Mathematical Statistician and Engineering Applications
ISSN: 2326-9865
DOI: https://doi.org/10.17762/msea.v70i1.2303

Relevance Feature Discovery for Text Mining

Vikrant Sharma

Asst. Professor, Department of Computer Science, Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Article Info
Page Number: 225-233
Publication Issue:
Vol. 70 No. 1 (2021)

Article History
Article Received: 25 January 2021
Revised: 24 February 2021
Accepted: 15 March 2021

Abstract

Due to large size words also data patterns, it is difficult to ensure the quality of relevant characteristics that are found in text documents that describe user preferences. Most widely used text mining and classification techniques now in use have embraced term-based strategies. However, polysemy and synonymy issues have affected them all. The theory that pattern-based approaches should outperform term-based ones in performance in expressing user preferences has been often held throughout the years, however text mining still struggles with how to employ large-scale patterns successfully. This research introduces a novel methodology for relevance feature discovery to address this hard problem. It finds higher level features in text texts that are both positive and negative patterns and uses them instead of low-level features (terms). Additionally, it organised terms into categories and updates term weights according to the patterns and specificity of those distributions. Significant tests employing this model on the datasets RCV1, TREC themes, and Reuters-21578 reveal that it performs noticeably better than both the most advanced term-based approaches and pattern-based methods.

1. Introduction

Finding valuable characteristics in text documents, both relevant as well as irrelevant, for explaining text mining findings is the goal of relevance feature discovery (RFD). From both an empirical and a theoretical standpoint, this task in contemporary information analysis is exceptionally difficult. Professionals in the disciplines of online intelligence, information retrieval, machine learning, including data mining have paid attention to this issue since it is also crucial to many Web personalised applications. Employing pattern mining methodologies to uncover relevance characteristics in both relevant as well as irrelevant materials presents two difficult problems. The first issue is the lack of assistance. Long patterns are often more focused on a given topic, although they typically exist in papers with little or no support. Numerous noisy patterns can be found if the minimum support is lowered. The second difficulty is the misunderstanding problem, which results in the metrics (such as "confidence" and "support" employed in pattern mining) being unsuitable for employing patterns to solve issues. For instance, a pattern that appears frequently (and is typically brief) in both pertinent and unrelated materials may be a generic pattern.

The two complex problems in text mining have a number of established solutions. There are approaches for pattern taxonomy mining (PTM) that employ closed sequential patterns found in text paragraphs over a word space to weight valuable characteristics. Additionally, the concept-based model has been proposed as a way to find concepts using natural language processing (NLP) methods. In order to identify concepts in phrases, it suggested verb-argument frameworks. The efficacy of these pattern- or concept-based techniques has significantly increased. However, fewer noteworthy advancements are made in comparison to the best term-based method because it is still unclear how to efficiently incorporate patterns in both pertinent and irrelevant documents.

Many sophisticated term-based methods for text categorization, information filtering, including document ranking have been developed over the years. For text categorization, a number of hybrid techniques have recently been put out. Paper used two term-based models to learn term features from only relevant documents and unlabeled documents. It utilized a Rocchio classifier in the first step to separate a collection of trustworthy irrelevant texts from the unlabeled set. It created an SVM classifier in the second step to categorise text documents. It was also shown in a two-stage model that combining rough analysis (a term-based model) with pattern taxonomy mining is the best way to build a two-stage model for information filtering systems.

In order to classify words into three categories—"positive specific terms," "general terms," along with "negative specific terms"—this research presents a novel definition of the specificity function for a term using two empirical factors. When both positive and negative patterns are present in the higher level features, the RFD framework will appropriately evaluate term weights based on the specificity overall distributions of the higher level features. According to experiments, combining particular phrases with some broad terms is the most effective strategy to increase the performance of relevance feature discovery.

In addition to reporting the experimental findings and debates, this section also addresses the testing setting. Additionally, it offers suggestions for choosing offenders and describes user information needs using both specific and general terms. The suggested model is a supervised method that requires a training set made up of both pertinent and pointless materials.

Reuters Corpus Volume 1, a sizable data collection, and Reuters-21578, a smaller one, were employed as two well-known data sets to evaluate the suggested model. RCV1 has 806,791 papers that deal with a wide range of problems or subjects.

The suggested technique, known as the relevant feature discovery model, comprises three main steps: term categorization, feature discovery and deployment, and term weighting. In the training set, it initially identifies both constructive and destructive patterns and phrases. Additionally, it divides terms into three groups based on parameters or the Algorithm F Clustering. Finally, it uses Algorithm W Feature to calculate the word weights.

The robustness of a model is used in this study to explore the qualities that describe its ability to function correctly even when the training data or the application environment are modified. If a model continues to perform satisfactorily even after having The use of its environment or training sets transformed, we refer to that model as being robust. Due to Reuters-21578's testing set will shrink if we increase training sets, we only use RCV1 for this evaluation? In order to expand the training sets for the revised training sets, we employed six loops for each topic, with each loop utilising a sliding window that included 25 documents drawn at random from the testing set. The equivalent testing set for the 25 papers was likewise eliminated.

Since finding relevant knowledge is the major objective of relevance feature discovery, we think that positive feedback is more constructive than negative feedback. To improve the efficiency of relevance feature identification, Negative feedback does, in our opinion, offer some crucial information that may help to determine the boundary between relevant and irrelevant information. Due to the overwhelming volume of negative information, the majority of irrelevant papers are not closed to the topic at hand, which is the obvious issue with employing them.

2. Literature Survey

In order to increase the efficiency of using and updating found patterns for locating pertinent and fascinating information, this study provides a novel and effective pattern discovery approach that incorporates the processes of pattern deploying and pattern evolving. Data mining is a crucial phase in the process of knowledge discovery, which is the process of extracting information from huge datasets. The goal of this study is to design a knowledge discovery model that can be applied to text mining to efficiently exploit and update the patterns that have been found. In order to address the problem of accurately locating knowledge in text documents, IR offered term-based approaches, however these methods are hampered by polysemy and synonymy. Low frequency and misunderstanding are two key problems with pattern-based methods. Misinterpretation occurs when the metrics employed in pattern mining are inappropriate, but low frequency is typically a broad pattern. This study proposes a powerful pattern discovery method to precisely assess the weights of practical aspects (knowledge) in textual data. Prior to evaluating term weights in accordance with the distribution of terms in the identified patterns, it first calculates the specificities of patterns that have been discovered. In order to identify ambiguous patterns and lessen their impact on the low-frequency problem, it additionally takes into account the influence of patterns from negative training instances [1].

In order to assess the degree of variation in the distributions of a phrase between a given category and the full corpus, this study suggests a novel method based on the t-test. Extensive comparison tests utilising three classifiers on two text corpora reveal that the novel method is on par with or significantly superior to state-of-the-art feature selection techniques in regards to macro-F1 and micro-F1. 19,905 papers make up the extensively utilised benchmark collection known as the Reuters corpus. In the concept space, each document is represented by a vector. Term weighting is determined using conventional methods, and the vector is then normalised to have a length of one unit. Our technique beats IG and MI approaches considerably on imbalanced text corpus, and is equivalent to or significantly superior to cutting-edge 2 and ECE with regard to of macro-F1 and micro-F1. Chi-Square Statistic (2), Information Gain (IG), Mutual Information (MI), plus Expected Cross-Entropy (ECE) are the four techniques employed. The assumptions behind the 2 statistic do not apply to the majority of textual analyses. MI is impacted by the marginal probabilities of terms, and 2 is unreliable for low-frequency words. Expected Cross-Entropy (ECE) solely takes into account terms that are included in a document and ignores phrases that are missing. Although IG has less discriminating power in TC tasks, it can be described as the variation among the original information requirement along with the new requirement [2].

This study suggests a different text-generation model that incorporates a log-frequency deviation model from a steady background distribution. This method has two major benefits: it can enforce sparsity and combine generating aspects by just adding them in log space, doing so without the need of latent switching variables. Standard Dirichlet-multinomial generative models waste training data by learning a certain probability distribution over the whole lexicon. In contrast to the Dirichlet-multinomial, the Sparse Additive Generative model (SAGE) is suggested in this study as a generative model of text. SAGE increases prediction performance and resilience to sparse training data by modelling the difference in log-frequencies from a background lexical distribution. Furthermore, it enables the construction of multifaceted latent variable models by merely adding SAGE component vectors. SAGE may be used in a variety of generative models and is designed as an optional substitute for the Dirichlet-multinomial [3].

We offer a technique for categorising documents that makes use of lazy learning from labelled terms. We measure the near sufficiency characteristic and demonstrate the viability of manual phrase labelling. The PBC system fared better than its predecessor in terms of coverage and precision when we used it to classify job titles. Current applications include LinkedIn's ad targeting product, and more are being created. Crowdsourcing is employed to swiftly annotate massive amounts of data and assess the categorization outcomes. Nevertheless, crowdsourcing outcomes sometimes have poor quality. In the method known as phrase-based multilabel classification, a collection of commonly used phrases is extracted, each document is mapped onto a subset of phrases, and each document is then assigned to a subset of classes based on the categorization of the phrases. When it comes to multi-label categorization, human annotators do better than machines because they naturally know how many classes a data instance belongs to and if it should be placed in any particular class [4].

The hypothesis that the most common words in pseudo-feedback texts are helpful for retrieval is reexamined in this study. It demonstrates that many expansion words discovered using conventional methods have no relation to the query and are detrimental to retrieval. In order to forecast the utility of expansion terms, we suggest integrating a term categorization method. Utilising term classification can significantly increase retrieval effectiveness, according to experiments on three TREC collections. SVM will be utilised for term classification, which includes various more factors including term closeness in addition to the term distribution criteria from earlier work. At the very least, the following benefits come with this strategy: 1) Instead of choosing expansion words only on the basis of term distributions and other factors that are tangentially linked to retrieval effectiveness, expansion terms are now chosen based on their ability to affect retrieval effectiveness. 2) The process of classifying terms can organically incorporate many criteria, offering a framework for absorbing diverse types of evidence. We compare our technique to the conventional methods and test it using three TREC datasets [5].

3. Proposed System

As previously indicated, pattern taxonomy models (PTM) make use of text documents' closed sequential patterns to get beyond the limitations of conventional term-based techniques. The main PTM difficulty, however, is how to efficiently deal with a wealth of identified patterns in order to extract precise characteristics. Many of the patterns that have been found are meaningless, but some of them may also contain broad information (such as words or phrases) regarding the user's topic. These patterns are distracting and frequently limit effectiveness. This chapter introduces a unique data mining approach for extracting user preferences or information needs from text sources. This approach makes use of pattern taxonomy mining to extract significant semantics data from a feedback collection of pertinent documents. To lessen the consequences of the noisy information that pattern mining gathered after then, a novel post-mining technique called pattern cleaning is used for relevance feature discovery.

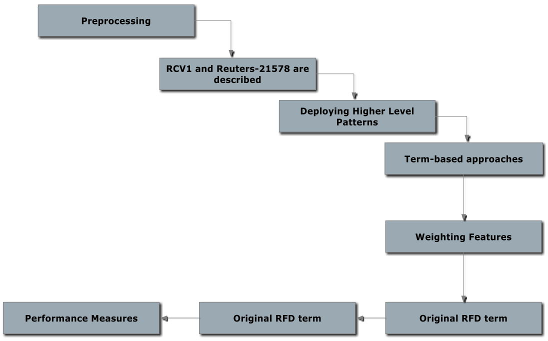


Fig 1: System Architecture

Closed sequential patterns (also known as positive patterns, or positive patterns for short) that are extracted from positive (related) publications for a specific topic collect prices of useful information for characterising user information demands. The quality of the extracted characteristics might, however, be readily impacted by the abundance of useless and irrelevant information present in the feedback papers. When a user has a particular demand, noisy information cannot be handled via closed pattern mining. For instance, specialised long patterns have low support whereas short patterns with strong support typically provide broad information for a certain topic. Getting rid of sounds from the discovery process is what pattern cleaning aims to do. The basic goal of pattern cleaning is to use irrelevant information to hone the knowledge that is pertinent to a certain topic. However, since they may frequently be gathered from other topics, using all negative documents may not be interesting and may increase noise. To overcome the aforementioned problem, we present the concept of offenders in this study. A bad document that is more similar to good ones is called an offender.

It is difficult to find beneficial characteristics in information retrieval and data mining to aid people looking for pertinent information. The most useful information source for learning about users' information demands is user relevance feedback. Nevertheless, excessive noise present in the real-world feedback data can negatively impact the quality of the features that are extracted. In order to lessen the impact of noisy features recovered by frequent pattern mining, the main research question in this thesis is how to extract relevant knowledge from user relevance feedback. An innovative pattern-based method for finding relevant features is presented in this thesis. We present the idea of pattern cleaning, which involves utilizing chosen non-relevant samples to improve the quality of frequently found patterns in pertinent documents.

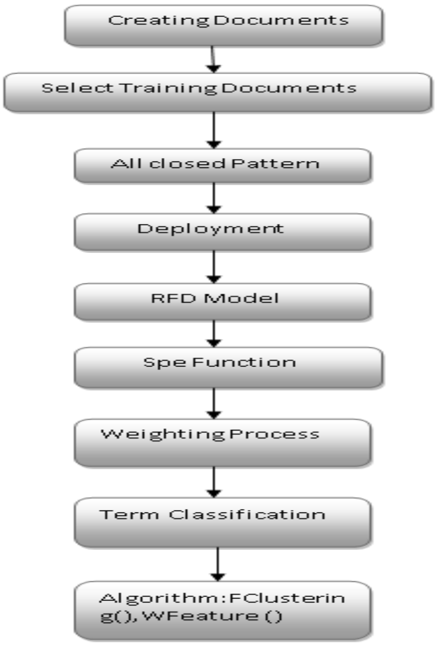


Fig 2: Flow Diagram

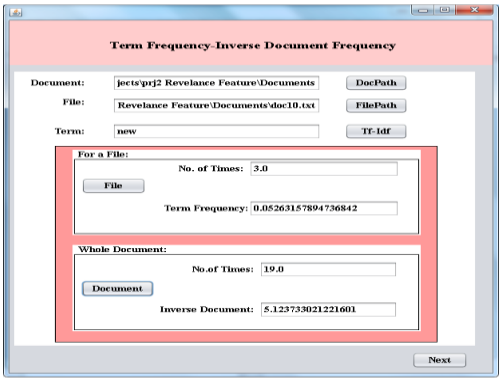
We demonstrate how the information from irrelevant samples may be used to increase the quality of certain characteristics and minimize noise in relevant documents in order to obtain correct information. The following benefits of the suggested strategy are listed:

- The fundamental premise of this study is that relevance features are utilized to define relevant documents alongside irrelevant materials are applied to ensure that extracted characteristics can be distinguished.
- Additionally, it offers suggestions for choosing offenders and describes user information needs using both specific and general terms.
- Positive and negative patterns are found in text texts as higher level features, and these features are deployed over low-level features.

4. Results

This study introduces a novel approach to relevance feature discovery that identifies both positive and unfavourable patterns in text texts as higher level features and applies them to words as low-level features. Significant tests employing this model on the datasets RCV1, TREC themes, and Reuters-21578 reveal that it performs noticeably better than both the most advanced term-based approaches and pattern-based methods. A unique data mining methodology uses pattern taxonomy mining to collect crucial semantics information in a feedback set of pertinent documents for gaining user information wants or preferences in text documents. For relevance feature discovery, a novel post-mining technique called pattern cleaning is used to lessen the consequences of noisy information that pattern mining has gathered.

Furthermore, the study presents a technique for identifying and categorising low-level characteristics based on the specificity and presence of these elements in higher-level patterns. Studies reveal that when compared to baseline models based on terms and baseline models based on patterns, the suggested model performs well. The study demonstrates the need of using irrelevance feedback to enhance relevance feature discovery model performance.



Term Frequency-Inverse Document Frequency

Document: jects\prj2 Revelance Feature\Documents

File: Revelance Feature\Documents\doc10.txt

Term: new

For a File:

No. of Times: 3.0

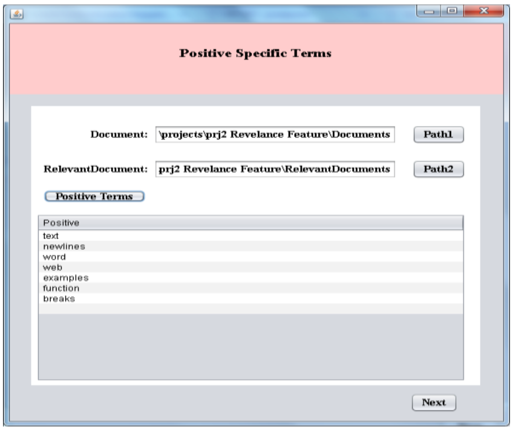
Term Frequency: 0.05263157894736842

Whole Document:

No.of Times: 19.0

Inverse Document: 5.123733021221601

Fig 3: Term Frequency-Inverse Document Frequency



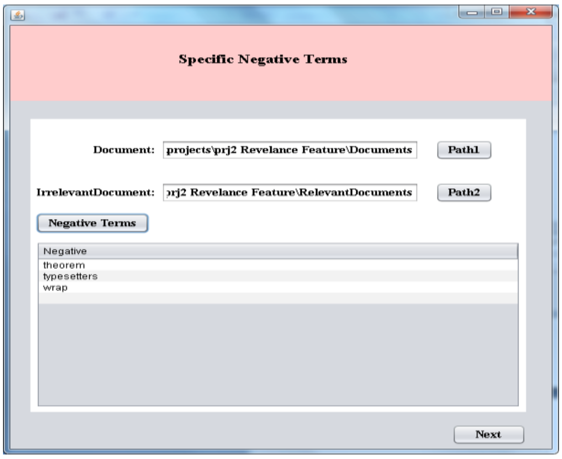
Positive Specific Terms

Document: \projects\prj2 Revelance Feature\Documents

RelevantDocument: prj2 Revelance Feature\RelevantDocuments

Positive
text
newlines
word
web
examples
function
breaks

Fig 4: Positive Specific Terms



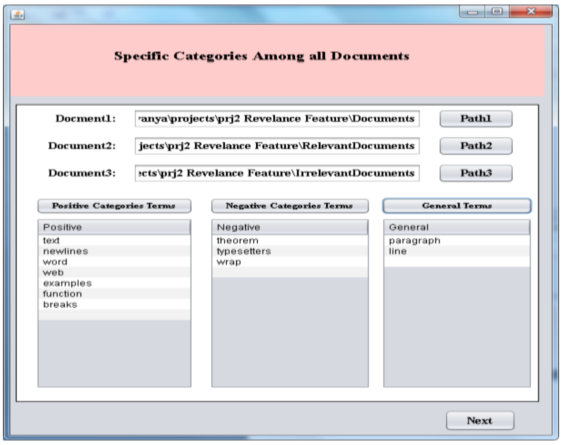
Specific Negative Terms

Document: projects\prj2 Revelance Feature\Documents

IrrelevantDocument: ›rj2 Revelance Feature\RelevantDocuments

Negative
theorem
typesetters
wrap

Fig 5: Specific Negative Terms



Specific Categories Among all Documents

Docment1: ‹anya\projects\prj2 Revelance Feature\Documents

Document2: jects\prj2 Revelance Feature\RelevantDocuments

Document3: ›cts\prj2 Revelance Feature\IrrelevantDocuments

Positive	Negative	General
text	theorem	paragraph
newlines	typesetters	line
word	wrap	
web		
examples		
function		
breaks		

Fig 6: Specific Category Among All Documents

5. Conclusion

The study suggests a different method for finding relevant features in text texts. It offers a technique for locating and categorising low-level elements based on both their specificity and how frequently they appear in higher-level patterns. Additionally, it introduces a technique for choosing pointless documents for weighting features. In this study, we further developed the RFD model and empirically demonstrated the validity of the suggested specificity function and the approximation of the word classification by a feature clustering technique. The first RFD model establishes the border between the categories using two empirical parameters. Although it performs as expected, a significant number of different parameter values must be manually tested. The new model automatically groups phrases into the three categories using a feature clustering approach. The new model is significantly more effective than the original model and also produced satisfying results. A series of experiments are also included in this publication.

These tests show that, when compared to baseline models based on terms and baseline models based on patterns, the suggested model performs well. The outcomes further demonstrate that the proposed feature clustering approach, recommended spefunction, and suggested models can all successfully approximate phrase categorization. This study shows that the suggested model was extensively examined, and the findings show that it is statistically significant. The study demonstrates the need of using irrelevance feedback to enhance relevance feature discovery model performance. It offers an excellent way for creating text mining models for finding relevant features according to both positive and negative feedback.

Reference

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,” inProc. 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,” inProc. 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, “A survey 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,” inJ. Mach. Learn. Res., vol. 3,pp. 1289–1305, 2003.
14. Y. Gao, Y. Xu, and Y. Li, “Topical pattern based document modelling 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,” inProc. 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.