Text Categorization Using Activation Based Term Set

M. Pushpa[1], Dr. K. Nirmala[2]

[1] **Research Scholar, Department of Computer Science, Bharathiar University
Coimbatore, Tamil Nadu, India**

[2] **Department of Computer Science, University of Madras, Quaid-E-Millath Government College for women
Chennai, Tamil Nadu, 600 002, India**

Abstract

Text classification is a challenging field in the current scenario and has great importance in text categorization application. Documents may be classified or categorized according to their subjects or according to their attributes. There is need to categorize a collection of text document into mutually exclusive categories by extracting the concept or features using supervised learning paradigm and different classification algorithm. In this paper we present a naïve based approach for the classification using semi-supervised text classification methodology with the help of Activation term sets. Such frequent term set can be discovered based on David Merrill's First principles of instruction (FPI) techniques. The system uses a pre-defined category group by providing them with the proper training set based on the activation of FPI We made an attempt to classify the document using FPI methodology, the algorithm involves the text tokenization, text categorization and text analysis

Keywords : *Text mining, Text characterization, Text Classification, Text tokenization, FPI and Instructional phase*

1. Introduction

A large portion of all available information today exists in the form of unstructured textual data. Books, magazines, articles, research papers, products, manuals, memorandums, emails and web content: all contain textual information in natural language form. The amount of text is simply too large to read and analyse efficiently. Manual analysing of huge amount of textual data requires a tremendous amount of processing time and effort in reading the text and organizing them in required format. This has led to the development of automated tools and techniques for analysing text to discover knowledge for various applications. These techniques are gathered under the name of **text mining.**

Automatic text categorization becomes more important for dealing massive data. The major problem with text categorization is the high dimensionality of feature space. Now-a-days there are many methods available to deal with text feature selection. In this paper we present an approach to deal with text feature selection based on bag of key words associated with the activation phase of First principle of instruction.

2. Text Mining

Text mining is a challenging task as it involves dealing with text data that are inherently unstructured and fuzzy. Text mining can be defined as knowledge discovery from textual database; it allows us to create a technology that combines human linguistic capability with the speed and accuracy of a computer. Text mining aims to analyse detailed information in the content of each document and to extract interesting information that can be provided only when multiple document's trends and significant features are viewed as whole for useful actions and decision making. Text mining is about analysing text for particular purposes and involves looking for regularities, patterns or trends in natural language text; associations among entities, predictive rules, etc.

The approach that extracts textual data generate by text mining tools, serves as a method to enrich the content of the documents. In this case text mining can be described as a way to extend mining methodologies by an automated process that creates structured data describing the documents. Text mining encompasses many areas like text categorization, text classification, text clustering, text summarization, conceptual navigation, feature extraction, ontology and topic detection, etc., Text mining can also extract concept from a large collection of documents without having to scan through a great number of files i.e. to uncover and discover valuable relationships between ideas and words contained in vast amounts of text information.

2.1 Text Classification or Text Categorization

In the recent information system the volume of documents continues to grow, the manual text classification becomes a very tedious process. Manual text classification is the process of classifying documents one by one without any inhuman expertise. Text classification is the process of classifying document into predefined categories. Document categorization or text categorization system attempts to reproduce human categorization judgement. One of the approaches to build a text categorization system is to manually assign some set of documents to categorize and then use inductive learning to automatically categorize to documents based on the words they contain.

2.1 Role of Text Categorization in Text mining

Text categorization becomes one of the important techniques for handling and organizing text data. It is now being applied in many context, ranging from document indexing based on a controlled vocabulary, to document filtering, automated metadata generation, word sense, disambiguation, population of hierarchical catalogue of web resources, and in general any application requiring document organization or selective and adaptive document dispatching. Text concept centric nature of text documents is also one of the reasons why the issues of text categorization are particularly challenging.

Text categorization based on machine learning methods need a training set and a test set. The training set is a set of documents, which is tagged manually by the experts. The performance of the system depends on good training set. Moreover the machine language approach to the text categorization is based on keyword matching. The use of concepts for text categorization increases its overall performance specifically when considering categorization of domain specific corpus. The motivation for the work described in this paper is the categorization of documents based on semi automated concept in addition to the keywords.

3. First Principles of Instruction

A first principle is an attempt to identify Reigeluth's basic methods. Principles method is a relationship that is always true under appropriate conditions regardless of program or practice. Properties of first principles of instruction learning from a given program will be facilitated in direct proportion to its implementation.

a) Analyze instructional theories, models, programs, and products to extract general first principles of instruction.

b) Identify the cognitive processes associated with each principle.

c) Identify empirical support for each principle.

d) Describe the implementation of these principles in variety of different instructional theories and models. and

e) Identify prescriptions for instructional design associated with these principles.

3.1 Instructional Phases

Many current instructional models suggest that the most effective learning environments are those that are problem-based and involve the student in four distinct phases of learning: 1) activation of prior experience, 2) demonstration of skills 3) application of skills and 4) integration or these skills into real world activities.

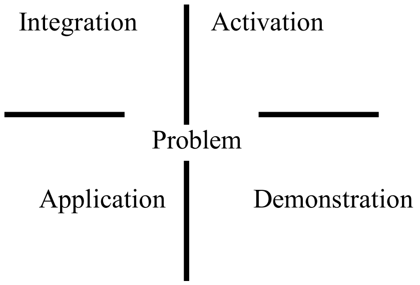

Fig. 1 First principles of instruction

Activation → Recalls the prior knowledge or experience and create learning situation for the new problem.

Demonstration → Demonstrate or show a model of the skill required for the new problem.

Application → Apply the skills obtained to the new problem.

Integration → Provides the capabilities and to show the acquired skill to another new situation.

3.2 Proposed System

The analysis of huge text collection usually aims at finding relevant text or text groups. It would be a tedious task of any information seeking user to scan all retrieved items. In order to facilitate this task, most text mining system characterizes their resulting text with various kinds of annotations. Keywords are helpful in the categorization process.

Activation is the starting phase in the learning process. The importance of activation of existing knowledge has been addressed by a number of educational psychologists. New knowledge builds on the learner's existing knowledge. Learners recall or apply knowledge from relevant past experience as a foundation for new knowledge. This could be from previous courses or job experiences undergone by the learner. For instance, the learner recalls the old relevant information such as dates, events and places. In Merrill's Activation phase, prior knowledge (or experience) is recalled and emotions are triggered in addition to the mental model.

Keywords are valuable means for characterizing texts. In order to extract keywords an efficient and robust, language and domain independent approach has been applied. The keywords are generated by the human judgment based on the repeated analysis of the text. The algorithm examines the first principle of instruction with the help of the bag-of-keywords as a feature term set. Using the bag-of-keywords thus generated as a feature set based on the semantic and syntactic functions of the words the proposed algorithm examines the FPI properties for the text categorization. The system is implemented with a set of processes like parsing and tokenizing.

3.3 Action verbs for the content analysis

Learning objectives communicate the expectations of both the instructor as well as the learner. Consequently, the learning objective has to identify the learning outcome, the appropriate depth or detail of ‘Problem’ or relevant topic to be instructed, and how the learner would be able to use the acquired knowledge.

Action verbs may be used to indicate the depth of understanding, expected from the learner. With the simple definition of the four phases (components) or abilities of Merrill’s model, several action verbs can be taken from the literature. For the purpose of arriving at the appropriate action verbs according to the four phases of David Merrill’s model the categories are simplified and defined according to the practical situation. Those action verbs are then used as the bag-of-keywords to categorize the text.

Activation (Where do I start)

Does the instruction direct learners to recall, relate, remember, repeat or recognize the knowledge from relevant past experience that can be used as a foundation for the new knowledge(problem)?

If learner have limited prior experience, does the instruction provide relevant experience that can be used as a foundation for the new knowledge?

Based on the above questions a set of actions verbs (Bag-of-keywords) for this phase are taken from the literature and used for categorizing text / document of this category.

The system process with the following observation:

- The term set uses bag-of-keywords for Activation phase
- Algorithm called FPI to find whether the document belongs to activation phase or not

In The proposed system the term is any sequence of characters separated from other terms. The term set associated with the activation phases defined by the FPI can be used for the task of classification. A well selected subset of the set of all term set can be considered for the classification of the document.

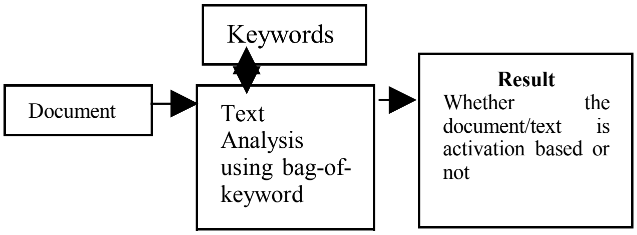
Keywords
Document
Text Analysis using bag-of-keyword
Result
Whether the document/text is activation based or not

Fig. 2 FPI based Text Mining System’s Architecture

Let D={ d1,d2, … d3} be a database of text document and T be the set of all terms related to the activation based action verbs occurring in the document D.

The following parameters were used

D → the number of documents

S → the number of sentences in the document

T → the number of unique term set that belongs to the activation phases

Tf → term frequency

Per→ Percentage of the activation phase per document

And The Steps involved in the FPI based algorithm is

1. Select term set of keywords (bag-of-words) based on activation phase ‘T’

2. Find the term frequency (tf) using the term set ‘T’ for each of the sentences ‘S’ in the document ‘D’

3. If the match does not encounter with the activation term set keyword allow the user to make the decision based on the Input

An implementation of extraction system based on this algorithm need to address the following points

- Which set of keywords need to be used as threshold parameter for the classification based on FPI's activation phase

- How should we resolve undefined cases?

4. Conclusions

This paper has presented a categorization algorithmic approach using FPI's activation based term set, which automatically discovers word sense from text. We also present a quantification methodology for measuring the percentage of activation feature supported by the document. Our manual evaluation for the categorization and quantification agreed with 80% of decision made by the automatic evaluation. The system can be applied to all or a specific portion of a text document. In the future, we will improve this algorithm further to categorize and quantify the text document using the other phases of FPI.

References

[1] Vikram pudi & P. Radha Krishna . "Data Mining"
[2] Salton G, McGill M. Introduction to modern Information Retrieval, McGrawHill,1983
[3] Tennyson R., Schott F. Seel N., Dijkstra S.(1997) Instructional Design: International perspective: Theory, Research & models.(Vol1) Mahwah,NJ: Lawrence Erlbaum Associates.
[4] Educ INF Technol(2009) 14:105-126 DOI 10.1007/s10639-008-9078-4 Categorizing computer science education research. Mike Jay, Jane Sinclair, Shanghua sun, Jirarat Sitthiworachart, Javier Lopez, Conzalez
[4] Amershi, S., Conati, C.(2006) Automatic Recognition of Learner Groups in Exploratory Learning Environments. Proceedings of ITS 2006, 8th International Conference on Intelligent Tutoring System.
[5] Merceron, A., Yacef,K.(2008) Interestingness Measures for Association Rules in Educational Data. Proceeding of the First International Conference on Educational Data mining.
[6] http://www.eurojournals.com/ejsr_22_2_10.pdf
[7] Moodle **http://moodle.ord/last consulted march.02.2008**
[8] **http://www.ibm.com/developerworks /data/techarticle/ dm_0809sigh/index.html**
[9] **http://www.autonlab.org/tutorial/.Retrieved**
[10] **http://en.wikipedia.org/wiki**
[11] **http://www.sciencedirect.com**

First Author M. Pushpa is pursuing PhD in computer science at Bharathiyar University, Coimbatore, Tamil Nadu, India. She is currently working as an Assistant professor in a reputed institution in India. Her area of interest is Artificial Intelligence, Data mining and Software Engineering.

Second Author K. Nirmala received her PhD degree in computer science from the university of Madras. She is at present an Associate Professor in the Department of Computer Science at Quaid-E-Millath Government College for Women in Chennai, Tamil Nadu, India. She has authored and co-authored many papers in an international Journal and international conferences.