Pattern-Based Web Mining Using Data Mining Techniques

Sheng-Tang Wu and Yuefeng Li

***Abstract*—In the last decade, many data mining techniques have been proposed for fulfilling various knowledge discovery tasks in order to achieve the goal of retrieving useful information for users. Data mining techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining and closed pattern mining. However, how to effectively exploit the discovered patterns is still an open research issue, especially in the domain of Web mining. In this study, we compare these data mining methods based on the use of several types of discovered patterns. The performance of the pattern mining algorithms is investigated on the Reuters dataset RCV1 for completing Web mining tasks. The experimental results show that the closed pattern methods, such as SCPM and NSCPM, have better performance due to the use of pruning mechanism in the pattern discovery stage.**

***Index Terms*—Web mining, data mining, pattern taxonomy model.**

I. Introduction

Web mining is the technique that helps users find useful information from the rich data on the World Wide Web. Due to the rapid growth of digital data made available in recent years, Web mining and data mining have attracted great attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users [1], [2]. Data mining is therefore an essential step in the process of knowledge discovery in databases.

In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining and closed pattern mining. Most of them are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame. With a large number of patterns generated by using the data mining approaches, how to effectively exploit these patterns is still an open research issue.

Manuscript received March 12, 2013; revised May 20, 2013.This work was supported in part by the National Science Council under Grant NSC 101-2221-E-468-029.

Sheng-Tang Wu is with the Department of Applied Informatics and Multimedia, Asia University, Taichung, Taiwan (e-mail: swu@ asia.edu.tw).

Yuefeng Li is with the Faculty of Science and Engineering, Queensland University of Technology, Brisbane, Australia (e-mail: y2.li@qut.edu.au).

The World Wide Web provides rich information on an extremely large amount of linked Web pages. Such a repository contains not only text data but also multimedia objects, such as images, audio and video clips. Data mining on the World Wide Web can be referred to as Web mining which has gained much attention with the rapid growth in the amount of information available on the internet. Web mining is classified into several categories, including Web content mining, Web usage mining and Web structure mining.

Most Web text mining methods use the keyword-based approaches, whereas others choose the phrase technique to construct a text representation for a set of documents. It is believed that the phrase-based approaches should perform better than the keyword-based ones as it is considered that more information is carried by a phrase than by a single term. Based on this hypothesis, Lewis [3] conducted several experiments using phrasal indexing language on a text categorization task. Ironically, the results showed that the phrase-based indexing language was not superior to the word-based one.

Although phrases carry less ambiguous and more succinct meanings than individual words, the likely reasons for the discouraging performance from the use of phrases are: (1) phrases have inferior statistical properties to words, (2) they have a low frequency of occurrence, and (3) there are a large number of redundant and noisy phrases among them [4], [5].

In order to solve the above mentioned problem, new studies have been focusing on finding better text representatives from a textual data collection. One solution is to use the data mining techniques, such as sequential pattern mining, for building up a representation with the new type of features [6]. 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 the feature set by removing noisy patterns. However, treating each multi-terms pattern as an atom in the representation seems likely to encounter the low-frequency problem while dealing with the long patterns [7]. Another challenge for the data mining-based methods is that more time is spent on uncovering knowledge from the data; consequently less significant improvements are made compared with information retrieval methods [8], [9].

The rest of this paper is structured as follows. Section II describes the related works and the terminology used in this study is presented in Section III. Following is the discussion of experimental results. Finally, Section V concludes this study work.

II. Literature Review

Most research works in the data mining community have focused on developing efficient mining algorithms for

discovering a variety of patterns from a larger data collection. However, searching for useful and interesting patterns is still an open problem [10]. In the field of text mining, data mining techniques can be used to find various text patterns, such as sequential patterns, frequent itemsets, co-occurring terms and multiple grams, for building up a representation with these new types of features [6].

Using phrases for the semantic representation still has doubts in increasing performance over domains of text categorization tasks [3], [4], meaning that there exists no particular representation method with dominating advantage over others [11], [12]. One solution is Pattern Taxonomy Model [8], which uses data mining technique to form the representation patterns.

Data mining techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining and closed pattern mining. The association rules mining, first studied in [13] for market basket analysis, is to find any association rules satisfying user-specified minimum support and minimum confidence [14]. An association rule is the discovery of the associative relationships among objects; i.e., the appearance of a set of objects in a database is strongly related to the appearance of another set of objects [15]. The basic problem of finding association rules is introduced in [13].

The problems of mining association rules from large databases can be decomposed into two sub-problems: (1) Find itemsets whose support is greater than the user-specified minimal support; (2) Use the frequent itemsets to generate the desired rules [16]. Much of the research has been focused on the former [13], [17]. Many variations of the Apriori algorithm and its applications have been extensively investigated in the literature [18]-[21]. Liu *et al*. [22] mined frequent itemsets from the Web to find topic-specific concepts and definitions. Maintaining frequent itemsets in dynamic databases is examined by Zhang et al. [23]. Mining Top-K frequent itemsets is suggested in [24].

Web mining is classified into several categories, including Web content mining, Web usage mining and Web structure mining[25]. Data mining on the World Wide Web can be referred to as Web mining which has gained much attention with the rapid growth in the amount of information available on the internet.

III. Pattern Taxonomy Model

The Pattern Taxonomy Model (PTM) [26] is proposed to replace the keyword-based methods by using tree-like taxonomies as concept representatives. Taxonomy is a structure that contains information describing the relationship between sequence and sub-sequence. In addition, the performance of PTM-based models is improved by adopting the closed sequential patterns. The removal of non-closed sequential patterns also results in the increase of efficiency of the system due to the shrunken dimensionality.

From the data mining point of view, the conceptual patterns are defined as two types: sequential pattern and non-sequential pattern. The definition is described as follows: Firstly, let $T = \{t_1, t_2, ..., t_k\}$ be a set of terms, which can be viewed as words or keywords in a dataset. A non-sequential pattern is then a non-ordered list of terms, which is a subset of T, denoted as $\{s_1, s_2, ..., s_m\}$ ($s_i \in T$). A sequential pattern, defined as $S = \langle s_1, s_2,...,s_n \rangle$ ($s_i \in T$), is an ordered list of terms. Note that the duplication of terms is allowed in a sequence. This is different from the usual definition where a pattern consists of distinct terms.

After mining conceptual patterns, the relationship between patterns has to be defined in order to establish the pattern taxonomies. Sub-sequence is defined as follows: if there exist integers $1 \le i_1 \le i_2 \le \ldots \le i_n \le m$, such that $a_1 = b_{i1}$, $a_2 = b_{i2}$,..., $a_n = b_{in}$, then a sequence $\alpha = \langle a_1, a_2,...,a_n \rangle$ is a sub-sequence of another sequence $\beta = \langle b_1, b_2,...,b_m \rangle$. For example, sequence $\langle s_1, s_3 \rangle$ is a sub-sequence of sequence $\langle s_1, s_2, s_3 \rangle$. However, sequence $\langle s_3, s_1 \rangle$ is not a sub-sequence of $\langle s_1, s_2, s_3 \rangle$ since the order of terms is considered. In addition, we can also say sequence $\langle s_1, s_2, s_3 \rangle$ is a super-sequence of $\langle s_1, s_3 \rangle$. The problem of mining sequential patterns is to find a complete set of sub-sequences from a set of sequences whose support is greater than a user-predefined threshold (minimum support).

We can then acquire a set of frequent sequential concept-patterns CP for all documents $d \in D^+$, such that $CP = \{p_1, p_2,\ldots, p_n\}$. The absolute support $supp_a(p_i)$ for all $p_i \in CP$ is obtained as well. We firstly normalize the absolute support of each discovered pattern based on the following equation:

$$\text{support: } CP \to [0,1]$$

such that

$$\text{support}(p_i) = \frac{supp_a(p_i)}{\sum_{p_j \in CP} supp_a(p_j)}$$

As aforementioned, statistical properties (such as support and confidence) are usually adopted to evaluate the patterns while using data mining techniques to mine frequent patterns. However, these properties are not effective in the stage of pattern deployment and evolution. The reason is the short patterns will be always the major factors affecting the performance due to their high frequency. Therefore, what we need is trying to adopt long patterns which provide more descriptive information.

IV. Experimental Results

We apply data mining patterns to the Web mining task using real Web dataset for performance evaluation. Several standard benchmark datasets are available for experimental purposes, including Reuters Corpora, OHSUMED and 20 Newsgroups Collection. The dataset used in our experiment in this study is the Reuters Corpus Volume 1 (RCV1) [27].

RCV1 includes 806,791 English language news stories which were produced by Reuters journalists for the period between 20 August 1996 and 19 August 1997. These documents were formatted using a structured XML scheme. Each document is identified by a unique item ID and corresponded with a title in the field marked by the tag <title>. The main content of the story is in a distinct <text> field

consisting of one or several paragraphs. Each paragraph is enclosed by the XML tag <p>. In our experiment, both the "title" and "text" fields are used and each paragraph in the "text" field is viewed as a transaction in a document. Fig. 1 shows RCV1 document as an example.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <newsitem itemid="26642" id="root" date="1996-09-01"
  xml:lang="en">
  <title>INDIA: At least 44 dead as vessel capsizes in
  India.</title>
  <headline>At least 44 dead as vessel capsizes in
  India.</headline>
  <dateline>NEW DELHI 1996-09-01</dateline>
- <text>
  <p> At least 44 people were feared drowned when their vessel
  capsized in the Nagavalli river in the southern state of Andhra
  Pradesh, the United News of India said on Sunday. </p>
  <p> It quoted official sources as saying the boat was carrying
  some 50 people, mainly tribespeople, when it sank on Saturday.
  </p>
  <p> Six people swam to safety, it said. </p>
  </text>
  <copyright> (c) Reuters Limited 1996 </copyright>
- <metadata>
```

Fig. 1. An example RCV1 document.

The primary result of pattern discovery using data mining method is shown in Fig. 2. The discovered patterns then can be used as representative features in the Web mining task.

```
<< 63261.xml >> => 1title + 4 paragraphs => 32 words
 (T)bill senat
 (1)bill theft trade secret foreign compani feder crime
    final action senat
 (2)senat version bill pass hous version pass hous final
    action hous
 (3)bill compani theft feder crime
 (4)foreign trade secret
====Found Patterns:
 [1Terms]:([senat](3)) Freq:3, rel_supp:0.6
 [1Terms]:([bill](4))Freq:4, rel_supp:0.8
 [1Terms]:([foreign](2))Freq:2, rel_supp:0.4
 [2Terms]:([bill](4),senat) Freq:2, rel_supp:0.4
 [2Terms]:([trade](2),secret) Freq:2, rel_supp:0.4
 [3Terms]:([bill,final](2),action) Freq:2, rel_supp:0.4
 [4Terms]:([bill,theft,feder](2),crime) Freq:2,
          rel_supp:0.4
 [4Terms]:([bill,compani,feder](2),crime) Freq:2,
          rel_supp:0.4
```

Fig. 2. An example of pattern discovery by using sequential pattern mining method.

The Sequential Pattern Mining (SPM) algorithm SPMining is depicted in Algorithm 1. In this algorithm, we apply the pruning scheme for the purpose of eliminating non-closed patterns during the process of sequential patterns discovery. The key feature behind this recursive algorithm is represented in the first line of the algorithm, which describes this pruning procedure. In this algorithm, all (*n-1*)*Terms* of length patterns are diagnosed to determine whether or not they are closed patterns after all *nTerms* of length patterns are generated from the previous recursion. 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.

Algorithm 1. SPMining(*PL*, *min_sup*)
Input: a list of *nTerms* frequent sequential patterns, *PL*; minimum support, *min_sup*.
Output: a set of frequent sequential patterns, *SP*.
Method:
1: $SP \leftarrow SP - \{Pa \in SP \mid \exists Pb \in PL$ Such That $len(Pa) = len(Pb) - 1$ And $Pa < Pb$ And $supp_a(Pa) = supp_a(Pb)\}$ //Pattern Pruning
2: $SP \leftarrow SP \cup$ PL //*nTerms pattern set*
3: $PL' \leftarrow \phi$
4: **For Each** *pattern p* In *PL* **Do Begin**
5: generating *p-projected* database *PD*
6: **For Each** frequent *term t* In *PD* **Do Begin**
7: $P' = p \oplus t$ //sequence extension
8: **If** $supp_r(P') \geq min\_sup$ **Then**
9: $PL' \leftarrow PL' \cup P'$
10: **End If**
11: **End For**
12: **End For**
13: **If** $|PL'| = 0$ **Then**
14: Return //no more pattern
15: **Else**
16: Call SPMining(PL', *min_sup*)
17: **End If**
18: Output *SP*

The algorithm SPMining is developed for the purpose of mining all frequent sequential patterns from documents. In addition to sequential patterns, non-sequential patterns mining (NSPM) from a set of textual documents is another application of the data mining mechanism. From the data mining point of view, non-sequential patterns can be treated as frequent itemsets extracted from a transactional database. Frequent itemset mining is one of the most essential issues in many data mining applications. The pseudo code of NSPMining is listed in Algorithm 2.

Algorithm 2. NSPMining(*NP*, *FT*, *min_sup*)
Input: a list of *nTerms* frequent non-sequential patterns, *NP*; a list of *1Term* frequent patterns, *FT*; minimum support, *min sup*.
Output: a set of frequent non-sequential patterns, *FP*.
Method:
1: $FP \leftarrow FP \cup$ NP //*nTerms* non-sequential patterns
2: $NP' \leftarrow \phi$
3: **For Each** pattern *p* in *NP* **Do Begin**
4: **For Each** frequent term *t* in *FT* **Do Begin**
5: $P' = p \cup \{t\}$ //pattern growing
6: **If** $supp_r(P') \geq min\_sup$ **Then**
7: $NP' \leftarrow NP' \cup P'$
8: **End If**
9: **End For**
10: **End For**
11: **If** $|NP'| = 0$ **Then**
12: Return //no more pattern
13: **Else**
14: Call NSPMining(NP', *FT*, *min_sup*)
15: **End If**
16: Output *FP*

The effectiveness of the Web mining model is evaluated by performing information filtering task with real Web dataset RCV1. The experimental results of data mining methods are compared to those of other baselines, such as PTM, using several standard measures. These measures

include Precision, Recall, *Top-k* (*k* = 20 in this study), Breakeven Point (*b/e*), F_β-measure, Interpolated Average Precision (*IAP*) and Mean Average Precision (*MAP*).

TABLE I: CONTINGENCY TABLE

		human judgement	
		yes	*no*
system judgement	*yes*	TP	FP
	no	FN	TN

The precision is the fraction of retrieved documents that are relevant to the topic, and the recall is the fraction of relevant documents that have been retrieved. For a binary classification problem the judgment can be defined within a contingency table as depicted in Table I. According to the definition in this table, the measures of Precision and Recall are denoted as TP/(TP+FP) and TP/(TP+FN) respectively, where TP (True Positives) is the number of documents the system correctly identifies as positives; FP (False Positives) is the number of documents the system falsely identifies as positives; FN (False Negatives) is the number of relevant documents the system fails to identify.

The precision of *top-K* returned documents refers to the relative value of relevant documents in the first *K* returned documents. The value of *K* we use in the experiments is 20, denoted as "*t20*". Breakeven point (*b/e*) 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.

Both the *b/e* and *F1*-measure are the single-valued measures in that they only use a figure to reflect the performance over all the documents. However, we need more figures to evaluate the system as a whole. Therefore, another measure, Interpolated Average Precision (*IAP*) is introduced. This measure is used to compare the performance of different systems by averaging precisions at 11 standard recall levels (i.e., recall=0.0, 0.1, ..., 1.0). The 11-points measure is used in our comparison tables indicating the first value of 11 points where recall equals to 116 Experiments and Results zero. Moreover, Mean Average Precision (*MAP*) is used in our evaluation which is calculated by measuring precision at each relevance document first, and averaging precisions over all topics.

The Fig. 3 shows the comparison of all data mining methods in precision at standard recall points on the first 50 RCV1 topics. It reveals that all data mining methods have the similar performance. The SCPM and NSCPM, which adopt closed patterns, perform a little better than others around the low recall situation. PTM outperforms data mining methods due to the use of pattern pruning and deploying techniques.

The comparison of number of patterns and runtime using data mining methods is indicated in Fig. 4. The NSPM method needs a lot of runtime since the huge amount of patterns generated during the projected stage. The closed pattern methods, such as SCPM and NSCPM, produce less number of patterns due to the adoption of pruning process during the pattern discovery stage. PTM also outperforms data mining methods since the less number of patterns is used.

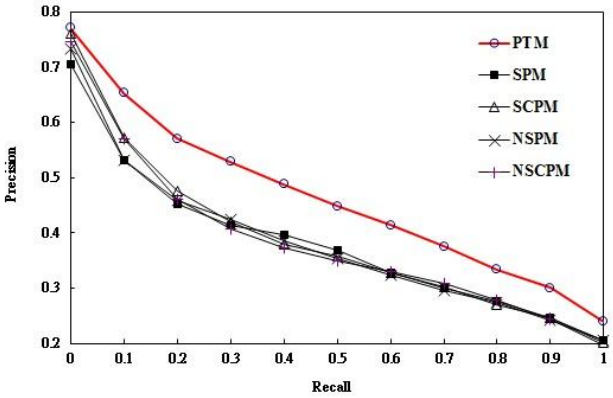

Fig. 3. Comparison of all methods in precision at standard recall points on the first 50 RCV1 topics.

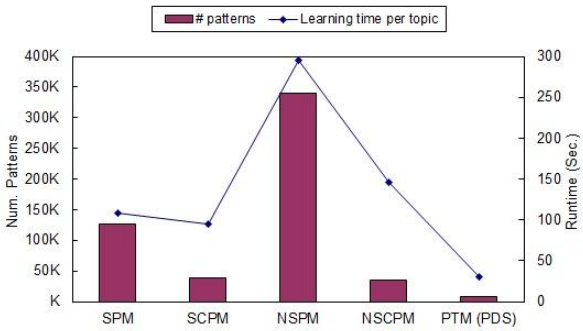

Fig. 4. Comparison of number of patterns and runtime using data mining methods.

Table II illustrates the experimental results on all measures for the performance comparison. It shows SCPM has better result on F_1 measure comparing to other data mining methods. However, NSCPM performs better on *t20* measure.

TABLE II: PERFORMANCE OF DATA MINING METHODS AND PTM

Method	*t20*	*b/e*	F_1	*IAP*	*MAP*
SPM	0.401	0.343	0.385	0.384	0.361
SCPM	0.406	0.353	0.390	0.392	0.364
NSPM	0.412	0.352	0.386	0.384	0.361
NSCPM	0.428	0.346	0.385	0.387	0.361
PTM	**0.490**	**0.431**	**0.440**	**0.465**	**0.441**

V. CONCLUSION

In general, a significant amount of patterns can be retrieved by using the data mining techniques to extract information from Web data. Many data mining techniques have been proposed in the last decade. These techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. However, how to effectively use these discovered patterns is still an unsolved problem. Another typical issue is that only the statistic properties (such as support and confidence) are used while evaluating the effectiveness of patterns. A comprehensive comparison of data mining methods applied for Web mining task is

performed in this study. The experimental results show that closed pattern methods, such as SCPM and NSCPM, have better performance due to the use of pruning mechanism in the pattern discovery stage. It also proves that the less number of patterns used by these two methods does not affect their effectiveness.

REFERENCES

[1] V. Devedzic, "Knowledge discovery and data mining in databases," in *Handbook of Software Engineering and Knowledge Engineering*, vol. 1, 2001, pp. 615-637.

[2] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: an overview," *AI Magazine*, vol. 13, pp. 57-70, 1992.

[3] D. D. Lewis, "An evaluation of phrasal and clustered representations on a text categorization task," *SIGIR*, 1992, pp. 37–50.

[4] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.

[5] N. Zhong, Y. Li, and S. T. Wu, "Effective Pattern Discovery for Text Mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 1, pp. 30-44, 2012.

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

[7] S. T. Wu, Y. Li, and Y. Xu, "An effective deploying algorithm for using pattern-taxonomy," in *Proc. iiWAS05*, 2005, pp. 1013-1022.

[8] S. T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in *Proc. ICDM*, 2006, pp. 1157-1161.

[9] S. T. Wu and H.-Y. Ho, "Using Negative Examples in Pattern Taxonomy Model for Knowledge Discovery," in *Proc. IETAC*, 2011, pp. 129-134.

[10] T. Y. Lin, "Database mining on derived attributes," *Rough Sets and Current Trends in Computing*, 2002, pp. 14–32.

[11] R. Kosala and H. Blockeel, "Web mining research: A survey," *ACM SIGKDD Explorations*, vol. 2, no. 1, pp. 1-15, 2000.

[12] S. Scott and S. Matwin, "Feature engineering for text classification," in *Proc. ICML*, 1999, pp. 379-388.

[13] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets and items in large database," in *Proc. ACM-SIGMOD*, 1993, pp. 207-216.

[14] K. Wang, Y. He, and J. Han, "Pushing support constraints into association rules mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 3, pp. 642-658, 2003.

[15] Y. J. Fu, "Data mining: Tasks, techniques and applications," *IEEE Potentials*, vol. 16, no. 4, pp. 18-20, 1997.

[16] J. Han and Y. Fu, "Mining multiple-level association rules in large databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no. 5, pp. 798-805, 1999.

[17] J. S. Park, M. S. Chen, and P. S. Yu, "An effective hash-based algorithm for mining association rules," in *Proc. SIGMOD*, 1995, pp. 175-186.

[18] L. Dumitriu, "Interactive mining and knowledge reuse for the closed-itemset incremental-mining problem," *SIGKDD Explorations*, vol. 3, no. 2, pp. 28-36, 2002.

[19] K. Gouda and M. J. Zaki, "Genmax: An efficient algorithm for mining maximal frequent itemsets," *Data Mining and Knowledge Discovery*, vol. 11, no. 3, pp. 223-242, 2005.

[20] J. Pei, J. Han, and L. V. S. Lakshmanan, "Pushing convertible constraints in frequent itemset mining," *Data Mining and Knowledge Discovery*, vol. 8, no. 3, pp. 227-252, 2004.

[21] M. J. Zaki and C.-J. Hsiao, "Charm: An efficient algorithm for closed itemset mining," in *SIAM International Conference on Data Mining*, 2002, pp. 457-473.

[22] J. Liu, Y. Pan, K. Wang, and H. Han, "Mining frequent item sets by opportunistic projection," in *Proc. KDD*, 2002, pp. 229-238.

[23] S. Zhang, X. Wu, J. Zhang, and C. Zhang, "A decremental algorithm for maintaining frequent itemsets in dynamic databases," in *Proc. International Conference on Data Warehousing and Knowledge Discovery (DaWaK05)*, 2005, pp. 305-314.

[24] R. C. Wong and A. W. Fu, "Mining top-k frequent itemsets from data streams," *Data Mining and Knowledge Discovery*, vol. 13, no. 2, pp. 193-217, 2006.

[25] Y. Li, X. Z. Chen, and B. R. Yang, "Research on web mining-based intelligent search engine," in *International Conference on Machine Learning and Cybernetics*, 2002, pp. 386-390.

[26] S. T. Wu, Y. Li, and Y. Xu, "An effective deploying algorithm for using pattern-taxonomy," in *Proc. the 7th International Conference on Information Integration and Web-based Applications & Services (iiWAS05)*, 2005, pp. 1013-1022.

[27] T. Rose, M. Stevenson, and M. Whitehead, "The reuters corpus volume1- from yesterday's news to today's language resources," in *Proc. Inter. Conf. on Language Resources and Evaluation*, 2002, pp. 29-31.

Sheng-Tang Wu received the MS and PhD degrees in the Faculty of Information Technology at Queensland University of Technology, Brisbane, Australia in 2003 and 2007, respectively. He is currently an assistant professor in the Department of Applied Informatics and Multimedia, Asia University, Taiwan. His research interests include data mining, Web intelligence, information retrieval, information systems, and multimedia. He also received the honor of the Outstanding Doctoral Thesis Award at Queensland University of Technology.

Yuefeng Li is the leader of the eDiscovery Lab in the Institute for Creative Industries and Innovation, and a professor in the Discipline of Computer Science, Faculty of Science and Engineering, Queensland University of Technology, Australia. He has established a strong reputation internationally in the fields of Web Intelligence, Text Mining and Ontology Leaning, and has been awarded three Australian Research Council grants.