CHAPTER 7

Conclusion and Further Work
Chapter 7

Conclusion and Further Work

7.1 Conclusion

In the present day world information exists in the form of text in various electronic media. So much of the documentation work is going on every minute around the world. This ocean of text has to be categorized accurately, effectively and in a faster manner. Hence finding an algorithm for doing this assumes higher priority. This forms the basis of this work.

Initially three feature selection algorithms were compared as part of the pre-processing task for the text categorization. The feature selection algorithms considered here were LVF, LVI and Relief. After this comparison LVF was found to have the better performance. Hence LVF algorithm was used implementing text categorization algorithms. Then the four text categorization algorithms (NB, NGrabs, KNN-SA, BPN) were analyzed and compared. Out of these four algorithms NGrabs showed promising results. Then a new algorithm called NGrabsSA was proposed. The comparison of this algorithm with the NGrabs algorithm shows that NGrabsSA algorithm produces better results.

7.2 Findings

The research work speaks about the various aspects of text categorization. In this work the pre-processing work required for the text categorization work is analyzed initially. This pre-processing work involves stop words removal, stemming and feature selection. After this the text categorization algorithms Naive Bayes, Back propagation
network, K-Nearest neighbourhood and NGramps were analysed and they were compared based on their performance. Finally a new algorithm called as NGrampsSA was designed and its performance was compared with that of the NGramps algorithm on the basis of execution time and precision.

The overall findings of this research work are provided here. The findings talk about the various aspects like pre-processing, text categorization algorithms and especially NGrampsSA algorithm. Also it has mentioned certain drawbacks in the algorithm. All these aspects become important in the point of view that text categorization is an emerging and a very relevant field in Information Technology. Following are the overall findings of this research work:

1. Feature Selection is a vital pre-processing step for the Text classification. Only if a better Feature Selection method is adopted along with the stemming and stop words removal, it is possible to get a high quality data for classification.

2. BPN is a very popular algorithm used for the text categorization. The performance of this algorithm is considered to be satisfactory with reference to the dataset used. But the performance of this algorithm depends on so many other factors like the dimensionality reduction. Also the activation function and the momentum terms play an important role in the performance of the BPN algorithm.

3. After comparing the four algorithms (NB, BPN, NGramps, and KNN-SA), NGramps algorithm is found to consistently produce lesser execution time than the others. BPN algorithm has slightly higher execution time than NGramps algorithm. Naive Bayes Algorithm has higher execution time than NGramps and BPN algorithms in most cases. It can also be noted that KNN-SA algorithm mostly has a very high execution time than rest of the three algorithms.
4. Considering the precision values produced by the four algorithms, NGrams algorithm has more precision on most cases. In 25% of the cases BPN has higher precision value than NGrams algorithm. In 20% of the cases KNN-SA has higher precision value than NGrams algorithm. In 35% of the cases Naive Bayes algorithms has better precision value than NGrams, BPN and the KNN-SA algorithms.

5. NGramsSA algorithm’s performance was compared against the NGrams algorithm. The execution time recorded for NGramsSA was smaller than the execution time of NGrams algorithm for all the cases given in chapter 6. In the execution time the variation between the NGrams and NGramsSA ranges from 0.06 second to 1.25 seconds.

6. NGramsSA algorithm’s precision values are found to be higher than the precision values of the NGrams algorithm’s precision values for the cases given in this chapter. The variation in the precision values of the NGramsSA and NGrams algorithms has a range between 0.039 and 0.051. If the average of the precision values is considered NGrams has 0.75 and NGramsSA has 0.80.

7. Going by the definition of n grams, there are more n grams in a given passage than there are words. This gives greater opportunities to collect enough n grams to be significant for matching.

8. In NGrams and NGramsSA, n grams frequency is used. This is found to be a language independent approach. It means that whatever be the language, English, Tamil, Chinese etc, this frequency approach will work.
9. NGrams and NGramsSA algorithms have certain advantages like working equally well with short and long documents and it has minimal storage and computational requirements.

10. The NGrams and the NGramsSA algorithms have certain problems too. A single misrecognized character throws off the statistics for a whole word. This may cause problems in finding the frequency of the n gram and in the calculation of the weight. This will have an effect in final categorization process.

11. Another difficulty in NGrams and NGramsSA is that when short passages in a document are considered they are too short to get representative subject word statistics. This also creates problem categorizing the text.

7.3 Suggestions

Before talking about the suggestions for the future work there are certain things that should be considered. One thing that should be emphasized here is that Text categorization is a fundamental task in document processing, allowing the automated handling of enormous streams of documents in electronic form. Though this is a basic task in text handling it has many difficulties. One of the important difficulty is that the quality of the document that should be classified. By saying quality of the document it means that the errors of different forms that may be present in the document. Spelling errors, grammatical errors are some of them. The text categorization process should carefully overcome these difficulties in a manner it categorizes the right document. The algorithms that are produced for categorizing the text should include these ideas.

As it has been emphasized many times in this work, text categorization is a very essential process for text document management. A poor text categorization method will
affect the text management processes like email sorting, topic identification, structure search and/or browsing etc. In this context there is a constant need to perform text categorization in a faster, error free and effective manner. Apart from forming new efficient algorithms to solve the text categorization problem there is always a need to review the methods performed in the earlier algorithms and pre-processing steps used. In a dynamically changing environment the new methods, new algorithms which will give optimum benefits, have to be found in a lesser time. Hence this research work demands further expansion in text categorization tools, methods and algorithms in the future.

Certain areas have been identified for doing further work in text categorization and they are listed below:

1. An important issue is that the quality of the training set that affects the matching performance. This is because the training set may not have included the current or “hot” topics in the particular category. For example, in games category the training set may be old and might not have included the current terms in use. To avoid this, the training data set should be updated and a testing should be conducted after this.

2. Another issue that can pose problems is that is about how to categorize an interdisciplinary document. There will be confusion in finding to which category the particular document actually belongs. To eliminate this, a good normalization scheme should be devised for these types of documents. This scheme will produce some sort of absolute measure in identifying how good a particular measure is. This normalized score will help in rejecting certain documents in a particular category.
3. The algorithm NGramSA has been tested in Pentium IV system which has clock speed 2.4GHz. Also it was implemented in VB.NET and SQL Server2005 as front-end and back-end respectively. Windows XP Professional was the operating system used. Now it is time to check whether the algorithm produces the same performance in the new technologies or there exists performance degradation.