of the rule takes place by the name of the mining supported SP methodology severally so as to look at the comparatively ordered patterns wherever the house of looking is got negotiated. currently moving towards the reciprocal documents that's the alternative of the conventional state of affairs based mostly documents wherever the shuffling involves in it and additionally the that provides support by the pattern D within the original kind. so as to the present a patterns generated by the noise gets invalidated as a result of the less frequency destined knowledge and is termed because the evolution of the inner patterns. Here the thought takes solely with relevance the inner phenomena wherever because the outer phenomena got at large from the system. currently here the main concern is comparable and non similar knowledge should be differentiated thus so as to separate between themselve a state should be set by the name of the brink severally. The shaping of the brink takes place by the assistance of the formatting of the setting D pattern. and thus the brink operate is delineate by the subsequent equation

$$Threshold(D\ ') = \min_{p \in D} \left( \sum_{(t,w) \in \beta(p)} support(t) \right).$$

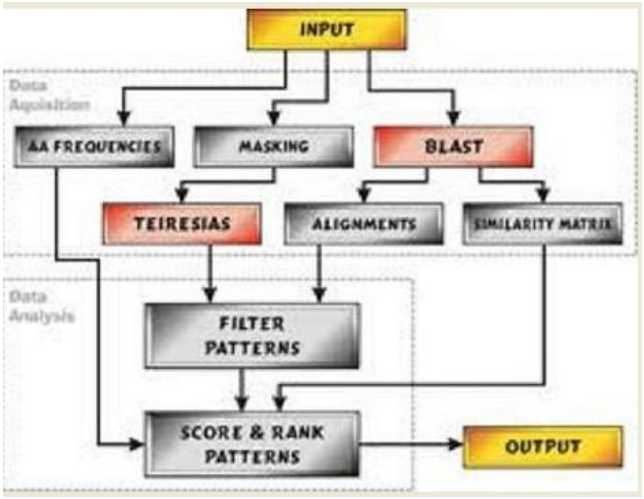
The tabular column shows the dataset containing itemsets in paragraphs respectively

| Paragraph | Terms |
|---|---|
| P1 | BOOKS,CD |
| P2 | DVD,VIDEO,GAMES |
| P3 | DVD,VIDEO,MOUSE,KEYBOARD |
| P4 | DVD,VIDEO,MOUSE,KEYBOARD |
| P5 | BOOKS,CD,KEYBOARD,SPEAKER |
| P6 | BOOKS,CD,KEYBOARD,SPEAKER |

## IV. EXPERIMENTAL RESULTS

In the gift section by the evolution of patterns within the model of taxonomy, experimental results of PTM approach ar bestowed and analyzed. The tabular column shows however ofttimes patterns occur in covering set and additionally displayed within the figure. As already mentioned earlier that mining of the information supported the item set wherever it's happy with fast generation however it suffers from the implementation analysis severally. Here the comparison takes place by the specialists and provides the information results accurately and within the far more economical manner. Our gift technique plays an important role on mining supported the terms and additionally on pattern dependent mining. and a few of them includes support vector machine and also the state of art.

**(a) Effective Inventory pattern:**




The above figure describes the step by step process of how the input has been converted to output i.e., inventory patterns.

**(b) Pattern Taxonomy model:**

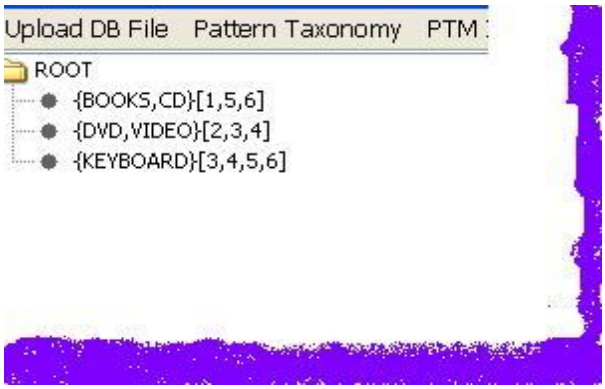
Upload DB File  Pattern Taxonomy  PTM IPE

| Frequent Pattern | Covering Set | Support |
|---|---|---|
| BOOKS,CD | P1,P5,P6 | 3.0 |
| DVD,VIDEO | P2,P3,P4 | 3.0 |
| BOOKS | P1,P5,P6 | 3.0 |
| CD | P1,P5,P6 | 3.0 |
| DVD | P2,P3,P4 | 3.0 |
| VIDEO | P2,P3,P4 | 3.0 |

After uploading the database, we will get the above figure, which shows how frequent the patterns appear in covering set

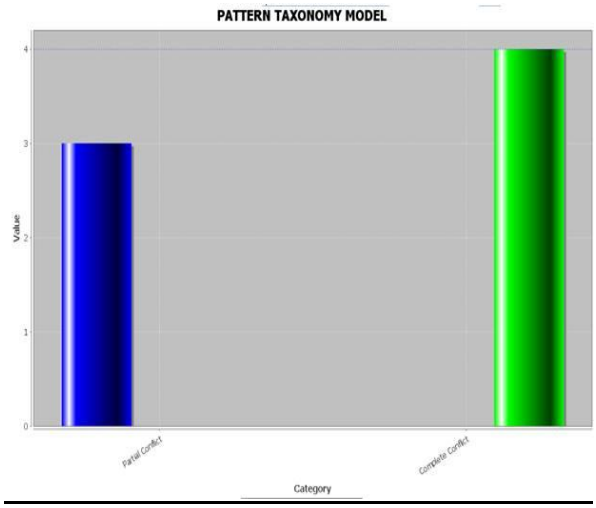
i.e., set of paragraphs and also shows the support value of each pattern.

**(c) Conflict tree:**




The above figure shows how set of different items appear in different paragraphs

**(d) Chart:**




above figure shows how the partial conflict differs from the complete conflict.

## V. CONCLUSION

Research are done on this specific drawback. several techniques are concerned before this system and a few of them area unit mining by association, ordered mining of the patterns, Pattern maximization, closeness of Pattern etc. The study of the on top of analysis orientating ideas involves heap of effort and could be a immense tedious job within the stream of mining text and lacks potency and is a smaller amount effective. additionally to the on top of drawback it\'s the frequency problems. it\'s low frequency elements. Low frequency means that patterns generated with most of them area unit little and ineffective. so as to beat these issues a brand new technique is enforced for learning low frequency knowledge and conjointly for learning mismatched issues severally. The technique that is planned during this paper deals with the evolution and preparation of the patterns within the mining of text. sensible approaches make sure that it not solely used for the information text however conjointly used for the state-of-art that\'s the support vector machine severally.

REFERENCES


[1] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile,"Kernel Methods for Document Filtering," TREC, trec.nist.gov/pubs/trec11/papers/kermit.ps.gz, 2002.

[2] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization," Proc. 14th Int'l Conf. Machine Learning (ICML '97), pp. 143-151, 1997.

[3] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 200-209, 1999.

[4] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

[5] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994.

[6] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrases in Automated Text Categorization,"

Technical Report IEI-B4-07- 2000, Instituto di Elaborazione dell'Informazione, 2000.

[7] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.

[8] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '00), pp. 1-12, 2000.

[9] M. Zaki, "Spade: An Efficient Algorithm for Mining Frequent Sequences," Machine Learning, vol. 40, pp. 31-60, 2001.

[10] Y. Yang and X. Liu, "A Re-Examination of Text Categorization Methods," Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '99), pp. 42-49, 1999.

[11] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization," Information Retrieval, vol. 1, pp. 69-90, 1999.

[12] X. Yan, J. Han, and R. Afshar, "Clospan: Mining Closed Sequential Patterns in Large Datasets," Proc. SIAM Int'l Conf. Data Mining (SDM '03), pp. 166-177, 2003.

[13] Y. Xu and Y. Li, "Generating Concise Association Rules," Proc. ACM 16th Conf. Information and Knowledge Management (CIKM '07), pp. 781-790, 2007.