Effective Pattern Discovery for Text Mining and Compare PDM and PCM

Yeshidagna Tesfaye Assegid [1], Rupali Gangarde[2]

[1]*Mtech student from the department of Computer Science, Symbiosis Institute of Technology Lavale Pune,India*

[2]*Assistant Professor in department of Computer Science, Symbiosis Institute of Technology Lavale Pune,India*

Abstract *— Due to the fast growth of digital data and increase the specific information needs of the users, the data mining task has a vital role to extract the useful information from that large amount of data. The extraction of these data can be achieved using different data mining techniques. The main objective of doing pattern mining is to develop knowledge discovery models for the effective utilize discovered pattern and apply it in area of text mining. In data mining community, most research work focus on developing an effective pattern discovering algorithm which include technique such as sequential pattern mining frequent item mining and close sequential mining for mining useful patterns. But there is a big challenge to discover and update effective pattern. In effective pattern discovery and use techniques there are two main problems. These are:*

- *Low frequency and*
- *Pattern misinterpretation problem*

The general overview of a proposed system is designed to address the problems of low frequency and pattern misinterpretation of pattern discovery method. This system tries to solve the existing approach problems and compare the result generated by pattern deployment and pattern deployment wit pattern co-occurrence methods

Keywords: Data Mining, Information Retrieval, Pattern Taxonomy Model, Text Mining, Association Rule, Sequential Pattern Mining, Close Sequential Pattern Mining, Pattern Deploying, pattern co-occurrence matrix.

I INTRODUCTION

In the past decades, several significant data techniques have been proposed. These techniques include association rule mining, frequent item set mining, sequential pattern mining, closed pattern mining and maximum pattern mining,. Using those pattern mining techniques is not sufficient because effectively using and updating a discovered pattern is still an endless research issue. The main objective of doing pattern mining is to develop knowledge discovery models for the effective utilize discovered pattern and apply it in area of text mining. In Information Retrieval (IR) there are several term based methods. These methods have a good statically properties, because it supports advanced theories for term weight. However term based methods suffered by synonymy, polysemy and homo nym where polysemy means two or more words has the same meaning; and synonymy one word has more than one meaning.

Over the years, phrase based mining approaches hypothesis have been proposed. Phrases could carry more semantics information than term because of that it may perform higher than the term based methods Even phrases are less ambiguous and carry larger information than individual terms, like terms, phrase has its own weakness i.e low frequency.

Like that of terms based methods, patterns enjoy good statistical property and used as an effective alternative to phrases. For solving the problems of phrase based approach, pattern mining method is suggested which uses closed sequential patterns. But the pattern based approach also has two main challenges. These include: pattern misinterpretation and low frequency problem.

II RELATED WORK

Knowledge discovery is the process of extracting important and none trailing formation from large digital data collection. This information may be implicitly present in the dataset or previously unknown potential useful for the users [6, 7].

A number of patterns are extracted from the database. But, all the patterns are not useful. Only those evaluated as interesting and for the user are become knowledge [12]. This depends on the assumer frame of reference defined either by the system itself or the user knowledge.

In general knowledge discovery has the following basic characteristics:

Interestingness: discovered knowledge must be interesting for the intended users and intended application

Accuracy: in knowledge discovery, the discovered pattern depicted the content of the data accurately that state the database

Efficiency: the process of knowledge discovery must be efficient. Especially if the data resource is very large

Understand ability: knowledge discovery can expressed using high level language.
Keyword based approach:
Based on IR (information retrieval), keywords (terms) are used as a representation unit. This representation used collection of words (terms) [1] in the form of attribute value form. Keyword representation has good computational statistical properties. However, the main drawback of key word repetition approach is, while considering single terms, it may suffer from synonyms and polysemy problem: where polysemy word which has more one words share same meaning and polysemy: a word which has more than one meaning. So, the relationship among words cannot be clearly defined and, this leads semantic ambiguity. Documents are classified and ranked based TFIDF classifier [1] [2] algorithms. This algorithm works on the frequency of terms that occurs in the whole document

i. Term Based Approach

Based on IR (information retrieval), keywords (terms) are used as a representation unit. This representation used collection of words (terms) [1] in the form of attribute value form. Keyword representation has good computational statistical properties. However, the main drawback of key word repetition approach is, while considering single terms, it may suffer from synonyms and polysemy problem: where polysemy word which has more one words share same meaning and polysemy: a word which has more than one meaning. So, the relationship among words cannot be clearly defined and, this leads semantic ambiguity. Documents are classified and ranked based TFIDF classifier [1] [2] algorithms. This algorithm works on the frequency of terms that occurs in the whole document.

ii. Phrase Based Approach

Even though, the term based approach has good computational properties, it suffers from different problems, such as: polysemy and synonymy [1],[2],[3] this prone to semantic ambiguity of terms. To overcome these problems, a phrase based approach has been proposed. Phase carry more specific information that terms, for example "search engine" has more specific meaning than "engine". This approach has more specific and clear meaning than single terms. But the phrase based approach that has no significant improvements than term based approach, because this approach has a low frequency, large number of noisy and unneeded phrases among them.

iii. Pattern Based Approach

To overcome the problems of keyword based and phrase based approaches, pattern based methods have been proposed [1] [2][3][4][5].This approaches focus on the pattern based mining, and the advantages of it over term based and phrase based one. As stated in the [1] paper, pattern mining methods use PTM models to classify the sequential pattern into closed sequential pattern and it uses PDM (pattern deployment models) to organize the closed pattern.

III Experimental Dataset

Many standard dataset are available in text data mining, including Reuter's corpus volume1 (RCV1), 20 new groups collection and OHSUMED. But rcv1 is the most popular dataset, which includes 806,791 English news and articles which is prepared by Reuter's journalist in the period between20 August 1996 and 19 August 1997. **B**ecause RCV1 contain the reasonable number of document and it is the latest one; these documents were prepared using structured XML schema. There two groups of topics (100 in total) for RCV1 [5].This is developed and provide by Text Retrieve Conference (TREC) filtering track. The first group includes 50 topics that were composed by human evaluators and the second group also includes 50 topics that were consisted artificially from combination topics. The total amounts of news documents are 8, 00,000. All experimental models use "title" and "text" content of XML documents only. The content in "title" is consider as a paragraph as the one in "text" which consists of from one or more paragraphs. To reduce the dimension of the term, stop word removal is applied and the Porter algorithm [1] is selected for convert term into their root format. When the system extracts the useful pattern, term which one number of frequency has discarded first

IIII Implementation Details

Design

The experiment has two phase, these are the training and testing phase.

a. System Architecture for Training

Figure1 gives training phase system's architecture of proposed system which divides work into modules. This proposed system uses porter, PTM, PCM, and then PDM and D-pattern algorithm. It takes RCV documents and 0.2 minimum supports. Fig. 1 shows system architecture in training phase. Concept vector is generated for all each RCV1 topic by using proposed system algorithms.

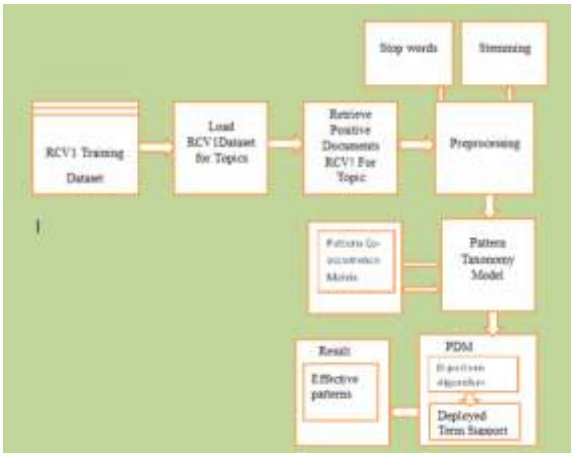

Fig. 1 Training Phase System Architecture

Retrieve and load positive documents:
In this module the system load and retriever all documents list into list that are relevant to the given topic and prepare them for the next module which is called preprocessing

.

Text data preprocessing:
In preprocessing module, each relevant (positive) document is processed using stop words and stemming technique. Stop word is removal of most common words (terms) such as articles, preposition, conjunctions punctuation mark, numbers, adjective, pronouns, adverbs, and verb to be in order to reduce the dimension of terms, and documents term are stemmed to its root format for reducing inflected (derived) terms by using Porter stemming algorithm [9].

Pattern Taxonomy Model (PTM):
It algorithm takes positive preprocessed documents from
the training set as input to PTM, each document is split into
set of paragraphs and each paragraph treated as individual
transaction which consists from collection of terms. PTM generates close sequential patterns using algorithm sp-mining

Pattern Co-occurrence Matrix:
PCM matrix removes ambiguous patterns by finding semantic relationship between them.
Input:
a list of close squential pattern P from positive document $d \in D^{+}$, mininmum support min_sup and Paragraphset $Ps(d) = \{d_{p1}, d_{p2}, d_{p3} \ldots d_{pm}\}$

Output: A pattern co-occurrence matrix, K_{n*n} total Pattern co-occurrence matrix function PCM

1 Let $n = |P|$;
2 Let $m = |PS(d)|$;
3 **for** $i = 1$ **to** n **do**
4 **for** $j = 1$ **to** n **do**
5 Let $A_{ij} = 0$;
6 **for** *pattern* $p_i \in P$ **do**
7 **if** $sup(p_i) \geq min\_sup$ **then**
8 **for** *paragraph* $dp \in PS(d)$ **do**
9 **for** *pattern* $p_j \in P$ **do**
10 **if** p_i *then* p_j *in* dp **then**
11 $A_{i,j} = A_{i,j} + 1$;
 //Count only one for each dp
12 **for** *pattern* $p_i \in P$ **do**
13 $W_R(p_i) = \sum_{j=1}^{n} p_{i,j}$;
14 $W_C(p_i) = \sum_{j=1}^{n} p_{j,i}$;
15 **for** *pattern* $p_i \in P$ **do**
16 $PCM(p_i) = \frac{W_R(p_i) + W_C(p_i)}{n*m}$;

Fig. 2 Pattern Co-occurrence Matrix [2]

Pattern Deployment Method (PDM):

This module is proposed to address the problem caused by the inappropriate evaluation of pattern discovered methods by Patter Taxonomy Model which are utilize discovered patters directly without any modification. These PDM method mainly used to: minimize the computational complexity in case of document evaluation; reduce the size of feature space, deploying specific pattern to emphasis the level of significance and to avoid the low frequency problems, and emphasizing specific pattern to reduce interference from general patters. It also accumulates the weight of terms in the overlap area to estimate the level of significance.

Doc	Pattern Taxonomies	Sequential Pattern
Doc1	PT(1,1)	<dole>12
	PT(1,2)	<democrat>5
	PT(1,3)	<teacher>6
	PT(1,4)	<edu>7
	PT(1,5)	<union>6
	PT(1,6	<school>9
	PT(1,7)	<Clinton>7
Doc2	PT(2,1)	<support>7
	PT(2,2)	<public>14
	PT(2,3)	<privat>4
	PT(2,4)	<school>6
	PT(2,5)	<deuce>4
	PT(2,6)	<poll>4
	PT(2,7)	<improv>4
	PT(2,8)	<percent>12
	PT(2,9)	<nation>3
	PT(2,10)	<support, public>3
	PT(2,11)	<public, school>4
	PT(2,12)	<improve, public>4
Doc3	PT(3,1)	<council>4
	PT(3,2)	<school>7
	PT(3,3)	<york>6
	PT(3,4)	<help>3
	PT(3,5)	<cathol>3
	PT(3,6)	<cit>8
	PT(3,7)	<valley>3
	PT(3,8)	<recen>3
	PT(3,9)	<fight>3

Fig. 3 Set of Positive Document that Consist from Pattern Taxonomies in proposed system

Pattern deployed on common set of terms using pattern deploying

$(df) \Rightarrow <(t_{f1}, n_{f1}, (t_{f2}, n_{f2}), ..,(t_{fm}, n_{fm})>$Where$(t_{f1},N_f)$ = (Term, Total supports from allpatterns)
For Table I the following vectors generated

$(d1) \Rightarrow$= (clinton ,1.0) (democrat ,1.0) (dole , 1.0) (educ ,1.0) (school ,1.0) (teacher ,1.0) (union ,1.0)

$(d2) \Rightarrow$ = (educ ,1.0) (improv ,2.0) (nation ,1.0) (percent ,1.0) (poll ,1.0) (privat ,1.0) (public , 4.0) (school ,2.0) (support , 2.0)

(d 3) $\Rightarrow$(cathol ,1.0) (citi , 1.0) (council , 1.0) (fight , 1.0) (help , 1.0) (receiv ,1.0) (school ,1.0) (vallon ,1.0) (york ,1.0)

Next step is merging patternto generate concept vector using composition operation

d=(educ , 0.6666667) (receiv , 0.33333334) (nation , 0.33333334) (clinton , 0.33333334) (dole , 0.33333334) (union , 0.33333334) (poll , 0.33333334) (improv , 0.6666667) (percent , 0.33333334) (vallon , 0.33333334) (help , 0.33333334) (democrat , 0.33333334) (teacher , 0.33333334) (cathol , 0.33333334) (privat , 0.33333334) (public , 1.3333334) (school , 1.3333334) (citi , 0.33333334) (council , 0.33333334) (york , 0.33333334) (support , 0.6666667) (fight , 0.33333334)

b. System Architecture for Testing Phase

Fig. 5 shows system architecture for testing phase

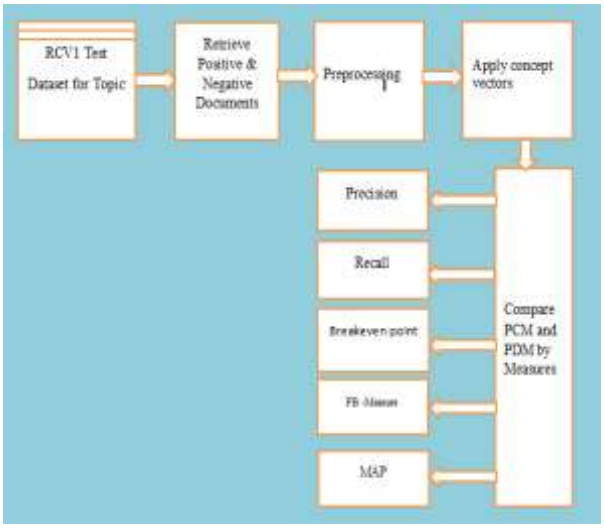

Fig. 5: system architecture for testing phase

- ***c.***
- Retrieving positive and negative documents: in this phase the system retrievers all the positive and negative documents that are relevant to the given topic and gives for preprocessing.
- Preprocessed using stop words and stemming technique. Stop word is removal of most common in order to reduce the dimension of patterns, and documents are stemmed to its root words by Porter stemming algorithm to minimize ambitious words [6].
- Apply Concept Vector: in this phase the system evaluate the term weight and also evaluate the document weight to determine the document status.

IIV performance Measure

Several standard measures are conducted based on precision and recall values. Precision is the proportion retrieved document set that are relevant to the given topic, which expressed as P= (relevant/retrieved) = TP/(TP+FP) and recall is the fraction of relevant documents that were found and expressed using the formula R=(retrieved/relevant) = TP/(TP+FN). Fig. 6 shows precision recall. Where TP is the number of the document the system correctly identify as a positive, FP is the number of document the system falsely identified as positive, FN is the number of relevant documents the system fails to identify. Based on the precision and recall value the system compares the result of PCM and PDM. Precision of first K returned documents top-K is also adopted in this paper. The value of K we use in the experiments is 20. In addition, the breakeven point (b=p) is used to provide another measurement for performance evaluation. It indicates the point where the value of precision equals to the value of recall for a topic. The higher the figure of b=p, the more effective the system is. The b=p measure has been frequently used in common information retrieval evaluations.
In order to assess the effect involving both precision and recall, another criterion that can be used for experimental

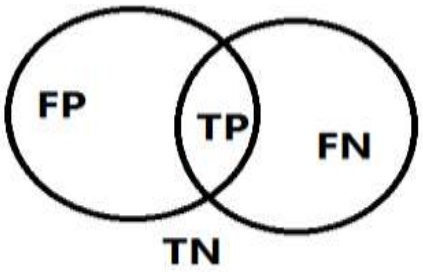

Figure 6: Relationship Between Recall and Precision

V. Experimental Result

In the last section we present the final experiment result which is returned by the proposed approach. In the proposed approach we compare the result getting by PTM (PDM) and PTM ((PCM) (PDM)). The result which discovered by the

approach is compared using the precision standard value. The overall compares on results presented in Fig. 7 Based on the precision value we get the following result.

Topic	Precision After PDM	Precison After PCM
R101	0.5320624	0.5650485
R102	0.5340136	0.56985295
R103	0.08888889	0.09570957
R104	0.48727274	0.48540145
R105	0.1953125	0.21276596
R106	0.1009772	0.10231023
R107	0.06779661	0.08405797
R108	0.04310345	0.048076924
R109	0.33854394	0.33617023
R110	0.052953158	0.053168735

Figure 7: the comparison result between PTM (PDM) and PTM ((PCM)(PDM))

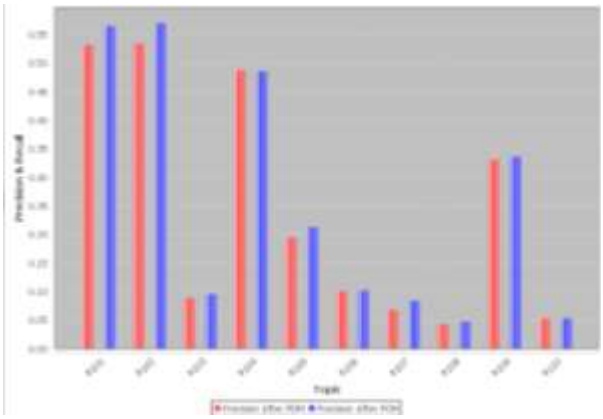

Fig. 8 Bar graph for comparison of PDM and PTM ((PCM)(PDM))

VI. CONCLUSION

Several data mining techniques have been proposed to discover effective patterns but suffered from pattern misinterpretation and low frequency problems. To overcome these problems, the proposed system use PDM for pattern deploying and pattern and Co-occurrence matrix to clean close sequential pattern in the pattern taxonomy model. Those methods increase the performance of a text mining process.

This paper focuses on research title effective pattern discovery for text mining and comparing PDM and PCM. RCV1 dataset used to conduct an experiment. The experiment has two phases: the training phase and testing. In training phase, we prove how to discover and use in text mining whereas in testing phase. We test the performance of method. Based on the test result we conclude that PTM ((PCM) (PDM)) improve the performance of the system and we get more efficient result.

VII. ACKNOWLEDGEMENT

I would like to express my special gratitude and thanks to Assistance Professor Rupali madam, my guide for all her guidance and encouragement throughout the research work. I would also like to thank you for my examiners for their wonderful previous comment and suggestion.

Special thank must go to symbiosis institute of technology which has provided me comfortable research environment with required infrastructure and support.

Many thanks also go to my respected family. This research work would not have been accomplished without the constant support of my family. I would like to dedicated this research to my lovely uncle shamble Menker Woldemeskel for his never ending encouragement for last two years

Finally, I would like to extend a heartfelt gratitude to the Aksum University, Ethiopian government, as well as Ethiopian people's for their ultimate assistance and support.

References

[1] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," Knowledge and Data Engineering, IEEE Transactions on, vol. 24, no. 1, pp. 30–44, 2012

[2] M. Albathan, Y. Li, and A. Algarni, "Using patterns co-occurrence matrix for cleaning closed sequential patterns for text mining," in Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01, pp. 201–205, IEEE Computer Society, 2012.

[3] S.-T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in Data Mining, 2006. ICDM'06.Sixth International Conference on, pp. 1157–1161, IEEE, 2006.

[4] L. Pipanmaekaporn, "Feature discovery in relevance feedback using pattern mining," in Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on, pp. 301–307, IEEE, 2013.

[5] Y. Li, A. Algarni, and N. Zhong, "Mining positive and negative patternsfor relevance feature discovery," in Proceedings of the 16th ACM SIGKDD international

conference on Knowledge discovery and data mining, pp. 753–762, ACM, 2010

[6] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic patterntaxonomy extraction for web mining," in Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on, pp. 242–248, IEEE, 2004

[7] R. Sharma and S. Raman, "Phrase-Based Text Representation for Managing the Web Document," Proc. Int'l Conf. Information Technology: Computers and Comm. (ITCC), pp. 165-169, 2003

[8] S. Wu, "Knowledge discovery using pattern taxonomy model in text mining," 2007.

[9]M.F,Porter,.An algorithm for suffix stripping. *Program*, *14*(3), pp.130-137.1980