Effective Classification of Text and Improving Learning Experience

Ms. 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*- Amount of data is increased in today world. We can extract the useful information which is generally in the unstructured form. More number of techniques available in text mining such as information extraction, clustering, classification, summarization, visualization are available under the text mining techniques. The main aim of this paper is to discover relevant feature in text documents (student answer sheets) and to evaluate the performance of student so that appropriate learning materials can be rendered to them.**

***Index Terms*-Textmining,Textclassification, information extraction**

I. INTRODUCTION

Text mining also refer to as text data mining, roughly equivalent to text analytics, refers to the process of deriving quality information from text. Quality information is typically derived through the devising of patterns through means such as statistical pattern learning. Text analytic software is used to transposing words or phrases in unstructured data into structured data. Text mining is used in more application such as information extraction, text classification, clustering and natural language processing. Natural language process is used to read and analyse the textual information. It is based on the queries. In the data mining pattern are extracted from the database. In web mining the input is structured. In sentence splitting is identified the sentence boundaries in the document. 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.

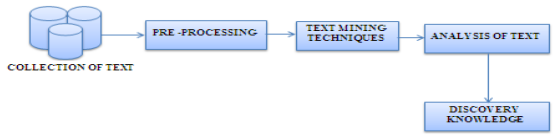


Fig 1. Preprocessing Steps

II. RELATED WORKS

Paper presented by Ning Zhong, Yuefeng Li, and Sheng-Tang [1] that presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance.

Paper presented by Charu C. Aggarwal, Philip S. Yu demonstrates [2] that segmentation of text data records is required in many applications such as document organization, new group filtering, and text crawling. The categorical data stream clustering problem of customer segmentation. Statistical summarization methodology an online approach for clustering massive text and categorical data strams is presented here.

Paper presented by Douglass Michael Steinbach George Karypis Vipin Kumar demonstrates [3] that two main to document clustering, agglomerative hierarchical clustering and k-means are compared here. Hierarchical clustering is always better quality clustering approach. K means have the time complexity. Sometime k means and agglomerative are merged and get the best of both world. K means technology is better than the K means approach as good as or better than the hierarchical approach. An explanation for these results that is based on analysis of the specific clustering algorithm and the nature of document data is proposed here.

Paper presented by S. Zhong demonstrates [4] that clustering data streams has been a new research topic, recently used in many real data mining applications, and has attracted a lot of research attention. However,

there is not much work on clustering high-dimensional streaming text data. This paper merges an efficient online spherical k-means algorithm with an existing scalable clustering strategy to achieve fast and adaptive clustering of text streams. The OSKM algorithm modifies the spherical k-means algorithm, using online update based on the well known. Winner Take all competitive learning. It has been shown to be as efficient as SPKM, but much superior in clustering quality. The scalable clustering strategy was previously developed to deal with very large data bases that cannot fit into a limited memory and that are too expensive to read/scan multiple times. Using this method, one keeps only sufficient statistics for history data to retain (part of) the contribution of history data and to accommodate the limited memory. To make the proposed clustering algorithm adaptive to data streams, a forgetting factor is introduced here that applies exponential decay to the importance of history data. The older a set of text documents, the less weight they carry. The experimental results demonstrate the efficiency of the proposed algorithm and reveal an intuitive and an interesting fact for clustering text streams one need to forget to be adaptive

Paper presented by yuanbin Wu, xuanjing Huang [5] that define an opinion unit as a triple consisting of a product feature, an expression of opinion, and an emotional attitude(positive or negative).They 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.

III. SYSTEM DESIGN

In this proposed systemwe 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. Concept based model is used to validate the base paper. Teacher prepares 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.

Methods:

- 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.

IV. 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 prepares 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.

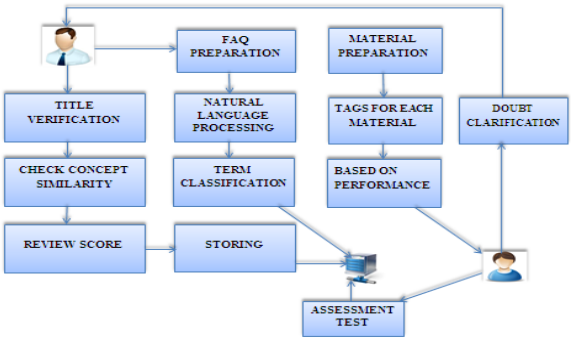


Fig 2 Architectural diagram

V. ALGORITHM

A. NAIVE BAYES APPROACH

Build the glossary 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 becomes the attributes, assuming that classification is independent of the position of the words.

Each Document in the training set becomes a record with frequencies for each word in the glossary. Train the classifier based on the training data set, by computing the prior probabilities for each class and attributes.

B. NATURAL LANGUAGEPROCESSING:

Natural language processing (NLP) is an artificial intelligence, and computational linguistics concerned with the interactions between computers and human languages. Most of the NLP techniques are used in the human - computer interaction. NLP includes two ways of processing; they are natural language understanding and natural language generation. Natural language understanding is described as taking some spoken or typed sentence and working out the means of it. And the natural language generation is the formal representation of what human want to say and working on the way to express it in natural form.

C. CONCEPT-BASED MINING MODEL

The concept based mining model can effectively discriminate between non important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. The mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence. This means that this term can have more than one semantic role in the same sentence. In such cases, this term plays important semantic roles that contribute to the meaning of the sentence. In the concept-based mining model, a labelled terms either word or phrase is considered as concept.

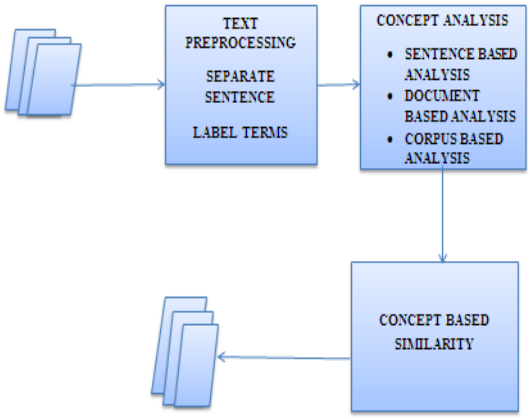


Fig 3. Concept Based Similarity

CONCEPT BASED SIMILARITY MEASURE

It is used to determine accurate similarity between the documents. It is used in three ways, First it analysed labelled term are the concept that capture the semantic structure of each sentence. Next the frequency of a concept to the meaning of the sentence, as well as to the main topics of the document. Finally the number of document that contains the analysed concepts is used to discriminate among document in calculating the similarity.

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

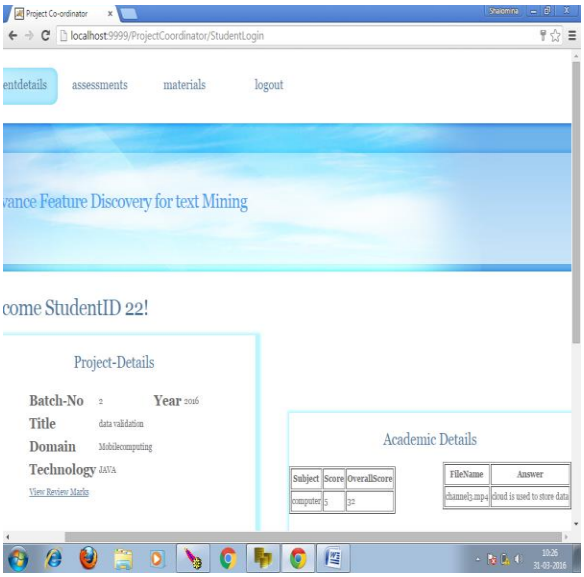
F. The dfi of each concept ci in each document d,where i=1,2,3,…..,m

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

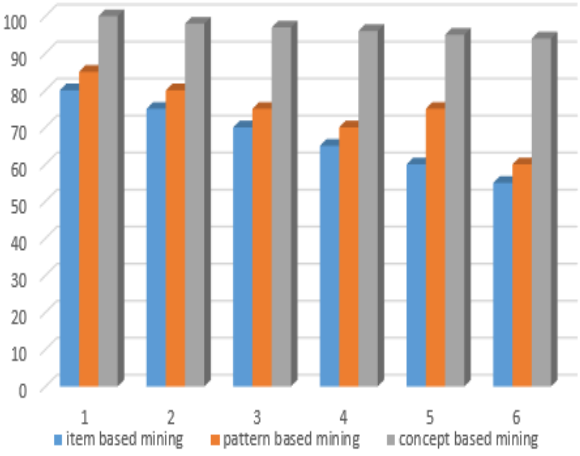
H. The length Lv of each verb argument structure which contains a matched document, and The total no. of documents, N, in the corpus.

VI. IMPLEMENTATION

In this module the co-ordinator verify the concept the base paper. 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. Video material is uploaded for student to reduce the learning difficulties. Teacher can get the question while login then they will analyze the question and clear the student doubts.



comparision between the various text mining techniques is as shown in the graph below,



VII. CONCLUSION

In this project, proposed a teacher and student application to evaluate the student answer sheet using text mining. First, the coordinator allocates the project and batch for students. Then developed a concept based mining model to evaluate the concept of the base paper. The teacher prepares the frequently asked question and then extracts the words and classifies the terms using NLP and wordnet tool. Student writes the assessment and servers validate the student answer and calculate the performance.

Teacher prepares materials based on student performance and also gives 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. so it is more useful to student and they can easily improve the learning experience.

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.

[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.