Discovering Compatible Top-K Theme Patterns from Text Based on Users' Preferences

Yongxin Tong[1], Shilong Ma[1], Dan Yu[1], Yuanyuan Zhang[2], Li Zhao[1], and Ke Xu[1]

[1] State Key Lab. of Software Development Environment
Beihang University, Beijing 100191, China
{yxtong,slma,yudan,lzh,kexu}@nlsde.buaa.edu.cn
[2] China Academy of Telecommunication Technology, Beijing 100191, China
yyzhang@catt.ac.cn

Abstract. Discovering a representative set of theme patterns from a large amount of text for interpreting their meaning has always been concerned by researches of both data mining and information retrieval. Recent studies of theme pattern mining have paid close attention to the problem of discovering a set of compatible top-k theme patterns with both high-interestingness and low-redundancy. Since different users have different preferences on interestingness and redundancy, how to measure the attributes of the users' preferences, and thereby to discover "preferred compatible top-k theme patterns" (PCTTP) is urgent in the field of text mining. In this paper, a novel strategy of discovering PCTTP based on users' preferences in text mining is proposed. Firstly, an evaluation function of the preferred compatibility between every two theme patterns is presented. Then the preferred compatibilities are archived into a data structure called theme compatibility graph, and a problem called MWSP based on the compatibility graph is proposed to formulate the problem how to discover the PCTTP. Secondly, since MWSP is proved to be a NP-Hard problem, a greedy algorithm, DPCTG, is designed to approximate the optimal solution of MWSP. Thirdly, a quality evaluation model is introduced to measure the compatibility of discovering theme patterns. Empirical studies indicate that a high quality set of PCTTP on four different sub text sets can be obtained from DBLP.

1 Introduction

Usually a large amount of text information is encountered in the application of text processing. Thus, how to discover a representative set of theme patterns automatically from a large amount of text to interpret their meaning is still concerned by researches of both data mining and information retrieval. Since recent researches in the domain of data mining show that a set of closed frequent patterns have many overlaps and redundant patterns, it is difficult for users to understand the meaning of huge number of patterns directly. The traditional frequent theme pattern mining from texts also confronts the same difficulties. Therefore, we should discover a small scale of theme patterns, with high-interestingness and low-redundancy simultaneously, from much text information. However, different users have different preferences on

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 130–142, 2009.

interestingness and redundancy, so it is hard to use one unified criterion to find patterns with both high-interestingness and low-redundancy. We give an example to make a further explanation that discovering "preferred compatible top-k theme patterns" (*PCTTP* for short) from a large amount of text is the urgent and interesting challenge in text mining post-processing.

An interesting example is to discover the hot research topics by analyzing the titles of papers in DBLP database (The DBLP is a well known and very popular search engine which is used to index papers in the computer science area.) Given the periods and the ranges of some conferences and journals, we can discover PPCTP with high-interestingness and few-overlap research topics based on users' preferences. For example, a researcher, especially a beginner, often wants to know what are the hottest research topics lately, that is, likely to pay more attention to the level of popularity, even there may be some overlap of terms in these topics. However, other researchers want to obtain many different topics even not the hottest ones, because the overlap among these topics is very low. Hence, discovering PCTTP from a bulk of titles of papers can precisely satisfy these requirements from different users.

It is important to discover PCTTP from the above example, however, mining compatible top-k theme patterns based on users' preferences has not been well addressed. Although recent studies of text pattern mining have mined redundancy-aware top-k patterns [8] and excluded the redundancy among theme patterns [1, 7, 9], they could not mine a set of patterns according to users' preferences of interestingness and redundancy. A detailed discussion of the related work is given in Section 6.

The rest of the paper is organized as follows. Section 2 introduces the problem formulation, including the evaluation function of preferred compatibility of every theme pattern, theme compatibility graph and MWSP (Maximal Weight Sum Problem). The DPCTG (Discovering Preferred Compatible Theme based Greedy) algorithm is proposed to approximate the MWSP in Section 3. A quality evaluation function is introduced in Section 4. The experimental results, the related work and the conclusion are given in Section 5, 6 and 7 respectively.

2 Problem Formulation

In this section, we firstly introduce some basic concepts of interestingness and redundancy of theme pattern, then, define the problem of discovering PCTTP.

2.1 Preliminaries

Given a sequence database D, $D = \{s_1, s_2, ..., s_n\}$ is a set of sequences. Each sequence is associated with an ID. The symbol $|D|$ represents the number of sequences in D. We define α is a sequential pattern. The support of a sequence sup(α) in D is the number of sequences in D which contains α. Given a minimum support threshold, denoted as *min_sup*, the set of **frequent sequential patterns** is a set of all the sequences whose support is no less than *min_sup*. The set of **closed frequent sequential pattern** is a set of sequences which have no *super-sequence* with the same support.

According to the frequent sequential patterns and the closed frequent sequential patterns, a text collection C is recognized as the above sequence database D, each sentence in C as a sequence in D, and the theme pattern and the closed theme pattern is defined as follows.

Definition 1 (Theme Pattern). Given a text collection C equal to a sequence database, a theme pattern is a frequent sequential pattern in C.

Definition 2 (Closed Theme Pattern). Given a text collection C equal to a sequence database, a closed theme pattern is a closed frequent sequential pattern in C.

Definition 3 (Theme Pattern Interestingness)[8]. Given a set of patterns P, there is a function which maps any pattern $p \in P$ to a real value and is used to measure interestingness about pattern p, denoted as *I(p)*. In this paper, the interestingness of theme pattern is weighted by a *tf-idf* scoring function, denoted as follows:

$$I(p) = \sum_{i=1}^{t} \frac{1+\ln(1+\ln(tf_i))}{(1-s)+s\dfrac{dl}{avdl}} \times \ln\frac{N+1}{df_i} \tag{1}$$

where tf_i equals the support of the pattern p, df_i is the inverse sentence frequency of a word, dl is the average sentence length associated with P, $avdl$ is the overall average sentence length, N is the number of sentences in the text collection, and s is an empirical parameter (usually 0.20).

Definition 4 (Jaccard Distance)[7]. A distance measure $Dis : P \times P \to [0,1]$ is a function mapping two patterns $p_m, p_n \in P$ to a value in [0,1]. In this paper, we use a Jaccard Distance to measure the pattern distance between p_m and p_n:

$$Dis(p_m, p_n) = 1 - |D_m \cap D_n| / |D_m \cup D_n| \tag{2}$$

Since the ideal redundancy measure $R(p_m, p_n)$ of any two theme patterns is generally difficult to obtain, we use the above Jaccard Distance to approximate the redundancy.

Definition 5 (Theme Pattern Redundancy). Given a set of theme patterns P, there is a function which maps any two theme pattern $p_m, p_n \in P$ to a real value and is used to measure redundancy between any two pattern p_m, p_n, denoted as $R(p_m, p_n)$

$$R(p_m, p_n) = 1/D(p_m, p_n) \tag{3}$$

According to definition (4) and (5), the redundancy ∞ (Jaccard Distance *0*) means two patterns are completely relevant, and redundancy *1* (Jaccard Distance *1*) means two patterns are completely independent.

2.2 Function of Preferred Compatibility

In this subsection, we describe how to measure the compatibility between interestingness and redundancy of theme patterns based on users' preferences (In what follows, we brief it as the preferred compatibility). In this paper, for simplicity of the problem, we assume that the users' preferences are evaluated by two categories: interestingness and redundancy.

Definition 6 (Function of Preferred Compatibility between Two Theme Patterns). Given a theme pattern set P, $I(x)$ measures the interestingness of a pattern belong to the set P. $R(x)$ measures the redundancy of every two patterns. l represents a proportion value between redundancy and interestingness in the users' preferences. $C(p_m, p_n)$, an evaluation function, denotes the value of the preferred compatibility between two patterns, which maps any two patterns $p_m, p_n \in P$ to a real value, shown as:

$$C(p_m, p_n) = I(p_m) + I(p_n) / R(p_m, p_n)^l \tag{4}$$

In formula (4), $I(p_m) + I(p_n)$ is the sum of interestingness of the two patterns and $R(p_m, p_n)$ is the redundancy of them. The function satisfies the feature that interestingness is inversely proportional to redundancy, namely, both increasing the interestingness and decreasing the redundancy will lead to the increase of $C(p_m, p_n)$, vice versa. In the followed Theorem 1, we will explain why l can measure the users' preferences. To prove the Theorem 1, we will firstly introduce the concept of elasticity.

Definition 7 (Elasticity) [6]. Elasticity of a differentiable function f at point x is the ratio of the incremental change of the logarithm of a function with respect to an incremental change of the logarithm of the argument, it is defined as:

$$Ef(x) = x / f(x) * f'(x) = d \ln f(x) / d \ln x \tag{5}$$

Theorem 1. The variable l in the formula (4) is the users' preference proportion both interestingness and redundancy of pattern.

Proof of Theorem 1. Let $I = I(p_m) + I(p_n)$, $R = R(p_m, p_n)$, then the formula (4) is substituted by a new function only including I and R. The new function is shown as:

$$G(I, R) = I / R^l \tag{6}$$

According to the concept of elasticity, we can get elasticity of S and R respectively:

$$\begin{cases} \dfrac{\partial G}{\partial I} \Big/ \dfrac{G(I,R)}{I} = \dfrac{1}{R^l} \dfrac{I}{I / R^l} = 1 \\ \dfrac{\partial G}{\partial R} \Big/ \dfrac{G(I,R)}{R} = -l \dfrac{I}{R^{l+1}} \dfrac{R}{I / R^l} = -l \end{cases} \tag{7}$$

From formula (7) we can see that the proportion of the elasticity of R to that of I with function $G(I,R)$ is just $|l|$. According to the concept of elasticity, when I and R change 1% respectively, the relative changes of $G(I,R)$ with them are just $|l|$ times different. Hence, $|l|$ represent the proportion of the relative changes of $G(I,R)$ influenced by I and R respectively. The single influence just denote the users' preferences proportion of interestingness and redundancy, so we get Theorem 1. □

2.3 Compatibility Graph

From the evaluation function of the preferred compatibility defined in the previous subsection, it is natural to think about how to measure compatibilities among n patterns. Since the redundancies of patterns are influenced by their interestingness and themselves, we should take interestingness and redundancy into account simultaneously rather than compute them respectively.

Previous researches have employed the redundancy graph to archive all information about interestingness and redundancy of patterns. A redundancy graph of a set of patterns is a weighted complete graph where every vertex corresponds to a pattern. The weight of vertex is the interestingness of pattern and the weight on the edge (m, n) is the redundancy of p_m and p_n. However, such redundancy graph may leads to separately considering the interestingness and the redundancy.

Since each theme pattern has a compatibility with any other theme patterns, in addition, the compatibility of two theme patterns are influenced by their interestingness, it is a crucial problem that how to distribute interestingness of every pattern into their compatibility. A reasonable solution is to partition the interestingness of every pattern by n-1 parts on average if the power of the set of patterns P is n. The distributed interestingness, $ID(P_i)$, is shown as:

$$ID(P_i) = 1/n-1 * I(P_i) \tag{8}$$

According to formula (4) and formula (8), given a set of n theme patterns P, the preferred compatibility between two theme patterns can be redefined as follows:

$$CD(p_m, p_n) = [ID(p_m) + ID(p_n)]/R(p_m, p_n)^l \tag{9}$$

By formula (9), we can propose a novel structure, called compatibility graph, which is used to archive the preferred compatibilities among n patterns.

Definition 8 (Theme Compatibility Graph). Given a set of theme pattern P, the power of the set is |P|, a compatibility graph of P is a weight complete graph $G(P) = G(V, E)$ where each vertex m in the vertex set V corresponds to a pattern P_m. The edge set of $G(P)$ is $E = \{e_{uv} = CD(p_u, p_v) \mid (u, v) : u, v \in V, u \neq v\}$.

2.4 Maximal Weight Sum Problem

According to the above introductions of the evaluation function and the compatibility graph, we have stored the preferred compatibilities of a set of patterns into the compatibility graph. The next crucial step is to make a reasonable problem formulation based on the compatibility graph.

Since we aim at discovering PCTTP, the result set ought to have K patterns. Let the total compatibility of K patterns be written as TC. In general, there are the redundancies with every two patterns. With the compatibility graph, a general total compatibility of K theme patterns is shown as:

$$TC == \sum_{i=1}^{k}\sum_{j=i}^{k} CD(p_i, p_j) = \sum_{i=1}^{k}\sum_{j=i}^{k} ID(p_i) + ID(p_j) \Big/ R(p_i, p_j)^l \quad (10)$$

In formula (10), $ID(p_i)$ represents the interestingness fused into the compatibility. The goal of discovering PCTTP is to maximize the result of formula (10). Hence, we firstly define what the maximal weight sum problem is as follows:

Definition 9 (Maximal Weight Sum Problem). Given a weighted complete graph G with n vertices, selecting K vertices from n vertices to make the sum of weight on every edge is maximal.

Hence, given a set of theme patterns P whose power is |P|, formulating the problem of discovering PCTTP is to the MWSP in the compatibility graph whose number of node is |P| correspondingly. However, we can obviously find it impossible to solve MWSP by enumerating. Actually, it is a NP-Hard problem. *Why the MWSP is a NP-Hard problem? The proof of it will be given in the next section.*

3 NP-Hardness

In this section, we show that the MWSP defined above is NP-Hard.

Theorem 2. The Maximal Weight Sum Problem is NP-Hard.

It is well-known that the optimization of one problem must be NP-Hard if its decision problem is NP-Complete. Hence, we firstly transform the MWSP to its corresponding decision problem, and then prove the decision problem is NP-Complete. In order to do it, we need the following definition.

Definition 10 (Decision problem of Maximal Weight Sum Problem). Given a weighted complete graph G which has n vertices, whether there exist k vertices in n vertices with the sum, no less than a given value M, of all edges for the k vertices.

The decision problem is written as WSP (Weight Sum Problem)

Proof of Theorem 2. We firstly show that WSP can be verified in polynomial time. Suppose we are given a graph G=(V, E), an integer k and a given value M. We use the result set of vertices V' ⊆ V as a certificate for G. The verification algorithm affirms that |V'| = k, and then it checks the total weighted sum of all edges between every two vertices in V' is no less than M. This verification can be performed straightforwardly in polynomial time.

Then, we prove that the MWSP is NP-hard by showing that CLIQUE $\leq_P$ WSP. This reduction is based on the concept of the "complement" of a graph. Given an undirected complete graph $G = (V, E)$, we define the complement of G as $\overline{G} = (V, \overline{E})$, $\overline{E} = \{(u,v) : u, v \in V, u \neq v, (u,v) \notin E\}$. Let the edges in E be weighted 1 and the edges in $\overline{E}$ is weighted 0, thus we get an undirected complete graph $G' = G + \overline{G}$.

Based on the complete weighted graph G' defined above, if there is a clique of k vertices, the weighted sum of all edges in the clique would be k(k-1)/2, namely, the

weighted sum of all edges between every two vertices of k vertices is maximal. Whereas, the graph including k vertices must be a clique, if the weighted sum of all edges between every two of k vertices is k(k-1)/2. Hence, the clique problem, a well-known NP-Complete problem, can be reduced to WSP. This implies that the decision problem of MWSP is NP-Complete and so we finish the proof of Theorem 2.

4 Discovering Preferred Compatible Theme based Greedy

In this section, we describe an algorithm for mining PCTTP. Since MWSP is a NP-Hard problem, it's natural to get the idea of developing an approximate algorithm to solve MWSP. We design a greedy algorithm, called DPCTG (Discovering Preferred Compatible Theme based Greedy), which approximates the optimal solution. The pseudo-code of DPCTG is shown in Algorithm 1.

The DPCTG algorithm contains two steps. The first one is to select a edge whose weight is maximal out of n(n-1)/2 edges of the compatibility graph with n vertices and then the edge will be archived into the result set. The second one is to iteratively select the k-2 vertices from remaining n-2 vertices. Each time choose one vertex from those remained that has the maximal sum of weight between itself and all vertexes in the present result set.

Algorithm 1. Discovering Preferred Compatible Theme based Greedy

Input: A set of n closed frequent theme patterns TP
A compatibility graph contained n vertices CG
Number of output closed frequent theme, k
Output: A result set of k patterns. RS
1. Selecting two theme patterns, p_m, p_n, which are contained in a edge whose weight is maximal in all edges of CG.
2. while (The size of RS is no more than k)
3. do
4. **search for a** vertex which maximize the sum of edge weight between it and all vertexes in the present result set
5. $RS \leftarrow RS \bigcup P_i$
6. return RS

5 Quality Evaluation Model

We have transformed the post-processing of mining frequent sequential pattern from the discovering PCTTP. However, how can we measure that the result set discovered by MPCTG is the best PCTTP based on users' preferences ? Since traditional evaluation approaches of mining frequent pattern can no longer apply to the interestingness and redundancy measuring of patterns simultaneously, we propose a quality evaluation model that is able to measure the compatibility of the discovering PCTTP.

Given a set of closed theme pattern, we are able to discover k theme patterns with the top-k interestingness, and discover other k theme patterns which have the minimal redundancy between any patterns. The prefect case is that the above two set including

k theme patterns are the same, however it is impossible generally. Thus, we define the extreme interestingness of k theme patterns by summing individual interestingness of the top-k interesting theme patterns, and define the extreme redundancy of k theme patterns by summing all redundancies between of k theme patterns with the minimal redundancy of every two patterns. Then, "the extreme average compatibility of k theme patterns" is proposed, which is the ratio between extreme interestingness and extreme redundancy.

Definition 11 (Extreme Interestingness of K Theme Patterns). Given a set of the closed theme patterns $P = \{t_1, t_2, ..., t_n, \}$, the top-k interesting theme patterns among P is $IK = \{t_1, t_2, ..., t_k, \}$, the extreme interestingness of k theme patterns of S is the sum of individual interestingness of theme patterns in IK, denoted as EIK.

$$EIK = \sum_{i=1}^{k} I(t_i) \tag{11}$$

Definition 12 (Extreme Redundancy of K Theme Patterns). Given a set of the closed theme patterns $P = \{t_1, t_2, ..., t_n, \}$, the k theme patterns with minimal redundancy between any patterns among S is $RK = \{t_1, t_2, ..., t_k, \}$. The extreme redundancy of k theme patterns of S is the sum of individual redundancies between any theme patterns in RK, denoted as ERK.

$$ERK = \sum_{i=1}^{k} \sum_{j=i+1}^{k} R(t_i, t_j) \tag{12}$$

Definition 13 (Extreme Average Compatibility of K Theme Patterns). Given a set of the closed theme patterns $P = \{t_1, t_2, ..., t_n, \}$.The extreme interestingness of k theme patterns in S is EIK and the extreme redundancy of k theme patterns in S is ERK. An extreme average compatibility of k theme patterns among S is denoted as: $AECK = EIK / ERK$.

According to the above definitions, for a set of the closed theme patterns P, extreme interestingness of k theme patterns and extreme redundancy of k theme patterns measure the most ideal interestingness and redundancy among the set S respectively. Thus, they can be regarded as two measuring criterions. Thus, the extreme average compatibility of k theme patterns is naturally regarded as the criterion measuring the compatibility of k theme patterns. In order to quantify the quality of the compatibility of k theme patterns, it is necessary to compute an approximation ratio with the extreme average compatibility of k theme patterns. Therefore, the average compatibility of k theme patterns and the approximation ratio of k theme patterns will be defined as follows.

Definition 14 (Average Compatibility of K Theme Patterns). Given a set including k theme patterns $KT = \{t_1, t_2, ..., t_n, \}$. An average compatibility of k theme patterns among KT is denoted as:

$$ACK = \sum_{i=1}^{k} I(t_i) \Big/ \sum_{i=1}^{k} \sum_{j=i+1}^{k} R(t_i, t_j) \tag{13}$$

Definition 15 (Approximation Ratio of K Theme Patterns). Given a set of the closed theme patterns $P = \{t_1, t_2, ..., t_n, \}$ and a set of the discovering k theme patterns $KT = \{t_1, t_2, ..., t_n, \}$ .The $AECK$ and ACK can be computed respectively. An approximation ratio of k theme patterns is denoted as: $AR = ACK / AECK$

To sum up, the quality evaluation model compute the approximation ratio of k theme patterns to quantify the quality of the compatibility of the discovering k theme patterns from text. The approximation ratio ranges in [0,1], and The more the ratio closer to 1,the better the compatibility of discovering k patterns is. The further discussion of the quality evaluation model will be shown in the next section.

6 Experimental Study

In this section we will provide the empirical evaluation for the effectiveness of our strategy and EPCEG algorithm for real-world tasks. The EPCEG algorithm is implemented in Java. The version of JDK is JDK1.52. All experiments are performed on a 2.0GHZ, 2GB-memory, Intel PC running Windows XP.

The four text datasets used in our experiment all come from subsets of the DBLP dataset. They contain papers from the proceedings of 25 major conferences, such as SIGMOD, SIGKDD, VLDB, SIGIR and etc., in Data Mining, Database and Information Retrieval. Considering the relations and the differences between the above four subjects, we classify these titles of papers into four text datasets, including database, data mining, database and data mining, data mining and information retrieval respectively. The detail information of datasets is shown in Table 1. In those experiments, we firstly use the tool Clospan [10] to generate closed sequential patterns, namely closed theme patterns, and the title terms are stemmed by Krovertz stemmer [2]. Then we discover the PCTTP based on the result set from the Clospan.

According to the quality evaluation model, we use TSP (mining closed top-k sequential pattern algorithm [5]) and DPCTG by different support thresholds, at the users' preferences proportion $l = 5/5 = 1$, on the four datasets respectively, and the approximation ratios are illustrated as curves in the figures. For the sparse datasets, the chosen support thresholds are low than 5%. From the figures, we can see that the compatibility of the set of discovering PCTTP is obviously better than that of the result set of TSP as the min_support is decreasing. Results of four datasets are shown respectively with the support threshold set 5 as a representative value. Four columns of each represent the 10 top results of TSP, and DPCTG with different l respectively. In ordering to explore the generalization of results, we assume l as 1/9, 5/5 and 9/1 in table2, l as 2/8, 5/5 and 8/2 in table 3, l as 3/7, 5/5 and 7/3 in table 4 and l as 1/9, 5/5 and 9/1 in table 5. From these tables, the italic represents the inclusion relations among this theme, and the bold represents the significant and focus theme. These tables indicate that our proposed strategy can meet the different demands when we adjust the users' preferences proportion.

To sum up, DPCTG can find effective result set of PCTTP from a large set of text. In addition, different users can obtain different result sets to satisfy the users' preferences with interestingness and redundancy by DPCTG.

Table 1. Information of four data sets

Table ID	Topic of data set	Number of Transaction	Conferences
Table 2	Data mining	984	AAAI ,SIGKDD, CIKM,SDM, ICDM, PKDD, PAKDD, ADMA
Table 3	Database	15280	SIGMOD, PODS, VLDB, ICDE, ICDT, CIDR, ER, EDBT SSDBM, DASFAA, WAIM
Table 4	Data Mining and Database	24404	AAAI ,SIGKDD, SIGMOD, PODS, VLDB, ICDE, ICDT, CIDR, ER, CIKM, SDM, ICDM, EDBT SSDBM, PKDD, PAKDD, DASFAA, WAIM, ADMA
Table 5	Data Mining and Information Retrieval	13531	AAAI, SIGKDD, SIGIR,WWW,WISE,ECIR , CIKM,SDM, ICDM, PKDD, PAKDD, APWeb, DMA

Table 2. Top-10 Theme Pattern on the Dataset of DM

Top-k	TSP	DPCTG (R/I=1/9)	DPCTG (R/I=5/5)	DPCTG (R/I=9/1)
1	*data*	*data*	data mining	**association rule**
2	*mining*	*mining*	**association rule**	data clustering
3	clustering	clustering	**frequent pattern**	**support vector machine**
4	*data mining*	*data mining*	database	knowledge discovery
5	database	database	classification	information retrieval
6	pattern	pattern	data clustering	**feature selection**
7	classification	classification	search	**text classification**
8	learning	learning	**time series**	**data streams**
9	discovery	association rule	**support vector machine**	mining database
10	rule	discovery	information retrieval	**neural network**

Table 3. Top-10 Theme Pattern on the Dataset of DB

Top-k	TSP	DPCTG (R/I=2/8)	DPCTG (R/I=5/5)	DPCTG (R/I=8/2)
1	data	data	database system	**database management system**
2	*database*	*database*	database management	**distributed database**
3	query	query	**distributed database**	**relational database system**
4	*system*	management	**query processing**	database design
5	management	*database system*	database design	data streams
6	processing	relational	data streams	**query optimization**
7	relational	query processing	xml	data model
8	model	data model	data model	data mining
9	*database system*	xml	data mining	query processing
10	xml	data stream	relational database	**concurrency control**

Table 4. Top-10 Theme Pattern on the Dataset of DM&DB

Top-k	TSP	DPCTG (R/I=3/7)	DPCTG (R/I=5/5)	DPCTG (R/I=7/3)
1	*data*	*data*	data	**association rule mining**
2	database	database	data mining	distributed database systems
3	*mining*	*query*	database	query processing
4	query	*data mining*	data clustering	**Support vector machine**
5	system	system	association rule	data clustering
6	clustering	information	Data stream	Data stream
7	information	*query processing*	query processing	time series
8	*data mining*	data clustering	information	knowledge discovery
9	management	data management	**distributed database systems**	**database management systems**
10	processing	relational	time series	mining pattern

Table 5. Top-10 Theme Pattern on the Dataset of DM&IR

Top-k	TSP	DPCTG (R/I=1/9)	DPCTG (R/I=5/5)	DPCTG (R/I=9/1)
1	data	data	information retrieval	**information retrieval system**
2	web	web	**data**	**association rule mining**
3	mining	mining	**data mining**	web search
4	*retrieval*	*retrieval*	database	knowledge discovery
5	*information*	*information*	classification	**support vector machine**
6	search	search	clustering	**web semantic**
7	clustering	clustering	web search	feature selection
8	*Information retrieval*	*information retrieval*	text	web mining
9	text	data mining	**web sematic**	text classification
10	learning	web search	**association rule mining**	**time series data**

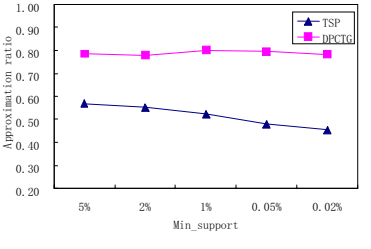

Fig. 1. Approximation Ratio on DM

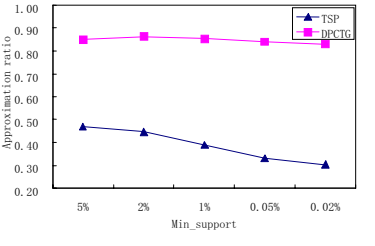

Fig. 2. Approximation Ratio on DB

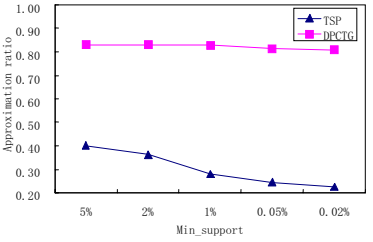

Fig. 3. Approximation Ratio on DM&DB

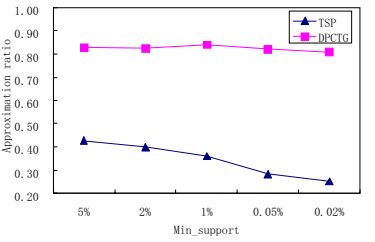

Fig. 4. Approximation Ratio on DM&IR

7 Relate Work

To the best of our knowledge, the problem of discovering PCTTP has not been well studied in current work. Most studies of frequent pattern mining still focus on the fast mining result set from the database and dose not pay attention to the post-processing of KDD. Recent studies in this field have paid close attention to mining compatible top-k patterns with both high-interestingness and low-redundancy, instead of finding top-k closed frequent patterns or compressing the size of frequent patterns to exclude redundancy among them separately, such as mining closed frequent pattern, compressing patterns, summarizing patterns and so on[1, 7, 9, 10].

References [5] [9] mine top-k closed frequent sequences and summarizing patterns respectively. But these papers only study the single factor, either interestingness or redundancy, not evaluate compatibility of the both to the effect of the results. Reference [8] mines top-k theme patterns considering interestingness and redundancy simultaneously, however, ignores to study the influences of different users on results, leading to the limits of the model.

In the field of text mining, semantic analysis is still a hotspot topic [3, 4]. However, there are few focuses on the problem of discovering theme patterns based on users' preferences considering interestingness and redundancy simultaneously.

8 Conclusion

Existing work of discovering a representative set of theme patterns from a large amount of text by frequent pattern mining usually generates a huge amount of theme patterns for the downward closure property. Thus, these discriminative theme patterns will be drowned in a large number of redundant patterns. Some recent post-processing studies of KDD introduced the technique of mining a set of top-k frequent patterns with both high-interestingness and low-redundancy. However, since different users have different preferences with interestingness and redundancy, it is hard to use one unified criterion to discover theme patterns with both high-interestingness and low-redundancy. The problem of discovering PCTTP has not been well addressed so far.

In this paper, the novel strategy of the post-processing in text mining is proposed, that is, discovering PCTTP based on users' preferences. For the problem discovering PCTTP, the evaluation function of preferred compatibility between every two theme

patterns is presented firstly. Then the preferred compatibilities are archived into a data structure called theme compatibility graph, and a problem called MWSP based on the compatibility graph is proposed to formulate the problem how to discover the PCTTP. In addition, since MWSP is proved to be a NP-Hard problem, a greedy algorithm, DPCTG, is designed to approximate the optimal solution of the MWSP. Empirical studies indicate that a high quality set of PCTTP can be obtained on the different text datasets from DBLP.

The proposed strategy and algorithm of this paper is general, however, we only study the theme patterns from a huge amount of text datasets, and use tf-idf approach to weight the interestingness of theme pattern. In the future work, we will study the text datasets which have the time attributes, and extend our strategy to other information retrieval models, such as the probabilistic model.

Acknowledgments

We would like to thank Zhiyuan Cheng and Yuhan Song for discussions regarding our work. This research was supported by National 973 Project of China under Grant No.2005CB321902.

References

1. Afrati, F.N., Gionis, A., Mannila, H.: Approximating a collection of frequent sets. In: KDD 2004, pp. 8–19 (2004)
2. Krovetz, R.: Viewing morphology as an inference process. In: Proceedings of SIGIR 1993, pp. 191–202 (1993)
3. Mei, Q., Xin, D., Cheng, H., Han, J., Zhai, C.: Generating semantic annotations for frequent patterns with context analysis. In: KDD 2006, pp. 337–346 (2006)
4. Mei, Q., Xin, D., Cheng, H., Han, J., Zhai, C.: Discovering Evolutionary Theme semantic annotations for frequent patterns with context analysis. In: KDD 2005 (2005)
5. Tzvetkov, P., Yan, X., Han, J.: TSP: Mining top-k closed sequential patterns. Knowledge and Information Systems 7, 438–457 (2005)
6. Varian, H.: Intermediate Microeconomics: A Modern Approach, 6th edn. W.W. Norton & Company Inc. (2003)
7. Xin, D., Han, J., Yan, X., Cheng, H.: On compressing frequent patterns. In: KIS 2007 (2007)
8. Xin, D., Han, J., Yan, X., Cheng, H.: Discovering Redundancy-Aware Top-K Patterns. In: KDD 2006, pp. 314–323 (2006)
9. Yan, X., Cheng, H., Han, J., Xin, D.: Summarizing Itemset Patterns: A Profile Based Approach. In: KDD 2005, pp. 314–323 (2005)
10. Yan, X., Han, J., Afshar, R.: Clospan: Mining closed sequential patterns in large datasets. In: SDM 2003, pp. 166–177 (2003)