TEXT MINING CLASSIFICATION FOR EFFECTIVE LEARNING EXPERIENCE

C.Pabitha[1], Dr.B.Vanathi[2]
[1]Assistant Professor, Dept of C.S.E, Valliammai Engineering College,
[2]Professor, Dept of C.S.E, Valliammai Engineering College,

Abstract
The voluminous amount of data are increased in day to day world. Extracting the useful information in unstructured form is found be very difficult. Text mining techniques such as information extraction, clustering, classification, summarization, visualization are available for effective processing of these unstructured data. The main aim of this paper is to relate the relevant feature in text documents for example student study materials and to verify the performance of student so that necessary relevant learning materials can be rendered to them. The materials are categorized and classification algorithms are applied so that the materials given to the students are as per the subject and latest updates are informed to them.
Keywords: Text mining, Text classification, information extraction.

1 INTRODUCTION

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. 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 and analyzed with traditional data mining techniques. 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.

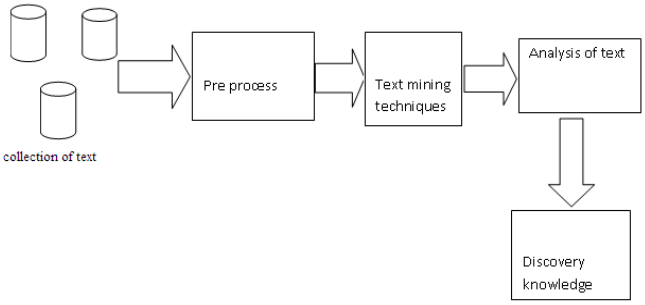

Fig No 1. Preprocessing steps

1.1 TRANSFORMATION STEPS

The conversion of text records to vectors of numeric attributes is a multi-staged process of feature construction, which employs several key transformations. There are many approaches to this process – one of the most common is to treat each text record as a "bag of words." Each unique word constitutes an attribute (a position in our example vector). The number of occurrences of a word in a record (frequency of occurrence) is the attribute's value (in our vector) for that document. Records are therefore represented as vectors of numeric attributes where each attribute value is the frequency of occurrence of a distinct term. This set of document vectors is often referred to as a vector space. Algorithms that operate on such representations are said to be using vector space models of the data.

1.1.1 SENTENCE SPLITTING

Identifying sentence boundaries in a document is not as trivial a process as it may seem. SEASR has components that achieve sentence splitting either using rules or statistical models

(or both). Once sentences are identified they are recorded as annotations in their own annotation set.

1.1.2 TOKENIZATION

Tokenization, simply put, is basically labeling individual words or sometimes word parts. This is important because many down stream components need the tokens to be clearly identified of analysis. Tokens are recorded as annotations in their own annotation set.

1.1.3 PART-OF-SPEECH (POS) TAGGING

Such components typically assign a POS tag to a token (the Penn Treebank project has provided a set of codes for this purpose that is widely used). Other data such as lemma, lexemes, and synonyms (to name a very few) may also be identified at this stage. POS information is stored as features of the token annotation.

1.1.4 STOP WORD FILTERING

Very common words like "and" and "the" are often filtered out to improve performance. This process is called stop word removal. SEASR has a components to perform this process. One basic approach is to remove all words that appear on a list of common words. Another approach is to remove words that occur with high frequency across most documents — these types of terms create "noise" that makes text records less distinguishable. The stop word filters will remove token annotations from a documents token annotation set or they can also mark such annotations as "stopped."

1.1.5 POS FILTERING

This component reads a document object as input and filters the tokens for that document based on part of speech tag information. A document object is taken as input. The token list is retrieved from the document and only those tokens with part of speech tags that match at least one value in the selected tag list are retained. The filtered list of tokens is placed into the document (replacing the old list) and the document is output. In the SEASR component the PoS Tags is a comma-delimited list of the part-of-speech tags that we want to retain. Tokens that do not possess one of these values will be removed from the annotation list or flagged as "stopped".

2 RELATED WORK

2.1 Mining Sequential Patterns by Pattern-Growth the PrefixSpan Approach

Sequential pattern mining is an important data mining problem with broad applications. However, it is also a difficult problem since the mining may have to generate or examine a combinatorially explosive number of intermediate sub sequences. Most of the previously developed sequential pattern mining methods, such as GSP, explore a candidate generation-and-test approach to reduce the number of candidates to be examined. However, this approach may not be efficient in mining large sequence databases having numerous patterns and/or long patterns. In this paper, we propose a projection-based, sequential pattern-growth approach for efficient mining of sequential patterns [1]. In this approach, a sequence database is recursively projected into a set of smaller projected databases, and sequential patterns are grown in each projected database by exploring only locally frequent fragments. Based on an initial study of the pattern growth-based sequential pattern mining, FreeSpan, we propose a more efficient method, called PSP, which offers ordered growth and reduced projected databases. To further improve the performance, a pseudo projection technique is developed in PrefixSpan. A comprehensive performance study shows that PrefixSpan, in most cases, outperforms the a priori-based algorithm GSP, FreeSpan, and SPADE (a sequential pattern mining algorithm that adopts vertical data format), and PrefixSpan integrated with pseudo projection is the fastest among all the tested algorithms. Furthermore, this mining methodology can be extended to mining sequential patterns with user-specified constraints. The high promise of the pattern-growth approach may lead to its further extension toward efficient mining of other kinds of frequent patterns, such as frequent substructures.

Disadvantage

1. Maximal sequential patterns, mining approximate sequential patterns, and extension of the method toward mining structured patterns.
2. The developments of specialized sequential pattern mining methods for particular applications, such as DNA sequence mining that may admit faults,

such as allowing insertions, deletions, and mutations in DNA sequences, and handling industry/engineering sequential process analysis are interesting issues for future research.

2.2 Phrase Dependency Parsing for Opinion Mining

In this paper, we present a novel approach for mining opinions from product reviews, where it converts opinion mining task to identify product features, expressions of opinions and relations relation between them. By taking advantage of the observation that a lot of product feature are phrases, a concept of phrase dependency parsing is introduced, which extends traditional dependency parsing to phrase level[2]. This concept is then implemented for extracting relation between product feature and expressions of opinions. Experimental evalutions shows that the mining task can benefit from phrase dependency parsing.

In this paper, we define an opinion unit as a triple consisting of a product feature, an expression of opinion, and an emotional attitude (positive or negative). We use this definition as the basis for our opinion mining task. Since a product review may refer more than one product feature and express different opinions on each of them, the relation extraction is an important subtask of opinion mining. Introducing the concept of phrase dependency parsing segment an input sentence into "phrases" and links segment with directed arcs. The parsing focuses on the :phrases" and the relation between them, rather than on the single words inside each phrase.Because phrase dependency parsing naturally divides the dependencies into local and global, a novel tree kernel method has also been proposed.

Disadvantage:

1. Product feature could not be represented by a single word, dependency parsing might not be the best approach here unfortunately, which provides dependency relation only between words.
2. Experimental results show that relation extraction task can benefit from dependencies within a phrase.
3. On relation extraction usually use the head word to represent the whole phrase and extract features from the word level dependency tree. This solution is problematic because the information provided by the phrase itself cannot be used by this kind of method.

2.3 A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data

A fast clustering-based feature selection algorithm (FAST) is proposed and experimentally evaluated in this paper [3]. The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent, the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated. Extensive experiments are carried out to compare FAST and several representative feature selection algorithms, namely, FCBF, ReliefF, CFS, Consist, and FOCUS-SF, with respect to four types of well-known classifiers, namely, the probability based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER before and after feature selection. The results, on 35 publicly available real-world high-dimensional image, microarray, and text data, demonstrate that the FAST not only produces smaller subsets of features but also improves the performances of the four types of classifiers.

In this paper, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

Disadvantage:

FCBF is a good alternative for image and text data but it is not support for different types of correlation measures.

2.4 SLPMiner: An Algorithm for Finding Frequent Sequential Patterns Using Length-Decreasing Support Constraint

In this paper they present an algorithm called SLPMiner that finds all sequential patterns that satisfy a lengh-decreasing support constraint. SLPMiner combines an efficient database-projection-based approach for sequential pattern discovery with three effective database-projection-based approach for sequential pattern discovery with three effective database pruning methods that dramatically reduce the search space.Our experimental evaluation shows that SLPMiner, by effectively exploiting the length-decreasing support constraint, is up to two orders of magnitude faster, and its runtime increase gradually as the average length of the sequential (and the discovered frequent patterns) increases[4].

We developed an algorithm called SLPMiner that finds all the frequent sequential patterns that satisfy a given length decreasing support constraint. SLPMiner serves as a platform to develop and evaluate various pruning methods for reducing the complexity of finding this type of patterns. Our design goals for SLPMiner were to make it both efficient and at the same time sufficiently generic so that any conclusions drawn from our experiments can carry through other database-projection-based sequential pattern mining algorithms.

2.5 Sentiment Classification by Sentence Level Semantic Orientation using SentiWordNet from Online Reviews and Blogs

In this method a domain independent rule based method for semantically classifying sentiment from online customer reviews and comments. The method is effective as it takes reviews, check individual sentences and decides its semantic orientation considering the sentence structure and contextual dependency of each word.KDT (knowledge discovery in text) or text data mining or text mining are terms used for the mining of unstructured or semi-structured data Retrieval of document relevant to the information needs of a user, is the primary concern of the traditional IR (perhaps a more appropriate name would be data retrieval);however , the user is left on his/her own to find the desired information in the document[5]. In Heart's opinion , data mining has not only directed dealing with the information, but it also attempt to uncover or glean previously unknown, information from the data (text).Different linguistic levels (words, sentence and document)highlighting the key difference between supervised machine learning methods, that rely on annotated corpora or corpus based, and unsupervised/lexicon-based methods in sentiment classification. Three main step are always involved in the process of text mining and sentiment classification; they are

a) Acquiring texts which are relevant to the area of concern usually called IR;
b) Presenting content collected from these texts in a format that can be processed, such as statistical modeling, natural language process.
c) Actually using the information in the presented format.

Disadvantages:

3 EXISTING SYSTEM:

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.

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 method.

RFD model for relevance feature discovery, which describes the relevant features in relation to three groups: positive 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.

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. Compared with the first model, the new model is much more efficient and achieved the satisfactory performance as well.

4 PROPOSED SYSTEM

In this proposed system we create student and teacher application. Teacher allocates the batch and project for each student and also validates the Project title and content. String matching algorithm is used to validate the title of the project. Content based model is used to validate the base paper. Teacher prepare FAQ's question and answer, then extract the words and classify the terms using NLP and wordnet tool. Student writes the assessment and servers validate the student answer and calculate the performance. After that teacher prepare materials based on student performance and also give tags to each material like good, average. Student getting material after assessment test is completion. Then student can view the material if they have any doubt send question to teacher. Teacher can get the question and clear the doubt through website in offline.

Modules for proposed system:

- Project Allocation
- Text mining in FAQ's preparation
- Student Assessment test and performance calculation
- Material preparation and Student Learning

A Project Allocation:

In this module coordinator allotted a project for each and every student, and also allocate batch for all project. If student having same project then we will validate the project title and if it is same then the project will not allotted for this students. String matching algorithm is used to validate the project title.

B Text Mining in FAQ's Preparation:

In this module project coordinator checks the project of content whether the students having the same content of paper for different batch. Teacher prepares FAQ's question and answer. Text mining process **natural language processing** and word net tools is used to extract the files and contents.NLP process is used to extract the literal meaning words in file content. Wordnet tool is used to give the related synonyms to literal word in that content. Teacher gives mandatory term, subordinate term, technical term for each answer, using the terms answer will check.

C Student Assessment Test and Performance Calculation

Reviewer gives the review marks for each student performance. Here we allotted the three reviews, and give marks for student based performance. Student login with his credentials and write the assessment test. Student answer is to be extract using NLP technique and wordnet tool, we evaluate separate terms. Machine will evaluate the answer using teacher terms. Depends upon student answer they will give marks and prepare progress report.

D Material Preparation and Student Learning

Teacher prepare the material for each subjects and also give tags(good, best).Here we upload the materials like video,text,pdf. Video transcoding is applied while material is uploaded for below average students. After finishing the assessment test, in student webpage they get the materials based on our test performance. If they have doubt in material, student can type the question sends to teacher. Teacher can get the question while login then they will analyze the question and clear the student doubts.

5 OVER ALL ARCHITECTURE

We create student and teacher application. Teacher allocates the batch and project for each student and also validates the Project title and content.String matching algorithm is used to validate the title of the project.Content based model is used to validate the base paper. Teacher prepare FAQ's question and answer, then extract the words and classify the terms using NLP and wordnet tool. Student writes the assessment and servers validate the student answer and calculate the performance. After that teacher prepare materials based on student performance and also give tags to each material like good, average. Student getting material after assessment test is completion. Then student can view the material if they have any doubt send question to teacher. Teacher can get the question and clear the doubt through website in offline

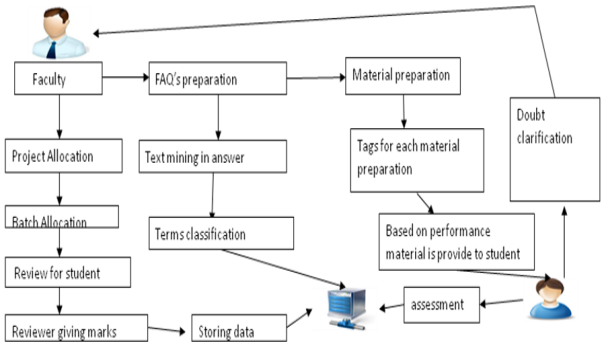

Fig No. 4.1 Overall Architecture

6 THE ALGORITHM USED IN THE PROPOSED SYSTEM

6.1 NAIVE BAYES APPROACH

Build the vocabulary as the list of all distinct words that appear in all the document of the training set. Remove the words and markings. The word in the vocabulary become the attributes, assuming that classification is independent of the position of the words. Each Document in the training set become a record with frequencies for each word in the vocabulary. Train the classifier based on the training data set, by computing the prior probabilities for each class and attributes.

6.2 NATURAL LANGUAGE PROCESSING:

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.

Current approaches to NLP are based on machine learning, a type of artificial intelligence that examines and uses patterns in data to improve a program's own understanding. Most of the research being done on natural language processing revolves around search, especially enterprise search.

Common NLP tasks in software programs today include:

- Sentence segmentation
- Part-of-speech tagging and parsing.
- Deep analytics.
- Named entity extraction

Natural Language Programming (NLP) is an ontology-assisted way of programming in terms of natural language sentences, e.g. English. A structured document with Content, sections and subsections for explanations of sentences forms a NLP document, which is actually a computer program.

6.2.1. Natural Language Understanding

Taking some spoken/typed sentence and working out what it means.

Raw speech signal

- Speech recognition
  Sequence of words spoken
- Syntactic analysis using knowledge of the grammar
  Structure of the sentence
- Semantic analysis using info. about meaning of words
  Partial representation of meaning of sentence
- Pragmatic analysis using info. about context

Natural language Processing types:

1. Query answering
2. Text categorization
3. Text summary
4. Data extraction

6.2.2. Natural Language Generation

Taking some formal representation of what you want to say and working out a way to express it in a natural (human) language (e.g., English)

Concept-Based Similarity measure

The concept-based measure exploits the information extracted from the concept-based

algorithm to better judge the similarity between the documents. The similarity measure is a function of the following factors:

A. No. of matching concepts m in the verb argument structures in each document d

B. Total no. of sentences sn that contain matching concept ci in each document d

C. Total no. of labeled verb argument structures v, in each sentence s

D. The ctfi of each concept ci in s for each document d, where i=1,2,3,….,m

E. The tfi of each concept ci in each document d, where i=1,2,3,….,m

F. The dfi of each concept ci in each document d, where i=1,2,3,….,m

G. The length l of each concept in the verb argument structure in each document d

H. The length Lv of each verb argument structure which contains a matched document, and the total no. of documents, N, in the corpus.

7 CONCLUSION

In this paper teacher and student application to evaluate their performance using text mining. Project and batch will be nominated by the coordinator. A concept based mining model is developed to evaluate the relevancy of the base paper. The coordinator prepare the frequently asked questions and then extract the words and classify the terms using Natural Language Processing and wordnet tool**.** The performance of the students is validated by the server and calculate their assessment marks. The coordinator prepare materials based on student performance and also give grades to each material like average,poor,good. Then student can view the material if they have any doubt send question to teacher. Teacher can get the question and clear the doubt through website in offline. The set up made will be demonstrating a smart classroom example where the teacher student can easily communicate whenever an explanation is needed.

REFERENCES

[1]shady shehata, Mohamed S. Kamel, "An efficient concept-based mining model for enhancing text clustering"IEEE transactions on knowledge and data engineering, vol. 22, no. 10, october 2010

[2] C. C. Aggarwal and P. S. Yu, ―A framework for clustering massive text and categorical data streams,‖ in *Proc. SIAM Conf. Data Mining*, 2006, pp. 477–481.

[3] .M. Steinbach, G. Karypis, and V. Kumar, ―A comparison of document clustering techniques,‖ in *Proc. Text Mining Workshop KDD*,2000, pp. 109–110.

[4] . S. Zhong, ―Efficient streaming text clustering,‖ *Neural Netw.*,vol. 18, no. 5–6, pp. 790–798, 2005.

[5] . YuanbinWu, Qi Zhang, Xuanjing Huang, LideWu Fudan "phrase Dependency parsing for opinion mining" University school of computer science.

[6] S. Guha, R. Rastogi, and K. Shim, ―rock: A robust clustering algorithm for categorical attributes,‖ *Inf. Syst.*, vol. 25, no. 5, pp. 345–366, 2000.

[7] Shubhangi V. Airekar, Prof. Dhanshree S. Kulkurni survey paper on text mining with side information.

Liwei Wei, Bo Wei, Bin Wang, "Text Classificat ion Using Support Vector Machine with Mixture of Kernel", A Journal of Software Engineering and Applications, 2012, 5, 55-58, doi:10.4236/jsea.2012.512b012 Published Online December 2012

[8]M. A. Hearst. What is text mining? http://www.sims.berkeley.edu/˜hearst/text-mining.html, Oct. 2003.

[9] Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky, "HierarchicalTopics: Visually Exploring Large Text Collect ions Using Topic Hierarchies", IEEE transactions on visualization and computer graphics, vol. 19, no. 12, december 2013

[10] Liwei Wei, Bo Wei, Bin Wang, "Text Classificat ion Using Support Vector Machine with Mixture of Kernel", A Journal of Software Engineering and Applications, 2012, 5, 55-58, doi:10.4236/jsea.2012.512b012 Published Online December 2012

[11] D.Cutting, D. Karger, J. Pedersen, and J. Tukey, ―Scatter/Gather:A cluster-based approach to browsing large document collections,‖in *Proc. ACM SIGIR Conf.*, New York, NY, USA, 1992,pp. 318–329.