isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 4, Issue: 3,Special Issue: 2 ,Apr,2016 ,ISSN_NO: 2321-3337

Relevance Feature Discovery for Text Mining

Laliteshwari [1],Clarish [2],Mrs.A.G.Jessy Nirmal[3]

Student, Dept of Computer Science and Engineering, Agni College Of Technology, India[1,2]

Asst Professor, Dept of Computer Science and Engineering, Agni College Of Technology, India[3]

ABSTRACT : *Relevance feature discovery for text mining is a big challenge to guarantee the quality of discovered relevance features in text documents for describing user preferences because of large scale terms and data patterns. Most existing popular text mining and classification methods have adopted term-based approaches. However, they have all suffered from the problems of polysemy and synonymy. Over the years, there has been often held the hypothesis that pattern-based methods should perform better than term-based ones in describing user preferences; yet, how to effectively use large scale patterns remains a hard problem in text mining.*

To make a breakthrough in this challenging issue, 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 (terms). It also classifies terms into categories and updates term weights based on their specificity and their distributions in patterns. Substantial experiments using this model on RCV1, TREC topics and Reuters-21578 show that the proposed model significantly outperforms both the state-of-the-art term-based methods and the pattern based methods.

Keywords— Text mining, text feature extraction, text classification

1.INTRODUCTION

The objective of relevance feature discovery (RFD) is to find the useful features available in text documents, including both relevant and irrelevant ones, for describing text mining results. This problem is also of central interest in many Web personalized applications, and has received attention from researchers in Data Mining, Machine Learning, Information Retrieval and Web Intelligence communities .

There are two challenging issues in using pattern mining techniques for finding relevance features in both relevant and irrelevant documents . The first is the low-support problem. Given a topic, long patterns are usually more specific for the topic, but they usually appear in documents with low support or frequency. If the minimum support is decreased, a lot of noisy patterns can be discovered. The second issue is the misinterpretation problem, which means the measures (e.g., "support" and "confidence") used in pattern mining turn out to be not suitable in using patterns for solving problems. For example, a highly frequent pattern (normally a short pattern)may be a general pattern since it can be frequently used in both relevant and irrelevant documents. Hence, the difficult problem is how to use discovered patterns to accurately weight useful features.

There are several existing methods for solving the two challenging issues in text mining. Pattern taxonomy mining (PTM) models have been proposed in which, mining closed sequential patterns in text paragraphs and deploying them over a term space to weight useful features.

Concept-based model (CBM) has also been proposed to discover concepts by using natural language processing (NLP) techniques. It proposed verb-argument structures to find concepts in sentences. These pattern (or concepts) based approaches have shown an important improvement in the effectiveness. However, fewer significant improvements are made compared with the best term-based method because how to effectively integrate patterns in both relevant and irrelevant documents is still an open problem. Over the years, people have developed many mature term-based techniques for ranking documents, information filtering and text classification.

Recently, several hybrid approaches were proposed for text classification. To learn term features within only relevant documents and unlabelled documents,used two term-based models. In the first stage, it utilized a Rocchio classifier to extract a set of reliable irrelevant documents from the unlabeled set.

In the second stage, it built a SVM classifier to classify text documents. A two-stage model was also proposed which proved that the integration of the rough analysis (a term-based model) and pattern taxonomy mining is the best way to design a two-stage model for information filtering systems.

2. PATTERN TAXONOMY MINING

Pattern taxonomy mining (a term-based model) and is the best way to design a two-stage model for information filtering systems. Text mining also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.

2.1. FEATURE SELECTION

Feature selection is a technique that selects a subset of features from data for modeling systems. Over the years, a variety of feature selection methods (e.g., Filter, Wrapper, Embedded and Hybrid approaches, and unsupervised or semi-supervised methods) have been proposed in various fields .Feature selection is also one of important steps for text classification and information filtering which is the task of assigning documents to predefined classes.

2.2.TEXT ANALYTICS SOFTWARE :

Text analytics software can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database.

With an iterative approach, an organization can successfully use text analytics to gain insight into content-specific values such as sentiment, emotion, intensity and relevance. Because text analytics technology is still considered to be an emerging technology, however, results and depth of analysis can vary wildly from vendor to vendor.

ANALYSING THE PERFORMANCE OF STUDENTS AND MARK ASSESSMENT

2.2. Existing System

A. To learn term features within only relevant document and unlabelled documents, paper used two term-based models. In the first stage, it utilized a **Rocchio classifier** to extract a set of reliable irrelevant documents from the unlabeled set. In the second stage, it built a SVM classifier to classify text documents. A two-stage model was also proposed in which proved that the integration of thorough analysis (a term-based model) and pattern taxonomy mining is the best way to design a two-stage model for information filtering systems.

B.WORDNET TOOL AND NLP TECHNIQUE

Teacher prepares questions and answers for student assessment. Text mining process is done by natural language processing and word net tools. Pos tagger is implemented to extract the important keywords in the answer given by staff before assessment is done.

Keywords are categorized mandatory keywords, subordinate keywords, and technical keywords. Wordnet tool is used to give the related synonyms to literal word in the subordinate terms.

C. **Concept-based model (CBM)** has also been proposed to discover concepts by using natural language processing (NLP) techniques. It proposed verb-argument structures to find concepts in sentences.

PROPOSED SYSTEM

The advantages of the proposed model include:

- Effective use of both relevant and irrelevant feedback to find useful features; and
- Integration of both term and pattern features together rather than using them in two separated stages.

TRANSCODING TECHNIQUES :

DEFINITION :

Transcoding is the direct analog-to-analog or digital-to-digital conversion of one encoding to another, such as for movie data files (e.g., PAL, SECAM, NTSC), audio files (e.g.,MP3, WAV), or character encoding

Transcoding is commonly a lossy process, introducing generation loss; however, transcoding can be lossless if the output is either losslessly compressed or uncompressed. The process of transcoding into a lossy format introduces varying degrees of generation loss, while the transcoding from lossy to lossless or uncompressed is technically a lossless conversion because no information is lost, however the process is irreversible and is more correctly known as destructive

Teachers prepare the material for each subject and also give tags (good, best) for student material recommendation. Here we upload the materials like video, text, pdf. Video transcoding is applied while video materials are uploaded for below average students. After finishing the assessment test, in student portal they get the materials based on overall performance calculated by server. If they have doubt while watching video content, students can interactively raise questions by simply clicking on the video frame. The video frames are previous indexed so that appropriate meta information's can be extracted for each frame.

The student's questions and meta information from the current frame are send to server and can be reviewed by the staff. Once the staff login they will be notified with the questions and then staffs can reply to the question. The Student can now be able to view the answers given by the staffs.

FFMPEG SOFTWARE

FFmpeg is a free software project that produces libraries and programs for handling multimedia data. FFmpeg includes libavcodec, an audio/video codec library used by several other projects, libavformat, an audio/video container mux and demux library,and the ffmpeg command line program for transcoding multimedia files. FFmpeg is published under the GNU Lesser General Public License 2.1+ or GNU General Public License 2+ (depending on which options are enabled).

Command line tools

- FFMPEG is a command-line tool that converts audio or video formats. It can also capture and encode in real-time from various hardware and software sources such as a TV capture card.
- ffserver is an HTTP and RTSP multimedia streaming server for live and recorded broadcasts. It can also be used to time shift live broadcasts.

MAJOR IMPLEMENTATION :

Project Allocation: In this module coordinator upload project base paper on behalf of each and every student, and also allocate batches for all projects. Batches were created by coordinator by selecting number of student in batch and student ID’s.

Text mining for assessment: Teacher prepares questions and answers for student assessment. Text mining process is done by natural language processing and word net tools. Pos tagger is implemented to extract the important keywords in the answer given by staff before assessment is done

Project Review and Student Assessment : Student login with his credentials and then uploads the review materials in server. Reviewer gives the review marks for each student based on performance. Here we allotted three reviews, and give marks for student based on review performance.

Student answers are evaluated later in server by extracting keywords using NLP technique and wordnet tool.

Material Recommendation and Interactive Student learning : Teachers prepare the material for each subject and also give tags (good, best) for student material recommendation. Here we upload the materials like video, text, pdf. Video transcoding is applied while video materials are uploaded for below average students

ARCHITECTURE

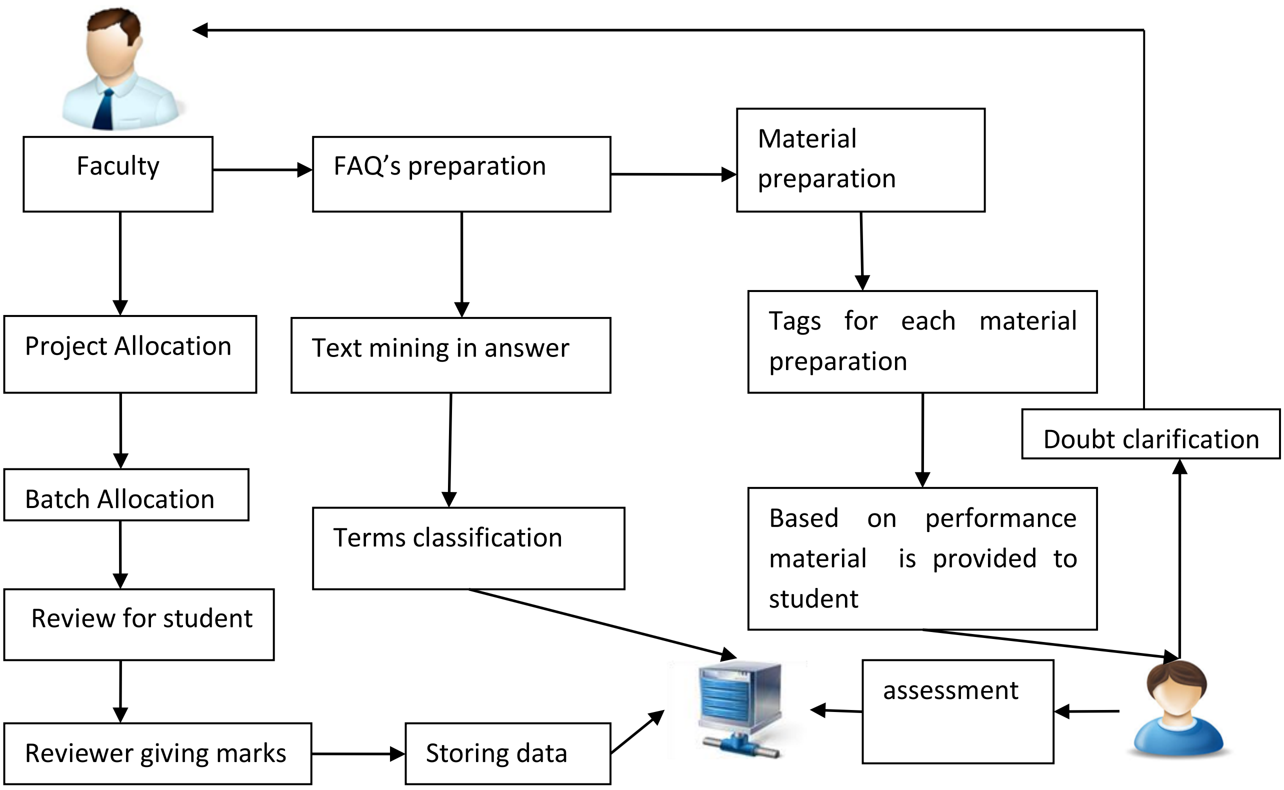

Fig 1 System Architecture

CONCLUSION

The research 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 patterns and their specificity. It also introduces a method to select irrelevant documents for weighting features. In this paper, we continued to develop the RFD model and experimentally prove that the proposed specificity function is reasonable and the term classification can be effectively approximated by a feature clustering method.

The first RFD model uses two empirical parameters to set the boundary between the categories. It achieves the expected performance, but it requires the manually testing of a large number of different values of parameters. The new model uses a feature clustering technique to automatically group terms into the three categories.

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 AC