Inadequate use of patterns derived from data mining techniques leads to the ineffective performance

Dr. Sunil Kumar Sah[1], Amit Kumar[2]

[1]*faculty Member ,Udca, Computer Centre, T. M. Bhagalpur University Bhagalpur.*
[2]*(Research Scholar), Dept.Of Stat. And Computer App., Tmbu, Bhagalpur.*

Abstract: *Many data mining techniques have been proposed for mining useful patterns in databases. However, how to effectively utilize discovered patterns is still an open research is- sue, especially in the domain of text mining. Most existing methods adopt term-based approaches.However, they all suffer from the problems of polysemy and synonymy. This paper presents an innovative technique, pattern taxonomy mining, to improve the effectiveness of using discovered pat- terns for finding useful information. Substantial experiments on RCV1 demonstrate that the proposed solution achieves encouraging performance.*

Keywords: *Text mining, Pattern Taxonomy, Pattern Evolving*

I. Introduction

Text mining is the discovery of knowledge in text documents. It is a challenging issue to find accurate knowledge in text documents to help users to find what they want. In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge, such as Rocchio and probabilistic models [1], rough set models [5], BM25 and SVM [9] based filtering models.However, term-based methods suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that phrase-based approaches should perform better than the term-based ones, as phrases are more discriminative and arguably carrying more “semantics”. This hypothesis has not fared too well in the history of IR [10]. Towards this direction, pattern-based approach PTM [12, 11] has been proposed, which adopts the concepts of closed patterns, and pruns non-closed patterns. This paper presents the technique of inner pattern evolution used in PTM, which considers the influence of patterns from negative training examples in finding ambiguous patterns and reducing their impact. The rest of this paper is structured as follows. Section 2 discusses the related works. Section 3 proposes the technique of inner pattern evolution in PTM. Following is the discussion of experimental setting and results. Finally, Section 5 concludes this study work.

II. Relatedworks

Pattern mining has been extensively studied in data mining communities for many years. These research works have mainly focused on developing efficient mining algorithms for discovering patterns from a larger data collection. However, searching for useful and interesting patterns and rules is still an open problem [6, 4, 13]. Moreover, the challenging issue is how to effectively deal with the large amount of discovered patterns. For this issue, we have used closed sequential pat- terns for text mining in [12], where we have firstly verified that the concepts of closed patterns in text mining were significant. We also proposed pattern taxonomy model in [11] to improve the effectiveness of using closed patterns in text mining. Also, pattern evolution technique was introduced in [7] in order to improve the performance of ontology mining.

III. Innerpatternevolution

Considering the influence of patterns from negative training examples, we re-shuffle supports of terms within normal forms of d-patterns NDP [7] based on negative documents in a training set. This technique is called Inner Pattern Evolution (IPE), since it only changes a pattern’s term supports within the pattern. A threshold is usually used to decide the relevance of incoming documents. A noise negative document nd in a set of negative documents D− is a negative document that the system falsely identified as a positive. In order to reduce the noise, we need to track which d-patterns have been used to give rise to such error. We call these patterns “offenders of nd”. Two types of offenders are de- fined:

(1) a complete conflict offender which is a subset of nd; and

(2) a partial conflict offender which contains a part of terms of nd. The basic idea of updating patterns is to remove complete conflict offenders from d-patterns firstly. For partial conflict offenders, we re-shuffle their term supports in order to reduce the effects of noise documents. The main process of inner pattern evolution is implemented by the algorithm IPEvolving (see Algorithm 1). The pattern composition operation $\oplus$ in Step (6) is used to compose updated normal forms together and is defined in [11]. The task of Re-Shuffling is to tune the support

1509

```
Input: A training set $D = D^+ \cup D^-$; and a set of d-patterns NDP.
Output: A set of term-support pairs np.
np ← ∅;
foreach negative document $nd \in D^-$ do
    if weight of $nd \geq$ threshold then
        Re-shuffling term supports in NDP.
        foreach pattern $p \in NDP$ do
            $np \leftarrow np \oplus p$;
        end
    end
```

Algorithm 1: IPEvolving

Table 1: Comparison of all methods on the first 50 topics.

Method	top-20		b/p	MAP	Fβ=1	IAP	
PTM(IPE)		0.493		0.429	0.441	0.440	0.466
Sequential ptns	0.401		0.343	0.361	0.385	0.384	
Freq. itemsets		0.412		0.352	0.361	0.386	0.384
Rocchio	0.416		0.392	0.391	0.408	0.418	
Prob		0.407		0.381	0.379	0.396	0.402
TFIDF		0.321		0.321	0.322	0.355	0.348
BM25		0.434		0.399	0.401	0.410	0.422
SVM		0.447		0.409	0.408	0.421	0.434

distribution of terms within a d-pattern. As a result, the complete conflict offenders (d-patterns) are removed since all elements within the d-patterns are held by the negative documents indicating that they can be discarded for pre- venting interference from these possible “noises”.

IV. Evaluationanddiscussion

In this study we use Reuters Corpus Volume 1 (RCV1) text collection to evaluate the proposed approach. For di- mensionality reduction, stopword removal is applied and Porter algorithm [8] is selected for suffix stripping. Terms with term frequency equaling to one are discarded. Sev- eral standard measures based on precision and recall are used. The 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, breakeven point (b/p), Fβ- measure [3], Interpolated Average Precision (IAP) [2] and Mean Average Precision (MAP) are also used for performance evaluation. The results of overall comparisons are presented in Table 1. The proposed approach PTM(IPE) archives an out- standing performance for text mining by comparing with the up-to-date pattern mining-based methods and the well- known term-based methods, including the state-of-the-art BM25 and SVM [9] models. The promising results can be explained in that the use of pattern taxonomies in PTM integrates well with the advantages of terms and phrases. Moreover, the inner pattern evolution strategy provides an effective evaluation for estimating each term’s significance in the hypothesis space based on not only the term’s statistical properties but also the pattern’s associations in the pattern taxonomies.

V. Conclusions

Many data mining techniques have been proposed in the last decade, including association rule mining, frequent item- set mining, sequential pattern mining and closed pattern mining. However, utilizing the discovered knowledge (or patterns) is difficult and ineffective. The reason is that a useful long pattern is of high specificity but short in support. However, not all frequent short patterns are useful. Hence, inadequate use of patterns derived from data mining techniques leads to the ineffective performance. In this research work, an effective pattern taxonomy mining model has been proposed, aiming to overcome the aforementioned problem. The experimental results show that the proposed model out- performs not only other pure pattern mining-based methods, but also term-based models including the state-of-the-art BM25 and SVM.

References

[1]. R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1999.

[2]. W. Lam, M. E. Ruiz, and P. Srinivasan. Automatic text categorization and its application to text retrieval. IEEE Transactions on Knowledge and Data Engineering, 11(6):865–879, 1999.

[3]. D. D. Lewis. Evaluating and optimizing automous text classification systems. In Proceedings of SIGIR, pages 246–254, 1995.

[4]. Y. Li, W. Yang, and Y. Xu. Multi-tier granule mining for representations of multidimensional association rules. In Proceedings of ICDM, pages 953–958, 2006.

[5]. Y. Li, C. Zhang, and J. R. Swan. An information filtering model on the web and its application in jobagent. Knowledge-based Systems, 13(5):285–296, 2000.

[6]. Y. Li and N. Zhong. Interpretations of association rules by granular computing. In Proceedings of ICDM, pages 593–596, 2003.

[7]. Y. Li and N. Zhong. Mining ontology for automatically acquiring web user information needs. IEEE Transactions on Knowledge and Data Engineering, 18(4):554–568, 2006.

[8]. M. F. Porter. An algorithm for suffix stripping. Program, 14(3):130–137, 1980.

[9]. S. Robertson and I. Soboroff. The trec 2002 filtering track report. In TREC 2002 (URL:trec.nist.gov/ pubs/ trec11/ papers/ OVER.FILTERING.ps.gz), 2002.

[10]. S. Scott and S. Matwin. Feature engineering for text classification. In Proceedings of ICML, pages 379–388, 1999.

[11]. S-T. Wu, Y. Li, and Y Xu. Deploying approaches for pattern refinement in text mining. In Proceedings of ICDM, pages 1157–1161, 2006.

[12]. S-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen. Automatic pattern-taxonomy extraction for web mining. In Proceedings of WI, pages 242–248, 2004.

[13]. Y. Xu and Y. Li. Generating concise association rules. In Proceedings of CIKM, pages 781–790, 2007.