CHAPTER II
LITERATURE SURVEY

2.1 ON THE NEED FOR AND APPLICATION OF INDEPENDENT, NAÏVE BAYES AND OTHER PROBABILITIES

Rules that depend on multiple keywords may not be flexible in sentences having combined keywords (Kim et al. 2012). They are also no robust to typographic errors. Hence erroneous results occur in automated procedures. Information based on combined words was extracted under domain specific applications, namely words occurring along with grant numbers (a key piece of bibliographic information). Detailed descriptions about grant numbers were considered for the purpose of extraction. An important observation was made like “Naïve Bayes theorem was used in such an application, as Naïve Baye’s classifier is based on statistics where several words used in sentences in complicated situations such as typographic errors” (Rish 2001). The ability of resolving such complex situations by Naïve Bayes was utilized successfully. Situation such as relevant and non-relevant of words in sentences was considered. Naïve Bayes classifier handles such complexity. The authors considered the conditional rule like the information such as organization; support word and grant format are tied with the grant numbers. The performance of the probability values under three measures was considered. Naïve Bayes classifier depends on statistical input with several words in the zones. The results proved to be reasonably accurate after overcoming situations such as typographic errors. Naïve Bayes is also used in situation where classification of texts is needed fast and easy to implement algorithms (Rennie et al. 2003). To find ways to improve the learned decision boundary weights, Naïve Bayes classifier was used (Ting 2011). A complement class was introduced to balance the amount of training examples. However problems with Naïve Bayes classifier have also been enlisted: There is a systemic problem with Naïve Bayes, as features are assumed to be independent. This results in a situation where each word of a sentence could independently contribute evidence. It was suggested that weights could be used with magnitudes worked out proportional to the dependencies either strong (more weight) or weak (less weight). To keep the classes with more dependencies, normalization technique was adopted on the classification weights. Results from Naïve Bayes were compared with real world information of text documents.
Comparative study results on feature selection methods in statistical learning of text categorization have been reported (Yiming Yang and Pedersen Jan 1997). Five methods, namely Document term frequency, Chi square, Information gain, Mutual information and Term strength were compared. It is reported that one of the major characteristics or difficulties of text categorization problem is the high dimensionality of the feature space. A caution note has been presented: “Naive Bayes model would be computationally intractable unless an independence assumption among features is made”. In the case of document term frequency, a threshold value on the document frequency was assumed and the document frequency for each unique term in the corpus of training terms was computed. This was removed from the feature space of those terms were less than the threshold value.

Relevant and irrelevant sentences of mathematics textual documents have been categorized using joint probabilistic model (Suleyman Centintas et al. 2010). An important observation noted namely “Intelligent tutoring system provide effective individualized instruction through adjustment of characteristics dynamically”. Adjusting components vary from difficulty levels, type of instructional flow and context etc. These kinds of adjustments cannot be manually done as analysis of student created problems puts a high burden on teachers. Math word problem types have been tried out to identify student created problems. This eventually leads to rhetoric concept identifications. This is an important observation for the research. Certain systems developed and reported in literature indicate that such systems would not handle problems with excessive verbiage or implicit information. In mathematics education, word problem is a term that is used to refer to numerical exercise on which significant background information is presented as text rather than mathematical notations. User generated problems have been analyzed through math word problem types in a rhetoric concepts (Cetintas et al. 2009). An important finding is: Correlation between sentence and question sentences in math problem significantly outperforms the traditional text categorization. Independent probability has been applied for bag of textual words in documents. Utilizing the correlation between sentences and question sentences along with relevant and irrelevant sentences with another irrelevant sentence could be correlated. In such an attempt, joint probabilistic model has been proposed.
2.2 ON FEATURES AND TEXT CATEGORIZATION

Clustering is needed for organizing large quantity of disordered text documents into meaningful and comprehensible representations (Anna Huang 2008). Partition of clustering has been demonstrated to be more useful rather than hierarchical clustering for processing large data set. Small number of consistent groups of document set is better than maintaining large number of unorganized sets for extractions (Jain et al 1999). Clustering in the form of meaningful grouping has been demonstrated to be more advantageous. This kind of clustering can benefit further analysis for information retrieval and extractions (Gerald Kowalski 1997).

Many methods are found in the literature on text categorizations with features. Gini index theory has been adopted for feature selection and an algorithm created (WenqianShang et al 2007). This theory is a non-purity split method, that involves sorting of binary systems and fits continuous numerical values. Gini’s minimum value is zero that means all members in the set belong to a same class. When all the samples in a set distribute equably for the class, then Gini will become maximum. This ensures maximum extraction of useful information. An important observation is that, ‘some words in a text document that do not appear have contributions to judge the class of text, but such contribution is far less significant when compared with the effort to consider the words that do not appear, as feature selection inclines high frequency words. Improvements on the Gini feature selection method have been made. Experiments showed that the improved Gini method performed better than many other methods and also demonstrated simpler computation.

Noise in text categorization refers to unwanted words other than key concept words after removing unnecessary stop words and not required stemming of words. The emerging trend of noise detection in textual data of systems is to perform on documents that are having linguistic and statistical features (Kontostathis et al 2009). The importance of human experts for reviewing systems for noise separation has been stressed. It was suggested to arrange streams of data of the concept words in the form of threads joined together with topically related material. Most of the learning algorithms per se do not imply exact concept extraction or understanding, but they would rather aid experts so as to view the data for confirming the concept. Hence for the purpose
of selection, such concept keywords can be listed automatically by algorithms. These automatic procedures may be used for selecting in supervised environments, for the purpose of selecting topics. Such automatic procedures can be used for deciding on which or how many of those concept words may be considered for comprehending the concept, and also for computing frequencies of occurrence.

A new concept based on document clustering that is based on repeated co-occurring of words has been reported (Wenquan Shang et al. 2007). The procedure for deriving concept words data before applying term frequency computation was enlisted. First step will be to remove unwanted stop words and then unnecessary stemming done before addressing synonyms. Such gathered words are then clustered. High frequency and low frequency words are then removed before analyzing the document for term frequencies and inverse document frequency.

Semantically closer synonym words cannot easily be distinguished with each other through automation, and a system has been suggested to list good synonym words or near synonym words from a given word through automation (Senellart et al. 2009). In view of this difficulty with semantic synonymous, many existing techniques do not treat semantically closer words, but offer words that are only similar in nature. When semantics were used in any automatic learning domain, a shift in perspective may be needed from information based content delivery system to a knowledge based system (Feng Tao et al. 2005). For the purpose of categorizing documents into hierarchy of topics, the document contents could be manually analyzed or through automatic procedures (MaluCastellanos 2009) although the manual method is most successful. But however for huge documents, manual methods used to be difficult and it needs supervision of domain experts. Even though many automatic efforts have been tried out and published in the literature, none of the automatic approaches has yielded results that are comparable with the ones obtained through manual categorization. It was hence recommended to analyze user logs for recording historical information, requested by them, instead of mining the entire contents of the document collection.
2.3 ON CONCEPT EXTRACTION

A concept is an entity which allows identifying from a set of queries that appear together in a repeated manner (Bruno Fonseca 2005). A concept could describe a synonymous relation with a concept that could describe a specific semantic relationship or a concept could be described in more generic information. Therefore, users that query using repeated terms can be identified and termed as concept keywords. The query techniques have however be limited to mining, identifying and labeling query relations.

Concept keywords could be derived from identifiers of software coding and applied with tf/IDF techniques (Masaru Ohba and Katsuhiko Gondow 2005). This technique however may not be accurate with natural languages. Instead of tf/IDF, concept keyword term frequency was developed instead of mere keyword term frequency. A concept keyword is a word that represents a key concept. Yet there may not be a clear cut definition of a concept keyword, since it is highly a subjective matter. Concept keywords could also be specified by human intervention in a supervised environment. Machine extracted concept keywords could be approximated and compared with human supervised ones.

For retrieval of near concepts content organization based on global indexing has been suggested (Wen-Chung Shih et al.2009). An ontology based semantic search has been adopted, for the purpose of indexing. A layered architecture for content-based information retrieval, so as to achieve precise retrieval of compliant data has been tried out. In this architecture, the query submitted by any user is used by the search engine on a global index for obtaining a list of desired documents. These documents are then kept in pre designed learning object repositories. The search engine then ranks them according to the similarity found between the query and the document of the repository. This methodology has worked well for retrieval in a fast and precise manner, for large number of learning contents. For translating knowledge into a process, ontology may be used. Human efforts may be needed to tag the resources so as to facilitate knowledge processing (Alaba Olumuyiwa et al.2008). Two types namely content and activity ontology were adopted for building instructional modules. It was concluded that semantics can be applied not just for displaying the words, but for automating, integrating and reusing the data for services across various applications.
Text classifications, particularly ‘words’ may be needed for extracting information, understanding of content based systems and transforming text to produce meaningful summaries (Kamruzzaman et al 2010). However relationships of such words in the form of rules may be needed for extracting features of the textual contents rather than pure words. Concept keywords may be formed as identifiers for computing frequencies using tf/IDF techniques (Masaru Ohba and Katsuhiko Gondow 2005). A concept keyword is a word that represents a key concept. However there is no clear cut definition for a concept keyword, as it is highly subjective. But a concept is one, that allows identification from a set of queries that appear together repeatedly (Bruno 2005). A concept could describe a specific semantic or it could describe more generic information or it could merely describe a synonym relation. Yet, the above techniques may not be accurate when compared with natural languages.

Combinations of small words called ‘quad grams’ have been tried out for characterizing texts for concept extraction (Ingrid Renz et al. 2003). A general measure of concept extraction based on pure keywords can hardly be termed as concept, as concepts are very subjective. Inspecting keywords of concepts showed that among top 10 keywords, around 30% were found to be quad gram-based keywords.

2.4 ON RHETORIC CONCEPTS

Instructional objectives and learning tasks constructed with rhetoric words are interdependent (Akpinar et al 2004, Friesen 2005, Krauss and Ally 2005, Churchill 2007). In some particular cases such as e-contents the learning tasks may be complex, but they are useful in arriving at learning objectives with rhetoric words. Rhetoric words that facilitate learning concepts may be utilized in online learning material (Nesbit and Li 2004, Nurmi and Jaakkola 2005). The effectiveness of rhetoric words in learning contexts needs to be evaluated through social studies from stakeholders, as learning materials consist of discrete units of learning rhetoric, founded on established standards (Ian Douglas 2001). The instructional design implications of these learning materials are needed to be examined so as to determine the extent of adaptation of instructional design methodologies (Parrish 2004, Nurmi and Jaakkola 2006). However, development of such learning materials in various contexts requires applications of
pedagogical knowledge (Hwee-Reei Chong and Chien-Sing Lee 2006). Learning theories are hence needed to be studied prior to understanding pedagogical designs of instructional materials. Some theories address sequencing of materials for effective instructions.

Application of taxonomical words, for analyzing learning materials that found wider acceptance in the instructional design community, has been suggested (Gagne Robert 1985). However such words need not be unique for a particular instructional or rhetoric condition. Whereas for each category of learning objective, unique condition for learning have been suggested (Merrill 1994). Such mental condition could be based on information processing theory, i.e. input – process – output. Different types of knowledge and skill are required for different conditions of learning. ‘Conditions of Learning’ theory emphasizes on different activation phases, where new knowledge builds on the learner’s existing knowledge. Merrill has extended this theory with ‘Component Display Theory’ and this part also is based on information processing theory and therefore it depicts a reclaiming nature in the instruction process (Merill 1987). The approach of putting a real-life problem into the centre of an instructional episode is particularly suited for problem-based subject learning (Nordhoff Helga 2002). This combined approach as model is known as ‘First Principles of Instruction’ (FPI), which emphasizes the following stages: the use of a real-life problem situation in the central of an instructional event; the importance of activating previously learnt knowledge of the learner; the role of demonstration followed by guided problem solving event and integrating the new knowledge with existing knowledge.

Users who would like to query for concepts, usually provide short and ambiguous words (Bruno Fonseca et al. 2005). For better answers to such queries, the queries need to be improved. A query expansion technique that improves the ambiguous queries has been proposed. In a supervised manner, the concepts related to the current query are shown to the user for feedbacks on his choice. Past queries that were stored and compared with current query using associated rules. The definition of a concept (as per the authors) is that allows identifying groups of queries that appear together repeatedly. Four types namely synonyms; specialization; generalization and association are used for classifying the concepts. Whenever a vague query is placed by the user that has occurred in the past, the user is provided with reasonable and meaningful suggestions through the proposed framework.