Using Patterns Co-occurrence Matrix for Cleaning Closed Sequential Patterns for Text Mining

Mubarak Albathan
Science and Engineering Faculty
Queensland University of Technology
Brisbane, Australia
mubarak.albathan@student.qut.edu.au

Yuefeng Li
Science and Engineering Faculty
Queensland University of Technology
Brisbane, Australia
y2.li@qut.edu.au

Abdulmohsen Algarni
Science and Engineering Faculty
Queensland University of Technology
Brisbane, Australia
a1.algarni@qut.edu.au

***Abstract*—With the overwhelming increase in the amount of texts on the web, it is almost impossible for people to keep abreast of up-to-date information. Text mining is a process by which interesting information is derived from text through the discovery of patterns and trends. Text mining algorithms are used to guarantee the quality of extracted knowledge. However, the extracted patterns using text or data mining algorithms or methods leads to noisy patterns and inconsistency. Thus, different challenges arise, such as the question of how to understand these patterns, whether the model that has been used is suitable, and if all the patterns that have been extracted are relevant. Furthermore, the research raises the question of how to give a correct weight to the extracted knowledge. To address these issues, this paper presents a text post-processing method, which uses a pattern co-occurrence matrix to find the relation between extracted patterns in order to reduce noisy patterns. The main objective of this paper is not only reducing the number of closed sequential patterns, but also improving the performance of pattern mining as well. The experimental results on Reuters Corpus Volume 1 data collection and TREC filtering topics show that the proposed method is promising.**

***Keywords*-Pattern co-occurrence matrix; Text mining; Closed Sequential pattern; Information retrieval**

I. Introduction

With the explosive growth of information sources available on the Web, search engines return large numbers of documents based on a keyword-matching approach, but most of the results are not relevant to what the user needs. It is becoming essential to provide users with tools that more effectively filter huge amounts of streamed text data in order to find accurate matches more quickly.

Various studies have been conducted in the area of pattern-based approaches, such as pattern discovery and relevance feedback, aiming to improve the retrieval of useful information needed by users. Pattern discovery is one type of data mining, and it is used as an effective technique for knowledge discovery in many applications. The key advantage of pattern discovery is that it can implicitly identify interesting patterns from given data without domain knowledge. An example of pattern discovery is the pattern taxonomy model, which discovers closed sequential patterns [1]. A closed sequential pattern aims to reduce redundant and noisy patterns in extracted frequent patterns.

Unlike normal keyword-based approaches, frequent patterns have different benefits for text mining, such as carrying more semantic information and being easy to obtain using a pattern-mining algorithm. Furthermore, experimental results have demonstrated encouraging improvements in the effectiveness of pattern-based models in comparison with keyword-based models. Despite these benefits, many patterns generated from text collection contain redundant, inconsistent and noisy information [2].

To filter the overwhelming output produced by pattern-based models, several approaches for text post-processing have been proposed that aim to improve the efficiency and quality of extracted knowledge by reducing the amount of extracted information. For example, some approaches selected k sets of the frequent sets [3] or summarised the collection of closed patterns [4]. Despite this, these methods continue to suffer from noisy and low-quality output.

In this paper, we propose a new method for pattern post-processing, which uses a pattern co-occurrence matrix to evaluate closed patterns and select a small set of closed patterns in order to improve the performance of pattern mining. The Pattern Co-occurrence Matrix (PCM) captures the relations between closed patterns based on their appearance in text paragraphs. The experimental results illustrate that the proposed method is promising.

The rest of this paper will be structured as follows. Section II presents a detailed overview of the related work, and different concepts of patterns will be introduced in section III. Section IV discusses the PCM model and how to calculate patterns and terms' co-occurrence weight. Following this is the discussion on the experiment's setting and results. Finally, section VI presents the conclusion of this paper.

II. Related Work

With the growing volume of published research and documents on the web, and therefore the underlying knowledge in these texts, text mining and Information Retrieval (IR) are aimed to assist researchers in extracting useful knowledge from a collection of text and improve search engines.

Closed patterns is one of the presence methods to be an alternative to phrases [5] because patterns enjoy good statistical properties, like terms. To avoid the disadvantages of using a phrase-based model, pattern-based models, such as Pattern Taxonomy Model (PTM) [5], have been developed, which use the concepts of closed sequential patterns and pruned nonclosed patterns that have shown effective improvement. Despite this, noisy and duplicated patterns still occur due to the data mining technique processes that occur to extract patterns.

Thus, to overcome this issue, various studies have used the co-occurrence matrix technique in text-mining pplications ranging from speech recognition and parse selection to IR [6]. The generic POPC algorithm [7] is a clustering approach, which calculates the co-occurrence matrix between patterns that are used as a similarity matrix. In another study, Weeds and Weir [6] tried to create a framework for lexical distributional similarity; this is called co-occurrence retrieval. This framework creates a co-occurrence retrieval matrix, which is a similarity matrix, to find relationships between words that might be found in a thesaurus, such as *synonymy*, *antonymy* and *hyponymy*.

For noisy and inconsistent extracted data, it is still an issue how to extract patterns and give them an accurate weight. The Co-occurrence matrix has been chosen to find and identify the importance of, and relation among, extracted closed sequential patterns in this paper. Based on these relations (patterns co-occurrence matrix), we can identify the important patterns among closed patterns and assign a suitable weight for the extracted patterns and their terms.

III. Pattern Definitions

Different pattern mining methodologies discover closed sequential patterns from frequent sequential patterns, such as Pattern Taxonomy Model (PTM) [5]. This model aims to find useful features, such as patterns, terms and their weight, from a training set *D*, which consists of positive documents D^+ and negative documents D^-, where each document *d* is represented as a set of paragraphs $PS(d)$. To clearly understand the concepts of patterns, we present the concepts of frequent patterns, closed patterns and closed sequential patterns in this section.

A. Frequent and Closed Patterns

Let $T = \{t_1, t_2, ..., t_n\}$ be a set of terms extracted from positive documents D^+. Given a *termset X*, a set of terms, in document *d*, *coverset*(*X*) is used to denote the covering set of *X* for *d*, which includes all paragraphs *dp*∈*PS*(*d*) such that *X*⊆*dp*, and its absolute support ($supp_a$) is the number of occurrences of X in *PS*(*d*), that is:

$$supp_a(X) \quad = \quad |coverset(X)| \tag{1}$$

Moreover, its relative support ($supp_r$) is the fraction of the paragraphs that contain the pattern, that is:

$$supp_r(X) \quad = \quad \frac{supp_a(X)}{|PS(d)|} \tag{2}$$

Therefore, *termset X* is called a *frequent pattern* if its $supp_a(X)$ or $supp_r(X)$ is greater than or equal to a minimum support (min_sup) [1].

On the other hand, given a set of paragraphs *Y* ⊆ *PS(d)*, we can define its *termset*, which satisfies:

$$termset(Y) = \{t | \forall dp \in Y \Longrightarrow t \in dp\} \tag{3}$$

and the closure of *X* is defined as

$$Cls(X) \quad = termset(coverset(X)) \tag{4}$$

Therefore, a pattern *X* is called closed if and only if *X* = *Cls(X)* [10].

B. Closed Sequential Pattern

The sequential pattern *X* is called a frequent pattern if its $supp_r(X) \geqslant min\_sup$. A frequent sequential pattern X is called a closed sequential pattern if there exists no frequent sequential pattern Y, such that $X \sqsubset Y$ and $supp_a(X) = supp_a(Y)$ [1], where the relation $\sqsubset$ represents the strict part of subsequence relation $\sqsubseteq$.

IV. Pattern Co-occurrence Matrix

Text co-occurrence matrices, such as co-citation, co-word and co-link matrices, can define concepts that occur within the same term in text [11], which provide us with useful information for understanding the structures of documents.

Not all extracted patterns are useful because extracted patterns usually contain noisy patterns and inconsistencies due to the different data mining processes that are used for extracting these patterns. It is clear that there are relationships between patterns in documents based on their appearances in paragraphs. The co-occurrence matrix method attempts to identify the semantic relationships between these patterns and identify the important relationships between them.

In this paper, the co-occurrence matrix has been chosen to study the Pattern Co-occurrence Matrix (PCM) in a document to find the relationship between patterns and identify the important relationships between them. Therefore, we can define the co-occurrence matrix in our research as *a matrix that is defined over a document to describe the co-occurrence relation between patterns*. For example, let A be the *n*n* pattern co-occurrence matrix, while the element A_{ij} is the number of times that the pattern A_j occurred after pattern A_i in the paragraphs of the document.

As mentioned earlier, closed sequential patterns are extracted from documents based on their support and confidence, while in this research we attempt to re-evaluate the extracted patterns based on the pattern co-occurrence matrix in order to reduce the noisy patterns.

A. Calculating the Pattern Co-occurrence Matrix (PCM)

This research applies the PCM on top of the closed sequential patterns, and tries to remove the noisy patterns which is in this experiment the patterns that have no relation with other patterns. Let $P = \{p_1, p_2, \ldots, p_n\}$ be a set of extracted closed sequential patterns with a *min_sup* (e.g. $min\_sup = 0.2$ in PTM) from all paragraphs *dp*∈*PS*(*d*) in document $d \in D^+$, where *PS*(*d*)= $\{dp_1, dp_2, \ldots, dp_m\}$.

$$A_{n*n} = \begin{bmatrix} & p_1 & p_2 & \ldots & p_j & \ldots & p_n \\ p_1 & A_{1,1} & A_{1,2} & \ldots & A_{1,j} & \ldots & A_{1,n} \\ p_2 & A_{2,1} & A_{2,2} & \ldots & A_{2,j} & \ldots & A_{2,n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ p_i & A_{i,1} & A_{i,2} & \ldots & A_{i,j} & \ldots & A_{i,n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ p_n & A_{n,1} & A_{n,2} & \ldots & A_{n,j} & \ldots & A_{n,n} \end{bmatrix}$$

As shown in matrix A_{n*n}, the pattern co-occurrence matrix *A* with size *n***n*, where $n = |P|$, is the number of extracted patterns and $A_{i,j}$ (read $p_i \rightarrow p_j$) is the number of co-occurrences of patterns p_j occur after p_i.

To calculate the co-occurrence of any two patterns in the matrix, such as patterns $A_{i,j}$, we run over all the document paragraphs *PS*(*d*), where the two patterns should be in the same paragraph and in the same order (p_j occur after p_i). In addition, the occurrence is only calculated once for each of the two patterns in each paragraph. Finally, to calculate the total co-occurrence of pattern p_i in document *d*, we first calculate the total co-occurrence $W_R(p_i)$ for row and $W_C(p_i)$ for column as follows:

$$W_R(p_i) = \sum_{j=1}^{n} A_{i,j} \quad , \quad W_C(p_i) = \sum_{j=1}^{n} A_{j,i}$$

And the total co-occurrence for pattern p_i will be:

$$PCM(p_i) = W_R(p_i) + W_C(p_i) \tag{5}$$

Finally, considering the length of the documents, we normalize the the total co-occurrence of a pattern as follows:

$$PCM(p_i) = \frac{W_R(p_i) + W_C(p_i)}{n * m} \tag{6}$$

where m is the number of paragraphs in the document.

For an example of calculating the pattern co-occurrence matrix as describes in the PCM, algorithm 1 describes the procedure for calculating the pattern co-occurrence matrix, the row co-occurrence, the column co-occurrence and the total co-occurrence PCM. It starts to initialize the pattern co-occurrence matrix $A_{n*n} = (0)$ (step 3 to step 5). It then calculates the value for each elopement $A_{i,j}$ (step 6 to step 11). Finally, it works out the co-occurrences (step 12 to step 16).

Algorithm 1: Calculating the Pattern Co-occurrence Matrix PCM

Input : A list of Closed Sequential Patterns P from document $d \in D^+$, Minimum Support; *min_sup*, and *PS(d)*= $\{dp_1, dp_2, \ldots, dp_m\}$.

Output: A pattern co-occurrence matrix, A_{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 ∈ P do
7      if sup(p_i) ≥ min_sup then
8          for paragraph dp ∈ PS(d) do
9              for pattern p_j ∈ 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 ∈ P do
13     W_R(p_i) = Σ_{j=1}^{n} p_{i,j};
14     W_C(p_i) = Σ_{j=1}^{n} p_{j,i};
15 for pattern p_i ∈ P do
16     PCM(p_i) = (W_R(p_i) + W_C(p_i)) / (n*m);
```

V. Evaluation

The main objectives of this research is to extract high quality patterns from text documents by introducing a method for weighting patterns, using the pattern co-occurrence matrix and cleaning the closed sequential patterns based on the pattern co-occurrence matrix. To support this idea, this section will illustrate the experiment environment, including the dataset that have been used, the baseline models and the results and discussion of the experiment results:

A. Data

In order to conduct the experiment, the Reuters Corpus Volume 1 (RCV1) will be used. These 100 topics from English language news stories produced by Reuters journalists between August 20, 1996 and August 19, 1997, comprise a total of 806,791 documents [10]. The format of these documents was structured in XML format. The first 50 topics were developed by human and the rest by intersections of pairs of Reuters categories, which divided into two sets: training and testing sets. Both of these sets consist of positive (relevant) and negative (irrelevant) documents [12].

Before applying the co-occurrence matrix, different operations have been conducted on the data, such as preprocessing

the documents and removing a given stop-words list. Also, the terms have been stemming by applying the Porter stemmed algorithm [13] for suffix stripping.

B. Baseline Models

Five baseline models have been used with $min\_sup = 0.2$ to reduce the number of extracted patterns with lower relative support. The first four are frequent patterns (Freq Patterns), frequent closed patterns (Fre Closed Ptns), sequential patterns (Seq Patterns) and closed sequential patterns (Closed Seq Ptns) [1].

The last one is the *n*-Gram model: *n* represent the length of sequence *S*, which indicates the number of words contained in *S*. In this paper, the length of sequence *n* is 3, i.e., 3-grams.

C. Evaluation Methods

In this study, we have a collection of documents and every document is known to be either relevant or irrelevant to the topic.To evaluate the effectiveness of this study, different means have been used, specifically precision *p* , the average precision of the *top 20* return documents, the $F_1 - score$ measure, and the *break-even point (b/p)*. Also, to evaluate the whole system, *interpolated Precision on 11-points* is used for comparison of the performance of different systems by averaging precisions at 11 standard recall levels which called Interpolated Average Precision (IAP). Moreover, Mean Average Precision (MAP) is used which is the average of precision of all experiment topics. These evaluation metrics are widely used in information retrieval research (for more information about these measures see [14]).

D. Results and Discussion

Closed sequential patterns are good alternatives for phrases (n-grams); however, they still struggle with some noisy and inconsistent patterns due to the common data mining process for extracting these patterns [1]. Moreover, support and confidence are not suitable to answer what users need. The Pattern Co-occurrence Matrix model (PCM) introduces a new way to weight and clean patterns. Thus, the PCM process consists of two main stages: weighting patterns using the co-occurrence matrix and cleaning closed sequential patterns based on the pattern co-occurrence matrix weights.

1) pattern co-occurrence matrix weight: Introducing PCM to find the co-occurrence relation between patterns helps to identify the important patterns and improve the efficiency of closed sequential patterns. Table I shows an excellent improvement in PCM model comparing with *n*-Gram and other patterns-based models. All the baseline models mentioned in this paper use support as the weighting technique, while this experiment uses the pattern co-occurrence matrix as the weighting technique for patterns. The results of this experiment show a significant improvement in all five measure factors over the 50 topics. It shows that PCM has 8% maximum and 2% minimum percentage changes on average for all measures when compared with the best pattern-based model closed sequential patterns (Closed Seq Ptns), even if we run the PCM to clean the patterns only, and use the support to weight the patterns instead of the co-occurrence weight, as shown in Table I. Furthermore, Figure 1 illustrates the improvement in performance between the PCM, the closed sequential patterns, and a PCM model that is used only to clean the patterns and using support weight, which shows that the result is similar to the closed sequential patterns.

Table I
COMPARISON OF ALL PATTERN (PHRASE) BASED METHODS ON 50 TOPICS

Method	*top-20*	*b/p*	*MAP*	$F_{\beta=1}$	*IAP*
PCM	**0.437**	**0.371**	**0.381**	**0.397**	**0.406**
Seq Patterns	0.401	0.343	0.361	0.385	0.384
Closed Seq Ptns	0.406	0.353	0.364	0.390	0.392
Freq Patterns	0.412	0.352	0.361	0.386	0.384
Freq Closed Ptns	0.428	0.346	0.361	0.385	0.387
n-Gram	0.401	0.342	0.361	0.386	0.384
PCM using support weight	0.382	0.341	0.343	0.374	0.371
$\%chagnge$	**+8%**	**+5%**	**+6%**	**+2%**	**+4%**

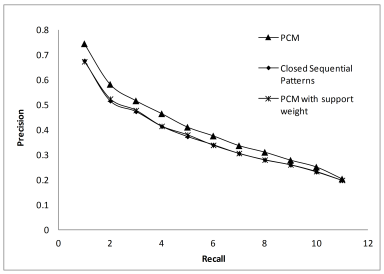

Figure 1. Average of Closed Sequential Patterns and PCM 11-point ($min\_sup = 0.2$)

In summary, the experimental results in this section show that using the pattern co-occurrence matrix is more suitable for weighting patterns than using support and confidence for weighting patterns. We will show in the next section that PCM is also suitable for cleaning documents.

2) Cleaning the Closed Sequential Patterns: Usually, long patterns are more important than short patterns, as proved in the Relevance Feature Discovery model (RFD) [10]. To extract long patterns from text documents using data mining methods, we have to use a very small minimum support (e.g., $min\_sup = 0.2$). However, low min_sup would generate a large number of patterns and most of them would be noise patterns. The closed pattern technique is one of the pruning methods that is used to remove some of the redundant and noisy patterns. To further reduce the number of noisy patterns, the PCM model studies the relationship among sequential closed patterns based on the co-occurrence matrix. As illustrated in Figure 2, if the min_sup is low, the

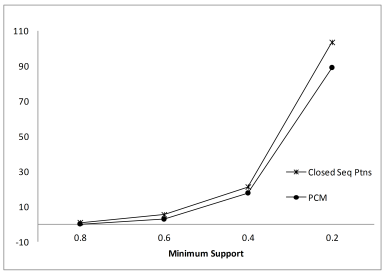

Figure 2. Average number of patterns in PCM and Closed Sequential Patterns

precision will be high; however, a large number of patterns will be generated, including some noisy patterns. This also shows that noisy patterns can be successfully reduced using PCM model.

Table II
COMPARESION OF NUMBER OF PATTERNS

Method	# extracted Patterns	% noisy patterns
Frequent Patterns	290.18	59%
Closed Seq Patterns	99.16	10%
PCM	**89.04**	

We observed that some patterns have no relationship with others, in other words; they have 0 co-occurrence ($PCM(p_i) = 0$). The patterns that have no relationship with others patterns will be considered as noise patterns. Those noisy patterns can be deleted from the patterns list. Table II shows the average number of extracted patterns for the PCM model, frequent patterns and closed sequential patterns patterns in the 50 topics. Closed sequential patterns clean about 59% of the redundant patterns in frequent patterns. Moreover, PCM model can further clean about 10% of the the closed sequential patterns.

Furthermore, in this experiment we observed the first 10 topics ($min\_sup = 0.2$), we found that the number of extracted patterns between the PCM and the closed sequential patterns is different. Some topics in closed sequential patterns have a large number of suspected noisy patterns, which between 50% and 2% of extracted patterns. Thus, the PCM model provides a promising method to significantly reduce the number of noisy patterns in the extracted patterns.

In summary, compared with other good pattern-based models, the PCM model has an excellent performance (see Table I). The PCM model can also identify patterns that have no relation with others using the co-occurrence matrix. It can then can largely remove noisy patterns (see Table II).

VI. CONCLUSION

This paper presents a new method to clean closed sequential patterns for text mining. It first identifies the relationship between extracted patterns by calculating the pattern co-occurrence matrix in the document. It also uses the matrix to re-evaluate closed sequential patterns in order to remove noisy patterns.

The proposed method is also tested in a standard data collection (RCV1) for 50 TREC topics and compared with five up-to-date baseline models on RCV1 and the 50 TREC topics. The experimental results show that the proposed method can significantly reduce noisy patterns in the extracted closed sequential patterns (10% reduced). It also shows that the proposed method can significantly improve the performance of pattern mining (the average percentage change is 5% for five measures).

REFERENCES

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

[2] 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*, dec. 2006, pp. 1157 –1161.

[3] F. Afrati, A. Gionis, and H. Mannila, "Approximating a collection of frequent sets," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*, 2004, pp. 12–19.

[4] X. Yan, H. Cheng, J. Han, and D. Xin, "Summarizing itemset patterns: a profile-based approach," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 2005, pp. 314–323.

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

[6] J. Weeds and D. Weir, "Co-occurrence retrieval: A flexible framework for lexical distributional similarity," *Computational Linguistics*, vol. 31, no. 4, pp. 439–475, 2005.

[7] T. Morzy, M. Wojciechowski, and M. Zakrzewicz, "Scalable hierarchical clustering method for sequences of categorical values," *Advances in Knowledge Discovery and Data Mining*, pp. 282–293, 2001.

[8] K. Wakabayashi and T. Miura, "Topics identification based on event sequence using co-occurrence words," *Natural Language and Information Systems*, pp. 219–225, 2008.

[9] O. Madani and J. Yu, "Discovery of numerous specific topics via term co-occurrence analysis," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 1841–1844.

[10] Y. Li, A. Algarni, and N. Zhong, "Mining positive and negative patterns for relevance feature discovery," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 753–762.

[11] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[12] T. Rose, M. Stevenson, and M. Whitehead, "The reuters corpus volume 1-from yesterdays news to tomorrows language resources," in *Proceedings of the Third International Conference on Language Resources and Evaluation*, 2002, pp. 29–31.

[13] M. Porter *et al.*, "An algorithm for suffix stripping," 1980.

[14] B. Liu, *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Verlag, 2007.