ACKNOWLEDGEMENTS

My special thanks go to Dr. Ruvan Weerasinghe of the University of Colombo, the principal supervisor of my project for all the support and guidance extended to me. I found his expertise and suggestions extremely valuable. My sincere gratitude is also extended to my supervisors Ms. G.S. Makalande and Mr. P. Dias for all their support and guidance. In addition, I would like to thank our coordinator of the Master Degree programme, Dr. Sunethra Weerakoon for her encouragement and support. I also would like to thank LAcNet in helping me to obtain access to news items from the SLNET archives.

Finally, I am very grateful to Mr. Piyasiri Vithanage, the Head of the Department of Economics, University of Ruhuna for his support and encouragement.

ABSTRACT

The advancement of information technology in the modern world has contributed towards enhancing the quality of life of people around the world. To keep pace with this rapid development, it is important to have links with the enormous network called the Internet. This enables people to have access to information resources, keep abreast of news, send timely E-mail, and have interactive remote conferences. However, a tremendous task facing information consumers today is to identify relevant news items speedily. Hence, designing an information filter for users of Internet news bulletins, is a dire need of the day. The main thrust of this study is, therefore, focused on designing an algorithm to identify news items available on the Internet and categorizing them according to their degree of similarity to each other.

The main concept exploited in obtaining a metric for computing the degree of similarity of two news items is based on calculating and comparing the percentage of proper nouns common to both news items. In order to extract proper nouns from a news item, a filtering process is employed to eliminate pronouns, articles, prepositions, "Be verbs",

determiners and other function words. Subsequently the frequencies in which the extracted words have occurred in the news item are calculated and analyzed. Statistical methods are used to confirm the above results before they are presented to the user. The proposed algorithm was tested and favourable results were obtained by using the news items downloaded from the LAcNet Sri lankan news archives available on the Internet.

Values of degree of similarity obtained for the test data was compared with human classification. Based on these results, it is demonstrated that the proposed algorithm is able to categorize news items into two classes, 'similar' and 'different', successfully. This achievement makes a significant contribution towards achieving the automatic categorization of news items available on the Internet into various topics.

CONTENTS