International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 20 Issue 3 – MARCH 2016.

EFFECTIVE CLASSIFICATION OF TEXT AND IMPROVING LEARNING EXPERIENCE

K.Surya, K. Kavitha, G.Joy princy

[#1]*Assistant professor/CSE, Dhanalakshmi college of Engineering*
[*2]*B.E.-Student, Dhanalakshmi college of Engineering*
[*2]*B.E.-Student, Dhanalakshmi college of Engineering*

suryaadce@gmail.com
kavithakrish95@gmail.com
joyprincy0911@gmail.com

***Abstract*--It is a large confront to assurance the excellence of exposed significance features in text ID for describing user preference because of big scale terms and data pattern. The majority existing well-liked text mining and categorization methods have adopt term-based approaches. However, they have all suffer from the evils of polysemy and synonymy. Over the years, there has been often held the theory that pattern-based methods should execute better ones in telling user preference; yet, how to efficiently use large scale patterns remains a hard problem in text removal.To make a get through in this difficult issue, this document presents an innovative model for sense feature discovery. It discover both optimistic and unenthusiastic patterns in book ID as higher level covering and deploy them over low-level features (terms).It also classify conditions into category and update term weights based on their specificity and their distributions in patterns. considerable experiment using this representation on RCV1, TREC topic and Reuters-21578 show that the future model drastically outperforms both the state-of-the-art term-based methods and the outline based methods.**

I. INTRODUCTION

The purpose of significance characteristic detection (RFD) is to locate the useful features obtainable in text ID, counting both pertinent and immaterial ones, for telling text removal consequences. This is a primarily demanding task in contemporary in order psychoanalysis, from both an experiential and hypothetical perspective [33], [36].

This involvedness is also of central awareness in many net modified application, and has received attention from researchers in Data removal, engine education, in order recovery and Web cleverness Communities [32].There are two demanding issues in using pattern mining techniques for judgment significance features in both relevant and irrelevant papers [32]. The first is the low-support difficulty. Given a topic, long patterns are typically more specific for the topic, but they usually appear in ID with low support or frequency.

If the smallest amount support is decreased, a lot of loud patterns can be open.The succeeding issue is the delusion problem, which means the measures (e.g., "hold up" and "selfassurance") used in prototype mining turn out to be not suitable in using patterns for solving problems. For example, a highly common pattern (usually a little outline)may be a universal pattern since it can be often used in both pertinent and immaterial documents. Hence, the hard difficulty is how to use discovered prototype to precisely heaviness useful features.There are several existing methods for solving the two demanding issues in text removal. Pattern classifi0cation removal (PTM) models have been future [59], [60], [70], in which removal closed chronological pattern in text paragraph and deploy them over a term space to weight usefulfeatures.Concept-based model (CBM) [50], [51] has also been prospect to find out concepts by using usual words processing (NLP) method.It future verb-argument structure to find concepts in verdict. These prototype (or concepts) based approach have shown an significant development in the efficiency [70]. However, fewer significant development are made compare with the best term-based method because how to effectively integrate patterns in both pertinent and irrelevant papers is still an open problem.Over the years, people have urbanized many mature term-based techniques for position papers, information filter and text categorization [37], [39], [44]. freshly, several hybrid approaches were proposed for text classification.To learn term features within only relevant papers and unlabelled documents, paper [27] used two term-based models. In the first stage, it utilized a Rocha classifier to extract a set of dependable immaterial ID from the unlabeled set. In the second stage, it built a SVM classifier to classify text papers. A two-stage model was also planned in [34], [35], which proved that the incorporation of the rough examination (a term-based model) and prototype classification mining is the best way to design a two-stage model for in order filtering systems.For many years, we have experiential that many terms with better weights are more universal because they are likely to be regularly used in both related and unrelated documents[32]. For example, word "LIB" may be more frequently used than word "JDK"; but "JDK" is more definite than "LIB" for recitation "Java indoctrination Languages"; and "LIB" is more general than"JDK"because"LIB" is also regularly used in other indoctrination languages like C or C++. Therefore, we advocate the consideration of both terms' distributions and specificities for relevance feature detection .Given a topic, a term's specificity describes the extent to which the term focus on the topic that users want [33].

However, it is very difficult to measure the specificity of terms because a term's specificity depends on users' perspective of their in order needs [55]. We future the first meaning of the specificity in [30], [31], which intended the specificity score of a term based on its looking exposed positive and negative patterns. However, this meaning necessary an iterative algorithm (three loops) in arrange to weight terms precisely.In order to make a get through in family member to the two demanding issues, we proposed the first account of the RFD

model in [32]. In agreement with the distributions of terms in a preparation set, it provided a new definition for thespecificity occupation and used two experiential parameters to group terms into three categories: "optimisticspecific terms", "general terms",and"unenthusiastic specific terms". Based on these definitions, the RFD model can accurately evaluate term weights according to both their specificity and their distributions in the higher level features, where the higher level skin include both positive and unhelpful patterns.The term cataloging method proposed in [32] requires physically setting two observed parameter according to difficult sets. In this paper, we continue to develop the RFD model, and experimentally prove that the proposed specificity job is reasonable and the term classification can be effectively approximated by a feature clustering method.We also design a total approach for evaluating the future models. In addition, we conducted some new experiments by using six new descending windows to adaptively update the preparation sets and also applying the RFD model for binary text categorization to test the heftiness of the proposed model..This paper proposes an ground-breaking technique for finding and classifying low-level terms based on both their appearances in the higher-level skin (patterns) and their specificity in a preparation set. It also introduces a method to select immaterial papers (so-called offenders) that are closed to the extracted features in the relevant documents in order to efficiently revise term weights. Compared with other methods, the compensation of the proposed model include:_ effectual use of both related and unrelated feedback to find useful features; and _ Integration of both word and pattern features together quite than using them in two separated stages.To give good reason for these claims for the proposed approach, we conducted substantial experiments on standard data collections ,namely, the Reuters Corpus quantity 1 (RCV1), TREC filtering evaluator topics, the records of Congress Subject heading (LCSH) ontology and Reuters-21578. We also used five events and the t-test to evaluate these experiments.The results show that the future specificity function is sufficient, the cluster method is effective and the proposed model is robust. The results also show that the proposed model considerably outperforms both the state-of -the- art term-based methods underpinned by Okapi BM25, Rocchio and verbal communication models, SVM and the pattern-based methods on most measures.The rest of this paper is organized as follows.Section 2 introduces a detailed overview of the related works. Section 3 reviews the concept of skin tone in text papers. Section 4 discusses the RFD model. Section 5 proposes a new feature clustering method based on the specificity function. To evaluate the presentation of the proposed model, we conduct considerable experiments on LCSH, RCV1, TREC filtering topics and Reuters-21578. The experiential results and discussion are reported in Section 6, follow by final comments in the last part.

II. RELATED WORK

Feature assortment is a method that selects a subset of features from data for modeling systems (seehttp://en.wikipedia.org/wiki/Feature_selection). Over the existence a variety of feature assortment methods (e.g., Filter, covering, Embedded and Hybrid approaches, and unsupervised or semi-supervised methods) have been future in various fields [6], [9], [17], [54], [69].Feature selection is also one of important steps for text classification and information filtering[1], [5], [47] which is the task of assigning documents to predefined classes. To date, many classifiers, such as Naïve Bayes, Rocchio, kNN, SVM and Lasso deteriorating [16], [26],[27], [28], [37], [62], [66] have been urbanized, in addition many believe that SVM is also a talented classifier [13].The categorization evils include the single class and multi-class problem. The most ordinary solution [71] to the multi-class problem is to go moldy it into some self-government binary classifiers, where a dual one is assigned to one of two predefined classes (e.g., pertinent category or immaterial category). Most conventional text feature selection methods used the bag of words to select a set of features for the multi-class problem [13]. There are some feature selection criteria for text classification, including document frequency (DF), the global IDF, in order gain, mutual information (MI), Chi-Square (x2) and term strength [1],[29], [37], [45], [67].In this paper we focus on pertinent feature selection in text ID. Relevance is a big research issue [25], [32],[65] for network search, which discusses a papers relevance to a user or a query. However, the traditional feature selection methods are not effective for selecting text features for solving significance issue because significance is a single class problem [13]. The imaginative way of feature assortment for relevance is based on a feature weighting function. A feature weighting function indicates the degree of information represented by the feature occurrence in a document and reflects the relevance of the feature. The popular term-based ranking models include tf*idf based techniques, Rocchio algorithm, Probabilistic models and Okapi BM25 [4], [24],[37], [44]. lately, one of the significant issues for multimedia data is the identification of the optimal feature set without any joblessness [69]; however, the challenging issue for text feature selection in text papers is the identification of which format or where the applicable features are in a text document because of the large amount of noisy in order in the document [2]. Text facial appearance can be simple structures (words), intricate linguistic structures or statistical structures. We mainly talk about three complex structures below for selecting relevant features: n-grams, concepts and patterns.

III. DEFINITIONS

For a given topic, the goal of relevance feature finding in text documents is to find a set of useful features, including pattern, terms and their weights, in a education set D, which consists of a set of relevant documents, Dþ, and a set of unrelated papers, D_. In this paper, we assume that all text papers, d, are split into paragraphs, PSðdÞ. In this part, we introduce the basic definitions about patterns and the deploying method. These definitions can also be found in [32], [34], [59].

3.1 Frequent and Closed Patterns

Let T1 ¼ ft1; t2; . . . ; tmg be a set of terms (or words) which are extract from Dþ, and termset X be a set of terms. For a given text d, coverset ðXÞ is called the covering set of X in d, which include all paragraph dp 2 PSðdÞ such that X _ dp, i.e., coversetðXÞ ¼ fdpjdp 2 PSðdÞ;X _ dpg. Its absolute support

is the number of occurrences of X in PSðdÞ, that is supaðXÞ ¼ jcoversetðXÞj. Its relative support is the fraction of the paragraphs that contain the pattern, that is, supr ðXÞ ¼ . A term set X is called a recurrent pattern if its supa (or supr) _ min sup, a given minimum support.It is obvious that a termset X can be mapped to a set of paragraphs coversetðXÞ. We can also map a set of paragraphs Y _ PSðdÞ to a termset, which satisfies Termset(Y) = {t|∀dp € Y=>t€dp}. A pattern X (also a termset) is called closed if and only if . Let X be a closed pattern. We have supaðX1Þ < supaðXÞ (1)for all patterns X1 _ X. All closed patterns can be structured into a pattern taxonomy by using the subset (or called is-a) relation [59].

3.2 Deploying Higher Level Patterns on Low-Level Terms

For term-based approach, weighting the usefulness of a given term is based on its appearance in ID.However, for pattern-based approaches, weighting the usefulness of a given term is based on its exterior in exposed patterns. To improve the efficiency of the pattern taxonomy mining, an algorithm, SP MiningðDþ;min supÞ [60], was proposed (also used in [34], [59]) to find closed sequential pattern for all papers 2 Dþ, which used the wellknown Apriori property to reduce the searching space. For all relevant papers di 2 Dþ, the Spinning algorithm discover all closed sequential patterns, SPi, based on a given min sup. We do not want to repeat this algorithm here because it is not the particular focus of this study.Let SP1, SP2; :::; SPjDþj be the sets of discovered closed Sequential pattern for all documents di 2 Dþði ¼ 1; . . . ; nÞ, where n ¼ jDþj. For a given term t, its d_support (deploying support, called weight in this paper) in discovered patterns can be described as follows:

$$d\_sup(t, D^+) = \sum_{i=1}^{n} sup_i(t) = \sum_{i=1}^{n} \frac{|\{p|p \in SP_i, t \in p\}|}{\sum_{p \in SP_i} |p|},$$

IV. RFD MODEL

In this section, we introduce the RFD reproduction for significance feature detection, which describes the relevant features in relation to three groups: optimistic specific terms, general terms and negative specific terms based on their appearances in a training set. We first discuss the concept of "specificity" in terms of the relative "specificity" in training datasets and the absolute "specificity" in domain ontology. We also present a way to understand whether the proposed relative" specificity" is reasonable in term of the absolute "specificity". Finally, we introduce the term weighting method in the RFD model. 4.1

4.1 Specificity Function

In the RDF model, a term's specificity (referred to as relative specificity in this paper) is defined [32] according to its exterior in a given teaching set. Let T2 be a set of terms which are extracted from D_ and T ¼ T1 [T2. Given a term t 2 T, its coverageþ is the set of relevant documents that contain t, and its coverage_ is the set of irrelevant documents that contain t. We assume that the terms frequently used in both pertinent documents and irrelevant documents are general terms. Therefore, we want to classify the terms that are more frequently used in the relevant documents into the optimistic specific category; the terms that are more frequently used in the irrelevant documents are classified into the negative specific category. Based on the above analysis, we defined the specificity of
a given term t in the training set D ¼ Dþ [D_ as follows:

$$spe(t) = \frac{|coverage^+(t)| - |coverage^-(t)|}{n},$$

where coverage þðtÞ ¼ fd 2 Dþjt 2 dg, coverage_ðtÞ ¼ fd 2 D_jt 2 dg, and n ¼ jDþj. speðtÞ > 0 means that term t is used more frequently in relevant documents than in irrelevant documents.

V. TERM CLASSIFICATION

RFD uses both exact skin (e.g., Tþ and T_) and general features (e.g., G). Therefore, the key investigate question is how to find the best partition (Tþ, G, T_) to efficiently classify relevant documents and immaterial documents. For a given set of features, however, this question is an N-P hard problem because of the large number of possible combinations of groups of features. In this section we propose an estimate approach, and proficient algorithms to refine the RFD model.

5.1 An Approximation Approach

The best partition (Tþ, G, T_) is used to clearly differentiate irrelevant ID from relevant ones. Assume that we have two characteristic functions f1, and f2, on all terms, such that f1ðtÞ is the near average weight of t for all relevant papers, and f2ðtÞ is the approximate average weight of t for all irrelevant documents. Therefore, the best partition (Tþ, G, T_) can maximize the following integration:
The above discussion motivates us to find adequate u1and u2 to make positive specific features move far awayf rom negative specific features. If we view the terms that have the same specificity score as a cluster and use the spe
function as the distance function, the new solution is to find three groups that can clearly divide the terms into three categories. Based on the above psychoanalysis, we can develop a cluster method to group terms into three categories automatically for each topic by using the specificity purpose. In the beginning, we allocate terms that appear only in irrelevant documents into the negative specific category T_. For the remaining terms, we initially view each term ti as a single cluster ci. We also represent each cluster ci using an interval ½minspeðciÞ; maxspeðciÞ_, where minspeðciÞ is the smallest spe value of elements in ci, and maxspeðciÞ is the largest spe value of the elements in ci. Let ci and cj be two clusters.

$$min\{|max_{spe}(c_i) - min_{spe}(c_j)|, |max_{spe}(c_j) - min_{spe}(c_i)|\}.$$

A bottom-up come up to is used to merge two cluster if they have the lowest amount difference.

VI. EVALUATION

This section discuss the testing surroundings, and reports the new consequences and the discussions. It also provides recommendations for offender selection and the use of precise terms and general terms for describing user in turn needs. The proposed model is a supervised approach that needs a education set including both relevant documents and unrelated documents.

6.1 Data

We used two well-liked data sets to test the proposed model: Reuters Corpus Volume 1, a very large data collection; and Reuters-21578, a small one. RCV1 include 806,791 papers that cover a broad field of issues or topics. TREC(2002) has urbanized and provided 50 reliable judge topics [44] for RCV1, aiming at tough robust information filtering systems. These topics were evaluated by human assessors at the National Institute of principles and Technology (NIST) [52]. For each topic, a subset of RCV1 documents is alienated into a training set and a testing set.

VII. CONCLUSIONS

The examine proposes an alternative approach for relevance feature discovery in text documents. It presents a method to find and classify low-level features based on both their appearances in the higher-level pattern and their specificity.
It also introduces a method to select irrelevant documents for

weighting features. In this paper, we continued to enlarge the RFD model and experimentally prove that the proposed specificity function is reasonable and the term arrangement can be effectively approximated by a feature clustering method.The first RFD model uses two experiential parameters to set the state line between the categories. It achieves the predictable performance, but it requires the physically testing of a large number of dissimilar values of parameter. The new model uses a feature clustering technique to mechanically group terms into the three categories. Compared with the first model, the new model is much more well-organized and achieved the satisfactory concert as well.This paper also includes a set of experiments on RCV1 (TREC topics), Reuters-21578 and LCSH ontology. These experiment exemplify that the proposed model achieves the best presentation for comparing with term-based baseline models and pattern-based baseline models. The results also show that the term categorization can be effectively approximated by the proposed feature clustering method, the proposed specifics function is reasonable and the proposed models are robust. This paper demonstrates that the future model was thoroughly tested and the results prove that the proposed model is statistically important. The paper also proves that the use of insignificance feedback is significant for improving the presentation of significance feature finding models. It provides a promising line of attack for developing effective text taking out models for relevance feature discovery based both positive and negative feedback.

ACKNOWLEDGMENTS

I. This paper was partially supported by Grant DP140103157 from the Australian Research Council (ARC detection Project). Y. Li is the matching author.

REFERENCES

[1] M. Aghdam, N. Ghasem-Aghaee, and M. Basiri, “Text feature selection using ant colony optimization,” in Expert Syst. Appl., vol. 36, pp. 6843–6853, 2009.

[2] A. Algarni and Y. Li, “Mining specific features for acquiring user information needs,” in Proc. Pacific Asia Knowl. Discovery Data Mining, 2013, pp. 532–543.

[3] A. Algarni, Y. Li, and Y. Xu, “Selected new training documents to update user profile,” in Proc. Int. Conf. Inf. Knowl. Manage., 2010, pp. 799–808.

[4] N. Azam and J. Yao, “Comparison of term frequency and document frequency based feature selection metrics in text categorization,” Expert Syst. Appl., vol. 39, no. 5, pp. 4760–4768, 2012.

[5] R. Bekkerman and M. Gavish, “High-precision phrase-based document classification on a modern scale,” in Proc. 11th ACM SIGKDD Knowl. Discovery Data Mining, 2011, pp. 231–239.

[6] A. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” Artif. Intell., vol. 97, nos. 1/2, pp. 245– 271, 1997.

[7] C. Buckley, G. Salton, and J. Allan, “The effect of adding relevance information in a relevance feedback environment,” in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1994, pp. 292–300.

[8] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, “Selecting good expansion terms for pseudo-relevance feedback,” in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 243–250.

[9] G. Chandrashekar and F. Sahin, “Asurvey on feature selection methods,” in Comput. Electr. Eng., vol. 40, pp. 16–28, 2014.

[10] B. Croft, D. Metzler, and T. Strohman, Search Engines: Information Retrieval in Practice. Reading, MA, USA: Addison-Wesley, 2009.

[11] F. Debole and F. Sebastiani, “An analysis of the relative hardness of Reuters-21578 subsets,” J. Amer. Soc. Inf. Sci. Technol., vol. 56, no. 6, pp. 584–596, 2005.

[12] J. Eisenstein, A. Ahmed, and E. P. Xing, “Sparse additive generative models of text,” in Proc. Annu. Int. Conf. Mach. Learn., 2011, pp. 274–281

[13] G. Forman, “An extensive empirical study of feature selection metrics for text classification,” in J. Mach. Learn. Res., vol. 3, pp. 1289–1305, 2003.

[14] Y. Gao, Y. Xu, and Y. Li, “Topical pattern based document modeling and relevance ranking,” in Proc. 15th Int. Conf. Web Inf. Syst. Eng., 2014, pp. 186–201.

[15] X. Geng, T.-Y. Liu, T. Qin, A. Arnold, H. Li, and H.-Y. Shum, “Query dependent ranking using k-nearest neighbor,” in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 115–122.

[16] A. Genkin, D. D. Lewis, and D. Madigan, “Large-scale Bayesian logistic regression for text categorization,” Technometrics, vol. 49,no. 3, pp. 291–304, 2007.

[17] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” in J. Mach. Learn. Res., vol. 3, no. 1, pp. 1157–1182, 2003.

[18] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000, pp. 1–12.

[19] Y.-F. Huang and S.-Y. Lin, “Mining sequential patterns using graph search techniques,” in Proc. Annu. Int. Conf. Comput. Softw. Appl., 2003, pp. 4–9.

[20] G. Ifrim, G. Bakir, and G. Weikum, "Fast logistic regression for text categorization with variable-length n-grams," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2008, pp. 354–362.

[21] N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2006, pp. 244–251.

[22] T. Joachims, "Transductive inference for text classification using support vector machines," in Proc. Annu. Int. Conf. Mach. Learn., 1999, pp. 200–209.

[23] T. Joachims, "Optimizing search engines using clickthrough data," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2002, pp. 133–142.