Semi Automated Text Categorization Using Demonstration Based Term Set

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

1. Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore,
push_surya@yahoo.co.in

2. Associate Proffessor, Department of Computer Science, Quiad-e-Millath Government Arts College for women, Chennai,
nimimca@yahoo.com

Abstract

Manual Analysis 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. In the current scenario, the major problem is with text categorization because of the high dimensionality of feature space. Now-a-days there are many methods available to deal with text feature selection. This paper aims at such semi automated text categorization feature selection methodology to deal with a massive data using one of the phases of David Merrill's First principles of instruction (FPI). It uses a pre-defined category group by providing them with the proper training set based on the demonstration phase of FPI. The methodology involves the text tokenization, text categorization and text analysis.

.

Keywords :

Text mining, Text characterization, Feature Selection, Text tokenization, FPI and Instructional phase

INTRODUCTION

Data mining technique are designed to operate on structured data bases. When data is structured it is easy to define the set of items and hence, it becomes easy to employ the traditional data mining techniques.

In the current scenario a large portion of all available information like books, magazines, articles, research papers, products, manuals, memorandums, emails and web content contains textual information and in the form of unstructured textual data and in natural language. The amount of text is simply too large to read and analyse efficiently.

Due to the continuous growth of the volume of text data automated extraction of implicit, previously unknown, and potentially useful information becomes more necessary to properly utilize this vast source of knowledge[1]. Identifying individual items or terms is not so obvious in a textual database. Thus unstructured data places a new demand on data mining methodology known as **text mining** have to be developed to process the unstructured textual data to aid in knowledge discovery.

DOI : 10.5121/ijcsea.2012.2408

In this paper we present a naive approach to deal with text feature selection based on bag of key words associated with the demonstration phase of First principle of instruction.

TEXT MINING

Text mining is the process of knowledge discovery from textual databases as well it is a process to create a technology that combines a human linguistic capability with the speed and accuracy of a computer. It aims to analyse more detailed information in the content of each document and to extract interesting information that may be a trigger for useful actions and decision making. Text mining can also be used for analysing text for particular purposes and involves looking for regularities, patterns or trends in natural language text.

TEXT CLASSIFICATION / CATEGORIZATION

Text classification is the process of classifying document into predefined categories. Manual text classification is the process of classifying documents one by one without any inhuman expertise. Because of the continuous growth in the recent information system the volume of documents continues to grow and the manual text classification/ categorization becomes a very tedious process.

Automatic text classification is the process of classifying/categorizing the text document into the most appropriate category by employing proper training term set. The categorization system is usually based on supervised learning or unsupervised learning or methodology. Text categorization system attempts to reproduce human categorization judgement [11].

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.

ROLE OF TEXT CATEGORIZATION IN TEXT MINING

- **Text categorization** is one of the important techniques for handling and organizing text data.
- Text categorization is the task of classifying text documents into categories or classes based on their content.
- The concept-centric nature of documents is also one of the reasons why the issues of document categorization are particularly challenging.
- It plays a vital role 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 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 be 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 motivation for the work described in this paper is the categorization of documents based on semi automated concept in addition to the keywords. The use of concepts for text categorization increases its overall performance specifically when considering categorization of domain specific corpus.

In the proposed semi automatic text classification model we made an attempt to classify/categorize the text document into the most appropriate category by employing proper training term set based on the Demonstration Phase.

FIRST PRINCIPLES OF INSTRUCTION

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.

INSTRUCTIONAL PHASE

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

&

4) Integration or these skills into real world activities.[10]

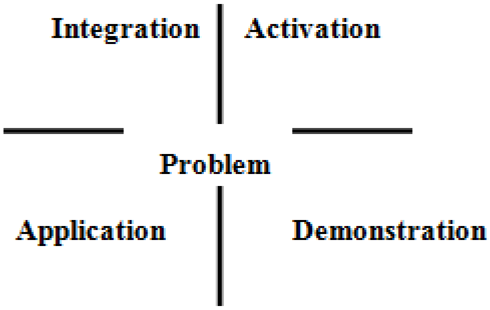

Fig. 1 First Instructional Principle

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.

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 item. In order facilitate this task, most text mining system characterize their resulting text with various kinds of annotations. Keywords are helpful in the categorization process.

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 can be generated by the human judgment based on the repeated analysis on the text. The algorithm is used to examine the first instructional principle with the help of the keywords.

New knowledge is demonstrated to the learner through this cognitive portrayal. Concept learners learn when the media (Say instructor or textual material or e-content media like voice or video) demonstrates what is to be learnt, rather than merely telling information about what is to be told. The learner observes or perceives while this portrayal is active. For example learning the principles involved in stack or queues. The media used in the process is expected to play a relevant instructional role. Explain with examples, understand information with meanings, predict consequences, order, group and infer causes are some samples for demonstration focuses the learner's attention on relevant information and promotes the development of appropriate mental models. It shows actions in a certain sequence, which can simplify complex tasks and facilitate learning.

We manually created a list of words from one of the four phases of FPI David Merrill's properties (Demonstration) based term sets for the text categorization; then we supplemented with the bag-of-words selected to categorize the textual contents. The system is implemented with a set of process as parsing and tokenizing.

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

Demonstration (Don't just tell me, show me!) Term Set

i) Does the instruction demonstrate (show example) of what is to be learnt rather than merely providing information about what is to be learnt?
ii) Are the demonstrations (examples) consistent with the content being perceived?

Based on the above questions a set of concept keywords for this phase are taken from literature and used as the Demonstration term set.

S → the number of sentences in the document
TDS → the number of unique term set that belongs to the Demonstration category
Tf → term frequency
Per→ Percentage of demonstration concept available in the Document 'D'

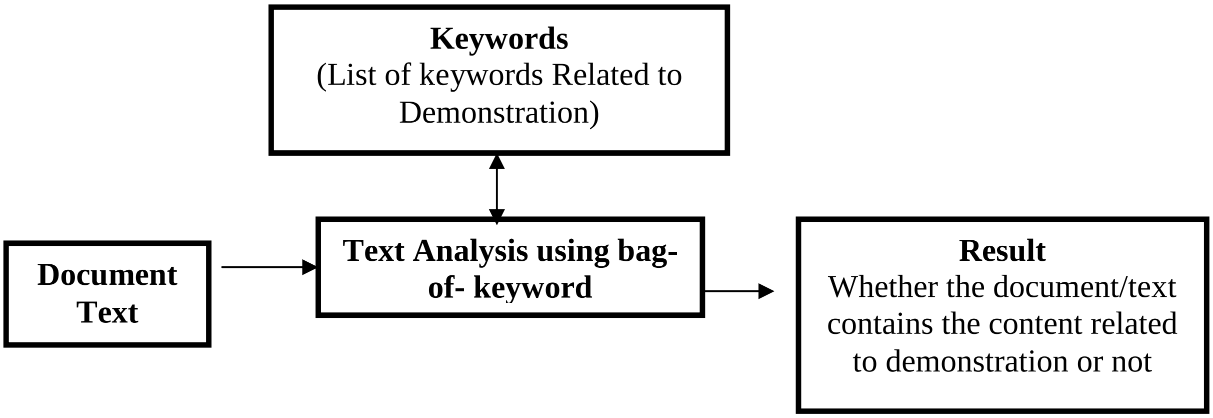
Keywords
(List of keywords Related to Demonstration)
Document Text
Text Analysis using bag-of- keyword
Result
Whether the document/text contains the content related to demonstration or not

Fig. 1 Demonstration based mining System Architecture

The system process with the following observation:

- The term set uses bag-of-keywords for Demonstration 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 demonstration phase 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.

Let D={ d1,d2, … d3} be a database of text document and T be the set of all terms occurring in the document D and 'K' categories of terms defined by the four cognitive phases. The following parameters were used

D → the number of documents

Implementation Steps

1. Select Set of terms (concept keywords) based on Demonstration ‘TDS’ and store that in a file

2. Input the document ‘D’

3. From the given document extract each sentence ‘S’ and proceed with the step 4

4. Find the term frequency (tf) using the term set in step 1 for the ‘S’. If the match does not encounter with the demonstration term set keyword allow the user to make the decision based on the sentence by displaying it on the screen

5. Repeat step 3 through step 4 until all the sentences are read from the input document.

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 clustering
- How should we resolve undefined cases?

CONCLUSION

This paper has presented a semi automated text categorization approach based on the demonstration property of the FPI. The system is used to check for the Percentage of demonstration supported by the document with the help of concept keywords. This technique requires adequate concept keywords generated based on human judgement that supports the property Demonstration. The categorization can be done only on the textual data base. The system can be applied to all or a specific portion of a document. In future the system can be extended to categorize textual document according to the properties of FPI.

REFERENCES

[1] Arun K. Pujari “Data mining Techniques”, Universities Press(India) Private Ltd.

[2] 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.

[3] Merceron, A., Yacef,K.(2008) Interestingness Measures for Association Rules in Educational Data. Proceeding of the First International Conference on Educational Data mining.

[4] Vikram pudi & P. Radha Krishna . “Data Mining”

[5] Salton G, McGill M. Introduction to modern Information Retrieval, McGrawHill,1983

[6] Tennyson R., Schott F. Seel N., Dijkstra S.(1997) Instructional Design: International perspective: Theory, Research & models.(Vol1) Mahwah,NJ: Lawrence Erlbaum Associates.

[7] 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

[8] S. Saraswathi “Design of Textual presentation from online information using hybrid approach”, ICTACT Journal on Soft Computing, Oct 2010,Vol01, Issue02,ISSN : 0976-6561(pp. 105-112).

[9] http://www.eurojournals.com/ejsr_22_2_10.pdf

[10] http://www.personal.psu.edu/users/y/z/yzx106/INSYS525/FirstPrinciple.html

[11] http://lilu.fcim.utm.md/Word_letter_compres.pdf

[12] Moodle http://moodle.ord/last consulted march.02.2008

[13] http://www.ibm.com/developerworks /data/techarticle/ dm_0809sigh/index.html

[14] Moore, A.(2005) Statistical Data mining Tutorials. http://www.autonlab.org/tutorial/.Retrieved June27,2008

[15] http://en.wikipedia.org/wiki

[16] http://aclweb.org/anthology-new/C/C00/C00-1066.pdf