**Information Retrieval Reading Assignment 2**

**The research problem and its significance:** Number of documents are growing in number and available online, so document information or data retrieval gets tougher without having indexing and content classifications. To avoid this, text categorization which categorize the documents based on their content is used. The research problem is the dimensionality reduction in feature selection techniques of text classification. Through dimensionality reduction we can reduce the unwanted dimensions and derive the new dimensions by combining existing dimensions to increase the performance and accuracy of text categorization.

**The research challenges this paper wants to address:** The research challenges is reducing the high dimensionality by selecting the features (variables) of documents for text categorization through removing the non-influential features without loss of information for categorization.

**A brief description of the approach:** To reduce the dimensionality, research paper assessed five methods like selection of Termrelay on **document frequency** [DF], **Mutual information** [MI], **Information Gain** [IG], **Chi-square [**CHI**]** and **Term Strength** [TS]. In Document Frequency thresholding method, the dimensionality reduction is achieved by removing the terms with low frequency documents considering as removing rare terms will increase categorization accuracy. while analyzing the text categorization techniques on large bulk of documents, IG and CHI-square it shows productive results. In IG, dimensional reduction done by calculating information gain for each term then whichever have low IG than a defined threshold are removed. In MI, rare terms are identified by calculating mutual information between terms and category. More than common terms conditional probability value is higher for rare terms .In CHI and MI main difference is using normalized values and terms for same category are compared in CHI-square. TS uses similarity functions to determine rare terms. IG uses k-nearest neighbor to remove 98% of unique terms which gives best accuracy of text categorization. IG,DF,CHI term values shows strong correlations. It indicates that DF have low computational cost and more reliable than IG or CHI.TS gives 50% term reduction but for when it applied on more terms it gives less percentage. Due to rare terms favoring and sensitive to probability estimation errors, MI gives comparatively bad performance than others. Classifiers used in Text categorization are: 1) KNN: K-nearest neighbors uses Cosine similarity to measure the similarity among documents and assign the predefined category to the document. 2) LLSF: Linear Least squares fit classifier uses regression model to words for predicting category weights.

**Unique contributions:**  Common terms are in fact more effective for text categorization, which gave IG,DF,CHI-square a very good performance. It is not possible for KNN,LLSF to have best performance of categorization, if the large amount of informative data is lost by dimensionality reduction. IG calculate number of bits based on existence of term in document. The strong correlation between Document Frequency and Information gain shows common terms are important in text classification which contradicts the statement common terms are not much useful for categorization. Categorical variables or data does not look like more important for good performance of feature selection process. DF does not use categorical variables for feature selection but shows similar performance as of IG and CHI-square.

**Strengths and weaknesses:** The strengths of the research papers are the performance of term selection in DF,IG,CHI is good and strong correlation exists between them. The Empirical results shows that common terms for text classification are important for DF, CHI, IG. It’s evident in paper that methods which shares same bias are having good performance. The paper experiments show that DF is a reliable method for selecting features. The weakness in paper are it does not explain about threshold values measuring in the methods. It does not discuss how to handles the bias in TS for to have best performance. IG and CHI a have lot of computations . MI is not suitable and underperforms for high dimensionality data and DF method does not consume space.