Volume 4, Issue 9, September 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Performance Enhancement for Effective Pattern Discovery & Concept Formation in Text Mining

Ms Sonali Gaikwad
Department of Computer Engg, DYPIET Pimpri
SavitriBai Phule Pune University, India

Prof Archana Chaugule
Department of Computer Engg, DYPIET Pimpri
SavitriBai Phule Pune University, India

Abstract— Due to rapid growth of digital data usages in recent database applications it is required to utilize effective text mining method for efficient performance. Text mining is research area because in many applications database content is not only in form of numerical data but also textual data form. To exploit hidden information from non-structured to semi-structured data form text mining is applicable. The challenging issue in text mining is to extract user required information in effectual manner. To perform this task text mining methods are used. Text mining methods categorization is based on how text document are analyzed. In these methods text document analyzed on the basis of term, phrase, concept and pattern. The pattern based approach can improve the accuracy of system for evaluating term weights because discovered patterns are more specific than whole documents. Objective is to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Proposed system increases efficiency of pattern discovery using different data mining algorithms with pattern deploying and pattern evolving method to solve misinterpretation and low frequency problem. Performance enhancement is achieved by applying multithreaded approach and concept formation from effectively discovered patterns.

Keywords— Sequential pattern mining, pattern deploying, pattern evolving, text mining.

I. INTRODUCTION

Now a day's abundant information is available on web it is good because it provides greater awareness and better knowledge. The goal of data mining is to discover hidden useful information in large databases. To extract useful information from various databases data mining technology are used. For numerical information data warehouses work well, but unsuccessful when it came to textual information. Text data mining refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. As text mining is extraction of useful information from text data it is also known as text data mining or knowledge discovery from textual databases. It is challenging issue to find accurate knowledge in text documents to help users to find what they want.

To large document collections text mining is an application of natural language processing. Nowadays most of the information in business, industry, government and other institutions is stored in text form into database and this text database contains semi structured data. Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. The modelling and implementation of semi structured data studies done in recent database research. On the basis of these researches information retrieval techniques such as text indexing methods have been developed to handle unstructured documents. In traditional search the user is typically look for already known terms and has been written by someone else. The problem is in result as it is not relevant to users need. This is the goal of text mining to discover unknown information which is not known and yet not written down.

As the natural language is not free from the ambiguity problem complexity of natural language is main challenging issue in text mining. Things can be understood in two or more possible ways and creates ambiguity so it cannot be entirely eliminated from the natural language as it gives flexibility and usability. There are various ways to interpret one phrase or sentence thus various meanings can be obtained. One word may have multiple meanings and multiple words can have same meaning. It is challenge to answer what user wants as semantic meanings of many discovered words are uncertain. The number of researches in resolving the ambiguity problem is going on the work is still not fully formed.

To instruct computers how to analyse, understand and generate text, technologies are produced by natural language processing. Text mining is discovery by computer for extracting new, previously unknown information and to automatically extract information from different written resources. The enormous amount of information stored in unstructured texts cannot simply be used for further processing by computers, which typically handle text as simple sequences of character strings. Therefore, specific pre-processing methods and algorithms are required in order to extract useful patterns.

In order to solve the above mentioned problem focus is on finding better text representation from textual data collection. One solution is to use data mining techniques, such as sequential pattern mining, for building up a representation with the new type of features. Such data mining-based methods adopted the concept of closed sequential patterns and pruned non closed patterns from the representation with an attempt to reduce the size of feature set by removing noisy patterns. However treating multi term pattern as an atom in representation seems likely to low frequency problem while dealing with long patterns. The technologies

like information extraction, summarization, categorization, clustering and information visualization are used in the text mining process.

II. METHODS AND MODELS USED IN TEXT MINING

Text mining exploit hidden information from text document to do so document analysis is basic process. Traditionally many techniques developed to solve the problem of text mining that is nothing but the relevant information retrieval according to user's requirement. Text mining methods categorization is based on how text document are analysed. In these methods of text mining text document analysed on the basis of term, phrase, concept and pattern. So according to the information retrieval basically there are four methods,

1) Term Based Method (TBM).
2) Phrase Based Method (PBM).
3) Concept Based Method (CBM).
4) Pattern Taxonomy Method (PTM).

A. Termed Based Methods

The Term in document is unit used to identify content of text. In Term Based Method each term in document is associated with value known as weight, which measure importance of term i.e. terms contribution in document. Word having semantic meaning is known as term and collection of such terms contributes meaning to document. Term based methods suffer from the problems of polysemy and synonymy. Polysemy means a word has multiple meanings and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want. Information retrieval provided many term-based methods like supervised and traditional term weighting methods to solve this challenge. The evolution of term weights is based on distribution of term in document.

The term frequency TF (t, d) is number of times term 't' occurs in document 'd'. The document frequency DF (t) is number of documents in which term't' occurs at least once. The inverse document frequency IDF(t) can be calculated from document frequency[11].

$$\mathrm{IDF(t)} = \log\left(\frac{|D|}{DF(t)}\right)$$

|D| is total number of documents. The inverse document frequency of term is low if it occurs in many documents and is highest if term occurs only in one document. The value of W_i i.e. weight of term of document 'd' calculated by product as,

$W_i = TF(t_i, d) * IDF(t_i)$

Using this mathematical model term based method analyze document to extract features by TFIDF feature selection approach.

B. Phrased Based Models

Title Phrases are less ambiguous and more discriminative than individual term so in phrase based method document is analyzed on phrase basis. In process of analysis of document phrases are profile descriptor of document. Phrases are collection of semantic terms so carries more information than single term. Over many years this is hypothesis that phrase based approach performs better than term based approach, as phrase may carry more semantic than term. Using data mining algorithms it is definite to obtain various phrases but it is difficult to use these phrases effectively to answer what user want. It is difficult because phrases have fewer occurrences in document and phrases comprise large number of noisy with redundant terms.

As phrases are collection of terms those can be considered as sequence of terms and hence to find sequence of terms sequential pattern mining algorithm is used. Algorithm extracts frequent sequential patterns, here pattern used as words or phrase which is extracted from document. First parameter PL is list of 'n' terms frequent sequential pattern. Second parameter is minsup is minimum relative support. In application of phrase based system algorithm is invoked recursively [6].

Algorithm: SPMining(PL, min sup)
Input: List of n Terms frequent sequential patterns (PL), minimum support (min sup)
Output: Set of frequent sequential patterns (SP).
Method:

1. SP ← SP − {Pa ∈ SP | ∃Pb ∈ P L such that len(Pa) = len(Pb) – 1∧Pa ` Pb ∧ suppa (Pa) = suppa (Pb)} // pattern pruning
2. SP ← SP ∪ P L
3. PL ← ∅
4. For each pattern p in P L do begin
5. generating p-projected database P D
6. for each frequent term t in P D do begin
7. P =pt
8. if suppr (P) ≥ min sup then
9. PL ← PL ∪ P
10. end if

11. end for
12. end for
13. if |P L | = 0 then
14. return
15. else
16. call SPMining(P L , min sup)
17. end if
18. output SP

Removal of meaningless terms is known as pruning process which is required for reducing cost of computation and improving effectiveness of system. First line presents pruning step which is to eliminate non-closed patterns. The algorithm starts to mine 'n' terms from projected database, if relative support of terms is greater than or equal to minimum support then will store those patterns. The algorithm repeats itself recursively until there is no more pattern discovered. As a result, the output of algorithm SPMining is a set of closed sequential patterns with relative supports greater than or equal to a specified minimum support. SPMining algorithms feature is it deals with several sequences at a time whereas others only handle one sequence at a time.

C. Concept Based Model

Most of text mining techniques are based on word and/or phrase analysis of text. It is important to find term that contributes more semantic meaning to document this concept is known as concept based method. Only the importance of term within document is captured in statistical analysis of term based method. In concept based method the term which contributes to sentence semantic is analysed with respect to its importance at sentence and document levels. The model tries to analyse term at sentence and document level by efficiently finding significant matching term rather than single term analysis. Conceptual term frequency (ctf) to analyse each concept at sentence level is proposed. The 'ctf' is number of occurrences of concept c of sentence's' and 'tf' is term frequency to analyse each concept at document level i.e. number of occurrences of concept 'c' in original document. The concept based Term Analyser algorithm describes how to calculate 'tf' and 'ctf' of matched concept in document [8].

Algorithm: Concept based Term Analyser

Input: Document

Output: Matched concept list L

1. ddoci is a new Document
2. L is an empty List (L is a matching concept list)
3. for each sentence s in d do
4. ci is a new concept in s
5. for each concept ci ∈ {c1, c2, ..., cn} in s do
6. Compute tfi of ci in d
7. Compute ctfi of ci in s in d
8. end for
9. for each dk, where k = {0, 1, .., doci − 1}, ci exist do
10. for each concept cj ∈ {c1, c2, ..., cm} in s do
11. if (ci == cj) then
12. Compute tfweight = avg(tfi, tfj)
13. Compute ctfweight = avg(ctfi, ctfj)
14. add new concept matches to L
15. end if
16. end for
17. end for
18. end for
19. output the matched concepts list L

Algorithm starts with processing of new document which has well defined sentence boundaries. The length of matched concepts and their verb argument structures stored concept based similarity calculation. For each sentence the concepts of the verb argument structures which represent the semantic structures of the sentence are processed sequentially. Each concept in the current document is matched with the other concepts in the previously processed documents. To match the concepts in previous documents is accomplished by keeping a concept list *L* that holds the entry for each of the previous documents that shares a concept with the current document. After the document is processed, *L* contains all the matching concepts between the current document and any previous document that shares at least one concept with the new document. Finally, *L* is output as the list of documents with the matching concepts and the necessary information about them.

The concept-based term analyzer algorithm is capable of matching each concept in a new document (*d*) with all the previously processed documents in *O*(*m*) time, where *m* is the number of concepts in *d*. The concept based similarity between two documents d1 and d2 is calculated by following formula,

$$sim_c(d_1, d_2) = \sum_{i=1}^{m} max(\frac{l_i}{S_{i_1}}, \frac{l_i}{S_{i_2}}) * weight_{i_1} * weight_{i_2}$$

where

$$weight_{i_1} = tfweight_{i_1} + ctfweight_{i_1}$$
$$weight_{i_2} = tfweight_{i_2} + ctfweight_{i_2}$$

The concept-based weight of concept *i*1 in document *d*1 is presented by *weight* $_{i1}$. In calculating *weight* $_{i1}$ the *tfweight* $_{i1}$ value presents the weight of concept *i* in the first document *d*1 at the document-level and the *ctfweight* $_{i1}$ value presents the weight of the concept *i* in the first document *d*1 at the sentence-level based on the contribution of concept *i* to the semantics of the sentences in *d*1. The sum between the two values of *tfweight*$_{i1}$ and *ctfweight* $_{i1}$ presents an accurate measure of the contribution of each concept to the meaning of the sentences and to the topics mentioned in a document. The term *weight* $_{i2}$ is applied to the second document *d*2.

D. Pattern Based Model

Figures In pattern based model document is analysed on pattern basis i.e. pattern of document is formed by analysing is-a-relation between terms to form taxonomy. Taxonomy is tree like structure The pattern based approach can improve the accuracy of system for evaluating term weights because discovered patterns are more specific than whole documents. To generate PTM document split into paragraphs. In pattern taxonomy the nodes represent frequent patterns and their covering sets. The edges are "is-a" relation. Smaller pattern in taxonomy are usually more general because they could be used in both positive and negative documents. Larger patterns in taxonomy are usually more specific since they may be used in positive documents. The semantic information will be used in the pattern taxonomy to improve the performance of using closed patterns in text mining.

The patterns which always co-occur with their parent patterns in the same transaction are redundant patterns and need to be discarded by pruning process. Data mining methods, such as SPM and NSPM, utilize discovered patterns directly without any modification and thus encounter the problem of lacking frequency on specific patterns. The Pattern Deploying Method [1] is proposed with the attempt to address the problem caused by the inappropriate evaluation of patterns, discovered using data mining methods.

Algorithm: PDM (D+, min sup)
Input: List of positive documents, D+; min sup.
Output: Set of vectors, Δ (d-pattern, supports of terms).
Method:

1. Δ←∅
2. foreach document d in D+ do begin
3. extract 1Terms frequent patterns PL from d
4. SP = SPMining(PL, min sup)
5. d←∅
6. foreach pattern p in SP do begin
7. d←d⊕p
8. end for
9. Δ ← Δ ∪ {d}
10. end for

In pattern deploying method frequent sequential patterns are generated by SPMining algorithm at line 4. The main process of pattern deploying occurs between line 6 and line 8 inclusively. The output of this algorithm is a set of vectors. The inputs of the algorithm PDM are a set of positive documents and a pre specified minimum support. In line 4 of this algorithm, a set of sequential patterns is discovered by calling the algorithm SPMining for each document. So far, only positive documents are considered and used in this approach. The use of information from negative documents is another issue with reference to pattern evolution which will be investigated and discussed in the next step in line 6 to 8, each pattern is firstly transferred into an expanded form and then merged into a temporary storage using pattern composition operator. As a result, the deployed pattern (i.e., the set of term weight pairs) for each document is obtained.

In research work, an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The pattern based technique uses two processes pattern deploying and pattern evolving [6]. This technique refines the discovered patterns in text documents. The experimental results show that pattern based model performs better than other data mining-based methods and the concept-based model, but also term-based models.

Table I: Performance comparison of Text mining Methods

Sr No	Method	Feature used to analyse document	Algorithm	F1 Measures
1	Term Based Method	Term	TFIDF	0.321
2	Phrase Based Method	Phrase	SPMining	0.406
3	Concept Based Method	Concept	Concept based Term Analyser	0.448
4	Pattern Based Model	Pattern	PDM	0.493

III. PROPOSED PATTERN MINING METHOD

A. Data Flow Diagram

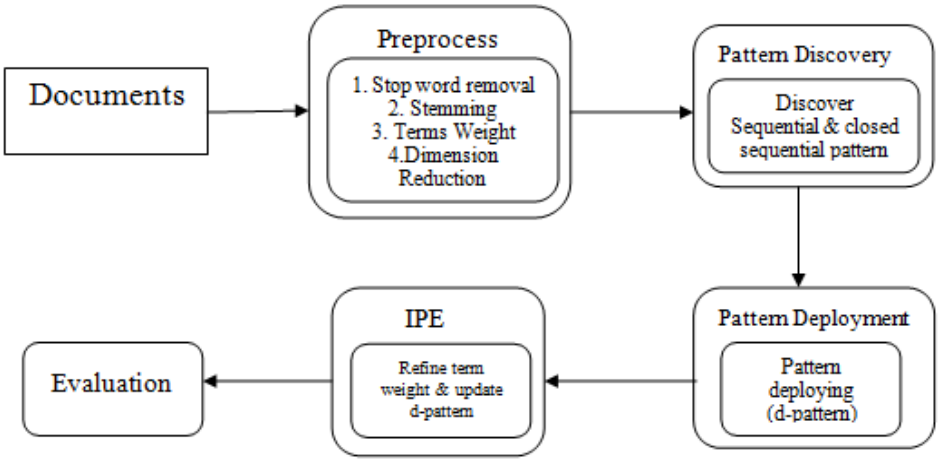

Fig. 1 Data Flow Diagram

Proposed pattern mining method includes modules like data transformation, pattern discovery, pattern deployment, pattern evolution, evaluation. Data flow diagram shows process details which starts with input documents. Here text document is in XML format. System uses medical data set which is divided into training set and test set. Documents in both the sets are either positive or negative. "Positive" means document is relevant to the topic otherwise "negative". System uses sequential closed frequent patterns as well as non sequential closed pattern for finding concept from data set. Each document is divided into paragraph first. Such documents are independently processed for pre-processing step. Pre-processing of XML file include stop word removal and text stemming, and term weight calculation. With minimum support dimension reduction process is done to get feature set. Sequential pattern and closed sequential pattern is discovered by pattern discovery model. Pattern discovery also evaluates specificities of patterns and then evaluates term weights according to the distribution of terms into the discovered patterns. The proposed approach can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents. User interface is reduced as independent processes are executed in multithreaded approach for efficient pattern discovery. A stoppage occurs at dependant processes only. Proposed system implements concepts on the basis of pattern discovery, deployment and evolution. For these three different tests common steps from document pre-processing to pattern discovery are executed.

The first approach of concept formation is from discovery of patterns using sequential closed pattern mining method i.e. is achieved after pattern discovery step and concepts are formed with the help of discovered patterns. Patterns discovered by using SCPM and NSPM. SCPM**:** Finding sequential closed patterns using the algorithm SPMining NSPM**:** Finding non-sequential patterns using the algorithm [6].

The second approach of concept formation is from deployed patterns i.e. after pattern discovery pattern deploying method summarizes discovered pattern using PDM [1]. The d-pattern algorithm is used to discover all patterns in positive documents are composed. After pattern deploying process concepts are formed on the basis of d-patterns.

Third approach is evolution in which inner pattern evolution is done. DPE: Inner pattern evolution is to identify the noisy patterns in documents and to update D-pattern by shuffling. To identify noisy document threshold value is calculated and is used to derive whether document is noisy or not. On the basis of refined pattern weights concept formation is done.

Algorithm: DPEvolving (Ω, D+, D−)
Input: Training set +D & -D, set of d=pattern, experimental coefficient
Output: Set of refined term weight pairs
Method:

1. d←∅
2. τ = Threshold(D+)
3. foreach negative document nd in D− do begin
4. if Threshold({nd}) > τ then
5. Δp = {dp ∈ Ω|termset(dp) ∩ nd = ∅}
6. Shuffling(nd, Δp)
7. end if
8. foreach deployed pattern dp in Ω do begin
9. d ← d ⊕ dp
10. end for
11. end for

Shuffling method removes noisy pattern from d-pattern [1]. Third approach gives more accurate results for concept formation. In all three approaches concept based term analyser algorithm is used for concept formation [8].

IV. RESULTS AND DISCUSSIONS

Results of all three approaches after discovery, deployment and evolution are compared against topic entered. Evaluation is to check system performance on the basis of three metrics like precision, recall and F1 measures. Using these metrics, different methods are compared to check the most appropriate method which gives maximum relevant documents to topic. Comparison of precision, recall and F1 measures for methods Pattern discovery, Pattern deploy and Pattern evolving is as shown in graph 1 fig 2.

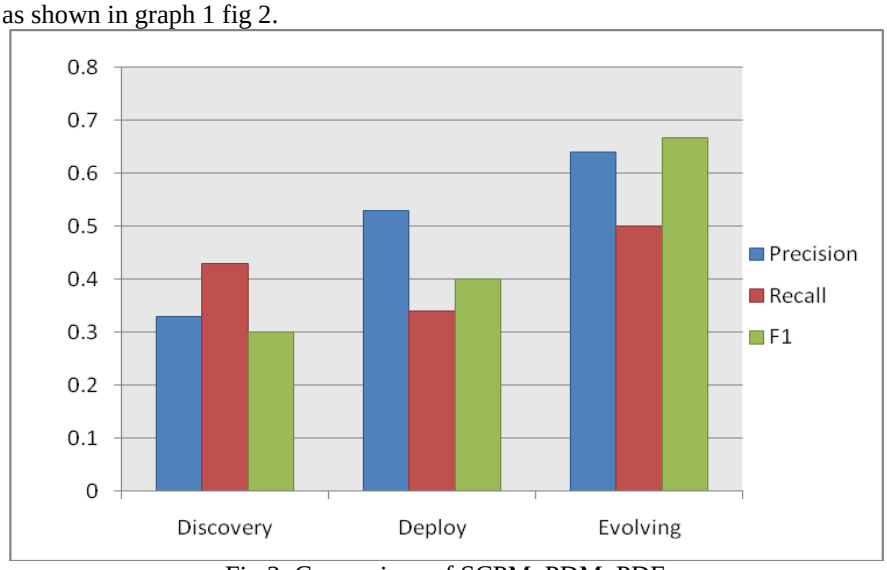

Fig 2: Comparison of SCPM, PDM, PDE

V. CONCLUSIONS

Text mining is the technique that helps users to find useful information from large amount of text documents. Data mining techniques provides pattern mining methods but to use these patterns and update in a way to solve misinterpretation and low frequency problem is achieved in this approach. Depending on processors capacity multiple threads are created to work separately on each process of architecture. Using proposed multithreaded approach stoppages are removed occurring at independent processes. In this paper methods of text mining are discussed out of which pattern based method outperform better than other methods as shown in table I. Patterns discovered by pattern based model are more specific than other method and hence helps to improves accuracy and efficiency. The results show that the implemented system using pattern deploy and pattern Evolving is superior to SCPM data mining-based method.

ACKNOWLEDGMENT

I would like to take this opportunity to acknowledge the contribution of certain people without which it would not have been possible to complete this paper work. I would especially like to thank Mrs. ARCHANA CHAUGULE my internal guide for providing me valuable insights into the subjects that I have chosen. Her constant motivation and enthusiasm helped me a lot. I would like to thank Principal Dr. R.K. JAIN, Head of the department Prof. PRAMOD PATIL, for constantly nourishing the thoughts on how work should be done and how to achieve and tackle all the intricacies. I would also like to thank Prof. JYOTI RAO, ME Coordinator for giving an opportunity to provide a podium and guiding thought this work. Last but not the least I would like to thank all reviewers for helpful comments**.**

REFERENCES

[1] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012

[2] Helena Ahonen, Oskari Heinonen, Mika Klemettinen and A. Inkeri Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections" IEEE 1998

[3] Yuefeng Li and Ning Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs,"IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 4, APRIL 2006

[4] Wai Lam, Miguel Ruiz, and Padmini Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 11, NO. 6. NOVEMBERIDECEMBER 1999

[5] Sheng-Tang Wu, Yuefeng Li, and YueXu, "Deploying Approaches for Pattern Refinement in Text Mining" Proceedings of the Sixth International Conference on Data Mining (ICDM'06) IEEE 2006.

[6] Sheng-Tang Wu Yuefeng Li Yue Xu Binh Pham Phoebe Chen, "Automatic Pattern-Taxonomy Extraction for Web Mining" Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence.

[7] Yuefeng Li, Wanzhong Yang, Yue Xu, "Multi-Tier Granule Mining for Representations of Multidimensional Association Rules" Proceedings of the Sixth International Conference on Data Mining IEEE 2006.

[8] Shady Shehata Fakhri Karray Mohamed Kamel, "Enhancing Text Clustering using Concept-based Mining Model" Proceedings of the Sixth International Conference on Data Mining IEEE 2006.

[9] Shady Shehata, Fakhri Karray, and Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, OCTOBER 2010

[10] Shady Shehata, Fakhri Karray, and Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, OCTOBER 2010

[11] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management:An Int'l J., vol. 24, no. 5, pp. 513-523, 1988.

[12] Andreas Hotho , Andreas N¨urnberger "A Brief Survey of Text Mining" May 13 2005

[13] Kejersti Aas and Line Eikvil, "Text Categorization : A survey" June 1999

[14] Dasa Munkova, Michal Munk and Martin Vozar, "Data Pre-Processing Evaluation for Text Mining:Transaction/Sequence Model" 2013 International Conference on Computational Science

