Feature based Text Classification using Application Term Set

K. Nirmala
Phd, Associate Professor
Department of Computer Science
Quaid-E-Millath Government College for Women,
Chennai 600 002

M. Pushpa
Research Scholar
Department of Computer Science
Bharathiar University, Coimbatore

ABSTRACT
In the present world of information, text classification is a more challenging process due to the larger number of training cases and feature set present in text data. One of the most difficult tasks in the text classification problem is high dimensionality of the feature space. As many real world text classifications are not modeled or too difficult to model, this paper aims at the real world text classification approach or model based on one of the properties of David Merrill's First principles of Instruction (FPI). The Objective is to introduce a method to improve text classifications effectiveness, efficiency and accuracy.
In this methodology we categorizes the text using a pre-defined category group by providing them with the proper training set based on the feature of Application phase in FPI. The algorithm involves the Parsing, text categorization and text analysis.

General Terms
Pattern Recognition, Text Mining, et. al.

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

1. INTRODUCTION
A large portion of all available information today exists in the form of unstructured textual data. Manual analyses huge amount of textual data requires a tremendous amount of processing time and effort in reading the text and organizing them in required format. An Automatic text categorization process is important to deal with massive data. Now-a-days there are many methods available to deal with text feature selection because of high dimensionality of feature space.
Feature selection makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary [16]. It also increases classification accuracy by eliminating noise feature. The uses of traditional algorithm for training classifier are incapable of handling the vast amount of textual information available.
Text classification can be performed at various levels like word level, sentence level, document level and application level. This paper deals with text feature selection based on the sentence level using subset of bag of keywords. The bag-of-keywords used are the subset feature that is associated with the application phase of First principle of instruction.

2. TEXT CLASSIFICATION

2.1 Text Classification
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. In the recent information system the volume of documents continues to grow, the manual text classification becomes a very tedious process.
Text classification system attempts to reproduce human categorization judgments. One of the approaches to build a text classification system is to manually assign some set of documents to a particular class and then use inductive learning to automatically categorize it to a category of documents based on the words they contain[.
Automatic text classification is the process of classifying/categorizing the text document into the most appropriate category by employing proper training frequent term set.

2.2 Feature selection
Feature selection is the process of selecting a subset of terms occurring in the text as the training set and using that as a subset, the text can be classified[19]. It increases classification accuracy by eliminating noise feature.
Feature selection is essential because of the reasons like many features are misleading or redundant. A major characteristic of difficulty of text classification problems is the high dimensionality of the feature space. Therefore feature selection or feature extraction is the important task and crucial step for the text classification.
Pre-eminent feature selection can reduce the dimensionality of feature space and also decrease the computing complexity to improve the accuracy rate of classification. There are two approaches for feature selection

i) select features before using them in a classifier in which evaluation of features is independent of classifier where each feature gets evaluated once
ii) Select features based on how well they work in a classifier in which evaluation of feature is by how they perform in actual use where features get evaluated iteratively.

In this paper we select the feature based on how well the work to classify the textual document. The feature set used is based on the Instructional Phases defined by David Merrill known as FPI.

3. INSTRUCTIONAL PHASES

3.1 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 [12].

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.

3.2 FPI - Instruction Phases

Present 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

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.

4. 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 are 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.

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[10]. The concept learner applies their acquired new knowledge to a numerical problem. This is the practice phase, where learners are required to use their knowledge and skill to solve relevant problem. Writing programs using 'stacks' and 'queues' for a specific problem is a good example. The purpose of a practice phase in the instructional events is to provide an opportunity for learners to develop proficiency and become experts. During application phase, cognitive processes come into play; and there is a search for meaningful patterns and mental programs occur in the learner's mind.

We select the set of keywords based on the application capability based on the semantic and syntactic functions by the human judgment. With the supplemented bag-of-words annotating application feature, we categorize the textual document. The system is implemented with a feature set of process such as parsing or tokenizing, categorizing and analyzing.

4.1 Action Verbs for the Text 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 feature based bag-of-keywords to categorize the text.

4.2 Application (Concept: "*Let me do it!*"):

- ✓ Do learners have an opportunity to practice and apply their newly acquired knowledge or skill?
- ✓ Are the application (practice) and assessment (tests) consistent with the stated or implied objectives?

Based on the above questions a set of application feature based keywords are taken and used as a feature set to categorize the textual document.

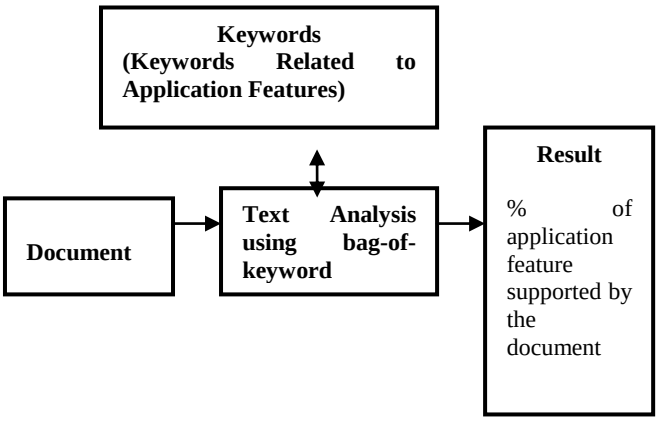

Fig 1: Application based Feature selection systems Architecture

The system process with the following observation:

The term set uses bag-of-keywords for Application phase

Algorithm called FPI to find whether the document belongs to application phase or not

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

Let D= {d1, d2, d3} be a database of text document and T is the set of all terms related to Application phase of the FPI occurring in the document D. The following parameters were used

D → Documents

S → Number of sentences in the document

T → Number of unique term set that belongs to the application feature

Tf → Term frequency

Per→ Percentage of Application phase in the 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?

Implementation steps

1. Select Set of terms 'T' (Feature term keywords) based on application features and store that in a file

2. Input the document 'D'

3. From the given document 'D' extract each sentence 'S' and go to step 4

4. Find the term frequency (tf) using the frequent term set 'T' in 'S'. If the match does not encounter with the application feature term set keyword allow the user to make 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

5. RESULT

The experiment deals with the independent textual instructional documents and the result with the manual method is been shown in Table 1

Table 1- Term frequency of instructional textual instructional document

Cognitive Portrayal	Term frequency extracted from the document in ratio of %	
	Semi Automated	Manual Method
Application	78%	100%

6. CONCLUSION

This paper is a simple demonstration for the feature selection approach of a semi automated text categorization methodology based on the application property of the FPI. The system is used to quantify the application feature supported by the document with the help of feature term set of keywords. This technique requires adequate set of keywords to support the concept of application in the document and those feature set keywords are generated with the help of human judgement. The system can be applied to the textual document of any size. In future the system can be used to analyse the learning material or any textual document with various features based on the properties of FPI.

7. 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] Manisha Pravin Mali, Mohammad Atique, "A review of Text Classification using Fuzzy logic", Proceeding of the International conference on Mathematics in Engineering and Business Management, Vol.2, pp.324-329, March 2012.

[9] Sadanandam Manchala1, D. Chandra Mohan & A. Nagesh "Word and Sentence Level Emotion Analyzation in Telugu Blog and News" **,** International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.3, June 2012 (pp.184-197)

[10] S. Saraswathi "Design of Textual Presentation from online information using hybrid approach", ICTACT Journal on Soft Computing, Oct 2010, Vol.01,Issue 02, ISSN:0976-6561(pp.105 -112)

[11] http://lvk.cs.msu.su/~bruzz/articles/classification/lewis94comparison.pdf

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

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

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

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

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

[17] http://nlp.stanford.edu/IR-book/pdf/13bayes.pdf

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

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

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

[21] http://nlp.stanford.edu/IR-book/html/htmledition/feature-selection-1.html

[22] http://www.ml.cmu.edu/research/dap-papers/ghani_ecoc-report.pdf