ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 1, Issue 6, November 2015

Relevance Feature Discovery for Text Mining using Dynamic Approach

Salve Bhavna B. [1], Prof. N. V. Alone [2]

PG Student, Dept. Of CSE, R.H. Sapat College of Engineering, Nasik, Maharashtra, India[1]

Professor, Dept. Of CSE, R.H. Sapat College of Engineering, Nasik, Maharashtra, India[2]

ABSTRACT— Text mining techniques helps users to find useful information from a large amount of digital text documents on a Web or databases engines. It is therefore crucial that a good text mining model should retrieve the information that meets users' needs within a relatively efficient time frame. Traditional Information Retrieval (IR) has the same goal of automatically retrieving relevant documents as many as possible while filtering out non-relevant ones at the same time. It was intended to generate dynamic models to classify multiple topics in a collection of documents. A fundamental supposition for these approaches is that the documents in the collection are all about one topic. To guarantee the quality of discovered relevance feature in text documents for describing the user preferences is very crucial and challenging task because of large scale terms and data patterns. Most existing text mining and classification methods adopt term based approaches which suffer from the problem of polysemy and synonymy. We all believe in hypothesis that the pattern based methods performs better than term based ones in describing user preferences. The challenging issue is large scale patterns remains as hard problem in text mining. To deal with the above mentioned limitations and problems, proposed model presents the dynamic approach for Relevance Feedback Discovery by classifying terms into different categories dynamically and updating term weights and their distribution in patterns efficiently by improving the performance of text mining. The proposed model significantly outperform both Term based Methods and Pattern based methods. To evaluate the effectiveness of the proposed model TREC data collection, Reuters Corpus Volume 1 and Reuters-21578 are used.

KEYWORDS:- Text Mining, Text Feature Extraction, Text Classification, Hierarchical Agglomerative clustering, Term polarity, Term frequency.

I. INTRODUCTION

The reason for importance function development (RFD) is always to chose the helpful functions available in text message docs, as well as each relevant in addition to immaterial people, with regard to explaining text message exploration results. This can be a in particular complicated task inside modern details research, by each a good empirical in addition to theoretical viewpoint [16]. This matter is also involving main interest in many Web personalized software, in addition to possesses obtained attention by research workers inside Data Mining, Device Learning, Info Access in addition to Web Learning ability residential areas [12]. You will find 2 complicated troubles inside employing structure exploration procedures for locating importance functions inside each relevant in addition to immaterial docs . The first is your low-support difficulty. Provided a topic, very long designs usually are much more

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 1, Issue 6, November 2015

unique with the subject, however they generally include docs having reduced help or perhaps regularity. In the event the minimal help will be reduced, many boisterous designs can be discovered. The 2nd matter will be the misinterpretation difficulty, which suggests your procedures (e. grams. “support” in addition to “confidence”) found in structure exploration come to be not really suited inside employing designs with regard to dealing with issues. By way of example, a highly repeated structure (normally a brief pattern) can be a general structure given that it may be frequently employed inside each relevant in addition to immaterial docs. Therefore, the difficult difficulty will be the way to utilize discovered designs to correctly weight helpful functions.

There are various active options for dealing with the two complicated troubles inside text message exploration. Style taxonomy exploration (PTM) models have been suggested, in which, exploration finished sequential designs inside text message grammatical construction in addition to implementing these individuals spanning a time period room to weight helpful functions. Concept-based product (CBM) has been recently suggested to learn methods by using pure terminology finalizing (NLP) techniques. That suggested verb-argument constructions to get methods inside phrases. These types of structure (or concepts) dependent methods have shown an important improvement from the performance [7]. Nevertheless, a lot fewer important improvements are created in contrast to the most beneficial term-based procedure because the way to correctly assimilate designs inside each relevant in addition to immaterial docs remains a good available difficulty.

Over the years, men and women have developed many adult term-based procedures for ranking docs, details selection in addition to text message category [14]. Recently, many hybrid methods had been suggested with regard to text message category. To know time period functions within just relevant docs in addition to unlabelled docs, papers employed 2 term-based models. Within the 1st point, that employed any Rocchio classifier to get a few dependable immaterial docs from your unlabeled collection. Within the next point, that constructed any SVM classifier to classify text message docs. Some sort of two-stage product had been also suggested inside which usually demonstrated which the integration of the abrasive research (a term-based model) in addition to structure taxonomy exploration will be the easiest method to pattern any two-stage product with regard to details selection systems. For quite a while, we've witnessed that many conditions having more substantial dumbbells are definitely more general because they're prone to always be frequently employed inside each relevant in addition to immaterial docs [12]. By way of example, expression “LIB” may be more often employed than expression “JDK”; however “JDK” will be much more unique than “LIB” with regard to explaining “Java Selection Languages”; in addition to “LIB” will be much more general than “JDK” because “LIB” is also frequently employed inside some other encoding different languages just like G or C++. Thus, we advocate your consideration involving each terms’ distributions in addition to specificities with regard to importance function development. Provided a topic, any term’s specificity talks about your extent to that the time period focuses on the niche that will end users wish [13]. Nevertheless, it's very tough to measure your specificity involving conditions just because a term’s specificity will depend on users’ perspectives of these details require [15]. We suggested your 1st meaning of the specificity inside [10], [11], which usually computed your specificity ranking of an time period determined by the look inside discovered beneficial in addition to negative designs. Nevertheless, that meaning needed a good iterative protocol (three loops) inside get to weight conditions correctly.

II. BACKGROUND

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 1, Issue 6, November 2015

One interesting theory is invented by Scientists Y. Li, N. Zhong Y. Li, N. Zhong's in 2006 on "Mining Ontology for automatically acquiring web user information needs". In this paper they had presented a novel approach to provide satisfactory structures for mining web user profiles. Automatically discover ontology from dataset to build complete concept models for web user information need [2].

According to theory presented by Scientists S. Shehata, F. Karray, M. Kamel in paper "Enhancing text clustering using concept based". It suggests the model consist of concept based analysis of terms and a concept based similarity measure [7].

Some similar theories are enlisted as follows

In 2011 the paper "Deploying approaches for pattern Refinements in text mining" is invented. This paper proposed two pattern refinement methods to improve effectiveness of pattern based method. These methods deploy discovered pattern into feature space which is used to represent concept of document [5].

Another Approach is invented in 2012 as "Effective pattern discovery for text mining" by N. Zhong, Y. Li, and S.T. Wu is that The relevance of a document can be modelled by a pattern-based model[4].

The Z. Zhao, L. Wang, H. Liu, J. Ye, had researched in "Onsimilarity preserving feature selection" in the year 2013 & This paper propose a new "Similarity Preserving Feature Selection" framework, which not only encompasses many widely used feature selection criteria, but also naturally overcomes their common weakness in handling feature redundancy.[3]

Another related study was made by Y. Li, A. Algarni, M. Albathan, Y. Shen, M. A. Bijaksana in 2015 & This paper presents an innovative model for relevance feature discovery. It discovers both positive and negative patterns in text documents as higher level features and deploys them over low-level features.

Function selection is usually a technique which prefers a subset regarding characteristics by facts regarding modelling techniques (see http: //en.wikipedia. org/wiki/Feature selection). Over the years, a selection of element selection procedures (e. h., Filtration, Wrapper, Set along with Hybrid techniques, along with unsupervised or semi-supervised methods) are proposed in various areas [6], [9]. Function selection is also certainly one of crucial ways regarding word group along with facts selection [1], [5] and that is the position regarding assigning papers to predefined instruction. Currently, a lot of classifiers, like Naive Bayes, Rocchio, kNN, SVM along with Lasso regression [6] are formulated, also a lot of believe that SVM is also a encouraging classifier [13]. The actual group problems add the one type along with multi-class problem. The most typical remedy towards multi-class problem is always to decompose this into a few self-reliance binary classifiers, when a binary is assigned to 1 of 2 predefined instruction (e. h., pertinent type or immaterial category). Many conventional word element selection procedures applied the tote regarding phrases to pick out a collection of characteristics regarding the multi-class problem [13]. There are several element selection criteria regarding word categorization, as well as doc regularity (DF), the international IDF, facts obtain, shared facts (MI), Chi-Square (x2) along with phrase toughness [1].

In this report we all concentrate on pertinent element selection in word papers. Relevance is usually a huge analysis difficulty[2], [5] regarding World wide web seek, which looks at a papers relevance to some end user or even a question. On the other hand, the standard element selection

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 1, Issue 6, November 2015

procedures will not be efficient regarding selecting word characteristics regarding resolving relevance difficulty because relevance is usually a one type problem [13]. The actual efficient means of element selection regarding relevance will depend on a feature weighting purpose. An element weighting purpose implies the degree regarding facts symbolized from the element events within a doc along with reflects the relevance in the element. The most popular term-based standing versions consist of tf*idf primarily based approaches, Rocchio protocol, Probabilistic versions along with Okapi BM25 [4].

III. PROBLEM DEFINITION

Model for Relevance Feature Discovery discovers both positive and negative patterns in text documents as higher level feature and deploys them over low level features (term). The selection of dataset is static, which limits the size of data. This model generates only three clusters as positive, negative and general for estimating relevance of documents in given dataset. Proposed model use novel approach for dynamically adding new dataset and generates more than three clusters to make relevance feature discovery more effective and to improve the performance of text mining. To guarantee the quality of discovered relevance feature in text document, this model gives relevance of the documents with the user preferences using both tem based methods and pattern based methods. For a given topic, the model finds set of useful features including patterns ,terms and their weights in the training set.

In many web personalized applications while text mining task it is important to find useful features available in text documents including both relevant and irrelevant ones for describing text mining results. This is a challenging task in modern information analysis. There are two challenging issues. First, Low support problem, Given a topic long patterns are generally more specific for the topic but they usually appear in documents with low support or frequency. Second, Misinterpretation problem means the measures (support and confidence) used in pattern mining turn out to be not suitable in using patterns for solving problems. Hence the difficult problem is how to use discovered patterns to accurately weight useful features. The Relevance feature Discovery is breakthrough these problems. The RFD model can accurately evaluate term weights according to both their specificity and their distribution in the higher level features which include both positive and negative patterns. Proposed model is an innovative technique for finding and classifying low level terms based on both their appearances in patterns and their specificity in training set. It also introduces method to select irrelevant documents. Proposed model shows following advantages. One, Effective use of both relevant and irrelevant feedback to find useful features. Second, Integration of both term and pattern features together. Third, Permission to add new dataset dynamically. Fourth, Improved performance of the model.

.

IV. PROPOSED SOLUTION

To guarantee the quality of discovered relevance features in text documents for describing user preferences the proposed RFD model use two algorithms. First algorithm helps to cluster the terms within specified limits in three categories as positive negative and general. It discovers both positive and negative patterns in text documents as higher level features and deploys them over low-level features. Second algorithm is used for calculating feature weights. The key research question is how to find the best partition for term categorization to effectively classify relevant and

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 1, Issue 6, November 2015

irrelevant documents because of large number of possible combinations of group of features. So we are going to refine the RFD model by using efficient algorithms such as using Hierarchical Agglomerative clustering for clustering documents into more than three (positive, negative, general) categories which benefits to form exact clusters allowing addition of new dataset dynamically with improved performance.

In this area, we all add the RFD model for meaning attribute discovery, which explains the appropriate characteristics in relation to 3 teams: positive certain terms, standard terms and negative certain terms determined by the hearings in the teaching arranged. We 1st go over the idea of "specificity" regarding the relative "specificity" with teaching datasets and also the utter "specificity" with sector ontology. We also current a way to comprehend perhaps the planned relative "specificity" can be fair with expression in the utter "specificity". Lastly, we all add the term weighting procedure within the RFD model.

specificity

Inside RDF type, a term's specificity (referred to be able to while family member specificity in this paper) is usually defined [2] in line with it's appearance in a very granted instruction established. Make it possible for T2 be a couple of words which are extracted coming from D in addition to Big t ¼ T1 [T2. Presented a period capital t a couple of Big t, it's coverage may be the number of related docs which contain capital t, and coverage may be the number of irrelevant docs which contain capital t. Most of us suppose that this words commonly used throughout each related docs in addition to irrelevant docs are generally standard words. Therefore, we should classify the particular words which can be more frequently used in the particular related docs to the good specific type; the particular words which can be more frequently used in the particular irrelevant docs are generally categorized directly into the particular bad specific type.

Any term's family member specificity explains the actual magnitude to be able to which in turn the word is targeted on individual of which customers want. It is very difficult to be able to gauge the actual family member specificity connected with phrases because a term's specificity is dependent upon users' sides in their information requirements [5]. As an example, "knowledge discovery" would have been a basic phrase from the files exploration community; nonetheless, it may be a selected phrase if we look at information technology.

Weighting Features

To explain importance attributes to get a presented topic, usually we all feel that particular terms are incredibly helpful so as to recognize the topic through some other subject areas. On the other hand, our own trials present which only using particular terms is actually not adequate enough to further improve the particular functionality regarding importance characteristic breakthrough discovery simply because user info requires can't simply be covered by documents that contain simply the unique terms. Consequently, the best way is to apply the unique terms blended with many of the general terms. Most of us go over this challenge inside the evaluation area. To enhance the particular efficiency, the particular RFD utilised unimportant documents inside the education arranged so as to eliminate the tones. The 1st concern with applying unimportant documents is actually the best way to select a suited set of unimportant documents given that a very large set of damaging trials is usually acquired. For example, the Yahoo and google lookup can easily return numerous documents; on the other hand, only a few of people documents may be

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 1, Issue 6, November 2015

regarding curiosity with a Net user. Naturally, it is not efficient make use of each of the unimportant documents. Nearly all versions can easily position document applying a couple of taken out attributes. If the unimportant document obtains an increased position, the particular document is actually termed the arrest [3] simply because it is a bogus breakthrough discovery. These offenders are defined as the particular top-K positioned unimportant documents. The essential hypothesis in this paper is actually which importance attributes are utilized to spell out pertinent documents, and also unimportant documents are utilized to ensure the particular elegance regarding taken out attributes. Consequently, RFD simply decides on many offenders (i. e., top-K positioned unimportant documents) quite compared to work with many unimportant documents. we all go over the particular functionality regarding applying different K prices, wherever K ¼ d only two acquired the top functionality.

V. EXPECTED RESULTS

- Understand of Specificity on LCSH Ontology
- RFD2 vs RFD1
- RFD2 vs Pattern-Based Models and n-Grams
- RFD2 vs Term Feature Selection Models
- Robustness.
- Term Classification and Specificity

VI. CONCLUSION

Your research proposes an alternative tactic with regard to importance feature development throughout text message paperwork. It gifts a method to come across and classify low-level capabilities depending on both their particular shows within the higher-level styles and their particular specificity. What's more, it brings out a method to pick out immaterial paperwork with regard to weighting capabilities. In this document, all of us carried on to be able to create the actual RFD model and experimentally demonstrate that this recommended specificity perform is usually sensible as well as the term class can be successfully approximated with a feature clustering method. The very first RFD model uses two empirical guidelines to be able to set the actual border between your categories. It accomplishes the actual predicted overall performance, but it demands the actual by hand screening associated with many various values associated with guidelines. The particular new model utilizes a feature clustering technique to routinely party conditions to the a few categories. In contrast to the primary model, the modern model is a lot better and accomplished the actual sufficient overall performance likewise. These kinds of studies underscore that this recommended model accomplishes the actual ideal overall performance with regard to looking at with term-based baseline models and pattern-based baseline models. The outcome also present that this term class can be successfully approximated because of the recommended feature clustering method, the actual recommended specificity perform is usually sensible as well as the recommended models are effective.

REFERENCES

[1] 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.

[2] Y. Li, A. Algarni, and N. Zhong, "Mining positive and negative patterns for relevance feature discovery," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2010, pp. 753–762.

[3] M. Seno and G. Karypis, "Slpminer: An algorithm for finding frequent sequential patterns using length-decreasing support constraint," in Proc. 2nd IEEE Conf. Data Mining, 2002, pp. 418–425.

[4] Z. Xu, and R. Akella, "Active relevance feedback for difficult queries," in Proc. ACM Conf. Inf. Knowl. Manage., 2008, pp. 459–468.

[5] S. Zhu, X. Ji, W. Xu, and Y. Gong, "Multi-labelled classification using maximum entropy method," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2005, pp. 1041–1048.

[6] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," in IEEE Trans. Knowl. Data Eng., vol. 25, no. 1, pp. 1–14, Jan. 2013.

[7] S. Shehata, F. Karray, and M. Kamel, "A concept-based model for enhancing text categorization," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2007, pp. 629–637.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 1, Issue 6, November 2015

[14] 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.

[15] 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.

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

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

[18] 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.