AN ANALYSIS OF HIERARCHICAL CLUSTERING AND NEURAL NETWORK CLUSTERING FOR SUGGESTION SUPERVISORS AND EXAMINERS

NURUL NISA BINTI MOHD NASIR

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

NOVEMBER 2005

Especially for Ayah & Mama…

Thanks for your love, guidance and blessings...

For my brothers & sisters… Along, Alang, Ateh & Adik…..

You are the best in the world…

For my Man…whom always there to support me…

Thanks for everything…

ACKNOWLEDGEMENT

Profound acknowledgement and special thanks to supervisor, Associate Professor Dr. Naomie Salim for encouragement, critics and friendship.

Thanks to my family, especially my beloved dad and mom (I know you are hearing me), for understanding and encouraged to continue my studies.

Not forgetting also to my college and friends for their support and assistance in helping me to achieve this major milestone of my life. Lastly, my thanks to Herman, for the continuous support throughout the years.

ABSTRACT

Document clustering has been investigated for use in a number of different areas of information retrieval. This study applies hierarchical based document clustering and neural network based document clustering to suggest supervisors and examiners for thesis. The results of both techniques were compared to the expert survey. The collection of 206 theses was used and employed the pre-processed using stopword removal and stemming. Inter document similarity were measured using Euclidean distance before clustering techniques were applied. The results show that Ward's algorithm is better for suggestion supervisor and examiner compared to Kohonen network.

ABSTRAK

Dewasa ini, kaedah pengelompokan dokumen banyak diaplikasikan dalam bidang Capaian Maklumat. Kajian ini akan mengadaptasikan pengelompokan dokumen berasaskan Rangkaian Neural dan juga Pengelompokan Timbunan Berhirarki. Hasil pengelompokan ini dianalisis bagi mencari kaedah terbaik dalam pemilihan penyelia dan penilai dan dibanding dengan pemilihan yang dilakukan oleh pakar. Dokumen-dokumen yang dikelompokkan menjalani pra-pemprosesan termasuklah penghapusan perkataan yang tidak membawa makna dan mempunyai kekerapan yang tinggi atau *stopword,* pembuangan imbuhan atau *stem*, dan seterusnya pengelompokkan kata nama supaya tiada pengulangan perkataan yang sama. Seterusnya, keserupaan dokumen-dokumen selepas pra-pemprosesan akan digambarkan menggunakan jarak Euclidean. Hasil yang diperolehi menunjukkan algoritma Ward's adalah lebih baik dalam pemilihan penyelia dan penilai berbanding algoritma Kohonen.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

ANN	-	Artificial Neural Network
BMU	-	Best Matching Unit
HAC	-	Hierarchical Agglomerative Clustering
IR	-	Information Retrieval
IRS	-	Information Retrieval System
NN	-	Neural Network
PCA	-	Principal Component Analysis
RNN	-	Reciprocal Nearest Neighbour
SOM	-	Self Organizing Map
STC	-	Suffix Tree Clustering

LIST OF SYMBOL

S_{D_i,D_j}	-	Similarity between document *i* and document *j*
$weight_{ik}$	-	*k*-th weight in document *i*
$weight_{jk}$	-	*k*-th weight in document *j*
ESS	-	Error sum squares
η	-	Learning rate

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.0 Introduction

IR is a discipline involved with the organization, structuring, analysis, storage, searching and dissemination of information. A compact definition of the basic function of an *information retrieval system* (*IRS*) has been given by Lancaster, (1968):

"An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his enquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request."

Much of the research and development in IR is aimed at improving the effectiveness and efficiency of retrieval. Document clustering was introduced to IR on the grounds of its potential to improve the efficiency and effectiveness of the IR process. Jardine and Van Rijsbergen (1971) provided some experimental evidence to suggest that the retrieval efficiency and effectiveness of an IR application can benefit from the use of document clustering. The efficiency and effectiveness of an IR application was expected to increase through the use of clustering, since the file organization and any strategy to search it, take into account the relationships that hold between documents in a collection (Croft, 1978). Relevant documents that might have otherwise been ranked low in a best-match search will be (through inter-

document associations) grouped together with other relevant documents, thus improving the efficiency and effectiveness of an IR system.

1.2 Problem Background

Document clustering has been applied to IR for over thirty years. The aim of research in the field is to postulate the potential of clustering to increase the efficiency and effectiveness of the IR process (Jardine & Van Rijsbergen, 1971; Croft, 1978). The literature published in the field covers a number of diverse areas, such as the visualization of clustered document spaces (Allen *et al*, 2001; Leuski, 2001), the application of document clustering to browsing large document collections (Cutting *et al*., 1992; Hearst & Pedersen, 1996), etc.

The main motivation for this work has been to investigate methods for the improvement of the efficiency and effectiveness of document clustering. One type of clustering employed in this study is hierarchical clustering; perhaps the most commonly used type of clustering in IR (Willett, 1988). This is a choice based on the more sound theoretical basis of hierarchical clustering. Jardine and Sibson (1971), Salton and Wong (1978) and Van Rijsbergen (1979) have identified three strengths of hierarchical methods. Firstly, such methods are theoretically attractive since they do not depend on the order in which documents are processed. Secondly, they are well formed, in the sense that a single classification will be derived from a given set of documents. And finally, hierarchic methods are stable, since small changes in the original document vectors will result in small changes in the resulting hierarchies.

The application of hierarchical methods to IR (e.g. group average, complete link and Ward's methods) was extensively investigated during the 1980s. The majority of the research work was carried out at Cornell University by Voorhees (1985a) and at Sheffield University by Griffiths et al. (1984, 1986) and also El-Hamdouchi and Willett (1989).

More recently, information science researchers have turned to other newer artificial intelligence based inductive learning techniques including neural networks. This newer techniques which are grounded on diverse paradigms have provided great opportunities for researchers to enhance the information processing and retrieval capabilities of current information storage and retrieval systems.

NN is another clustering technique applied in this study. Neural network models have many attracting properties and some of them could be applied to an IR system. Recently, there is a research tendency to apply NN in cluster document. Initially, Kohonen is an unsupervised NN which is mathematically characterized by transforming high-dimensional data into two dimensional representations, enabling automatic clustering of the input, while preserving higher order topology.

In neural network models, information is represented as a network of weighted, interconnected nodes. In contrast to traditional information processing methods, neural network models are "self-processing" in that no external program operates on the network: the network literally processes itself, with "intelligent behavior" emerging from the local interactions that occur concurrently between the numerous network components (Reggia & Sutton, 1988). It is expected that the research on the application of neural network models into IR will grow rapidly in the future along with the development of its technological basis both in terms of hardware and software (Qin He, 1999).

Neural networks computing, in particular, seem to fit well with conventional retrieval models such as the vector space model and the probabilistic model. Doszkocs et al. (1990) provided an excellent overview of the use of connectionist models in IR. A major portion of research in IR may be viewed within the framework of connectionist models.

Essentially, thesis focuses solely on the retrieval effectiveness and efficiency of document clustering for suggestion of supervisors and examiners for thesis since there is not much research in this domain.

Each Computer Science student enrolled in the master program should produce a thesis before finishing his/her studies. This thesis contains a complete report of a research.

Each thesis should have at least one supervisor and examiners to fulfil the requirement. This supervisor and examiners are selected either from FSKSM lecturers or any other person in who is expert in the thesis's subject.

A supervisor is responsible to guide student in doing research, and producing a valuable research whereas examiners evaluate the yield of research and to see whether students really understand his/her research. The evaluation from the supervisor and examiners shows the quality of a student's research.

Currently, the determination of supervisor and examiner is done manually by the coordinators. However, sometimes the coordinators are new and did not know much about the experience of lecturers in supervising and examining students in various areas. The selection process based on incomplete knowledge such as this sometimes may affect the quality of thesis produced by students. The major problem related to thesis performance is the student didn't get an effective guidance from his/her supervisor because the supervisor is not the expert in the thesis's subject.

The weaknesses of such a manual system may affect the quality of research in the long term.

Therefore, in this study, two clustering techniques are used, Kohonen clustering and Hierarchical clustering to give a better solution. Clustering result will be analyzed in order to find out the best techniques for the solution. Furthermore the implementation of mathematical algorithm makes the system more concrete without bias situation.

1.3 Problem Statement

- Can document clustering be used for determining supervisors and examiners of thesis effectively?
- Can Kohonen based document clustering perform better result than Ward's clustering (one type of hierarchical clustering) for determining supervisors and examiners of thesis?

1.4 Objectives

The objective of this study is as follows:

1. To represent index terms in document vector.

2. To apply two techniques of clustering, Kohonen clustering and Ward's clustering to improve the efficiency and effectiveness of suggestion for supervisors and examiners

3. To analyze Kohonen network based document clustering and Ward's based document clustering for suggestion supervisors and examiners

4. To compare clustering techniques to use in the domain of suggestion of supervisors and examiners in FSKSM, UTM

1.5 Project Scope

Two clustering techniques will be applied in this study that is Neural Network clustering and Hierarchical clustering. The result of these two clustering will be analysed to find out the best techniques in domain study. This study will be done in scope as stated below:

1. Title and abstract of 206 theses will be stored on the machine and will be used in information retrieval process. The theses are on master thesis from FSKSM, UTM only.

2. Porter stemming will be used to reduce a word to its *stem* or root form in the title and also the abstract of thesis

3. Indexing process will create a unique identifier of the documents by counting the frequency of each index terms before the *tfidf* weighting is calculated

4. Ward's clustering and Kohonen clustering will be applied to the indexed documents.

1.6 Significance of the Project

Results of the study will show whether NN based document clustering or Hierarchical based document clustering is effective for determining supervisors and examiners. It will also give insight on whether NN based is better than Hierarchical based document clustering in terms of suggestion of supervisors and examiners.

1.7 Organization of the Report

This report consists of five chapters. The first chapter presents introduction to the project and the background of problem on why is the study is being conducted. It also gives the objectives and scope of the study. Chapter 2 reviews on IR, pre-processing to achieve IR purpose, and document clustering also clustering techniques that will be used in this study. Chapter 3 discusses on the framework of this project in detailed including pre-processing phase further clustering algorithm that will be

applied in this study. Chapter 4 contains a cluster analysis based on Ward's and Kohonen performance in determining supervisors and examiners and Chapter 5 is the conclusion and suggestions for future work.

REFERENCES

Aldenderfer, M. S., & Blashfield, R. K. (1984). Cluster Analysis. Sage Publications, Inc.

Allan, J., Leuski, A., Swan, R., Byrd, D. (2001). Evaluating combinations of ranked lists and visualizations of inter-document similarity. *Information Processing & Management*, 37(3):435-458.

Anderberg, M.R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Azuaje, H. (2003). *Genomic data Sampling and Its Effect on Classification Performance Assessment*. Nothern Ireland, UK.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, ACM Press 1999.

Borgelt, C., and Nurnbeger, A. *Experiments in Document Clustering using Cluster Specific Term Weights*,

Botafogo, R.A. (1993). Cluster analysis for hypertext systems. In *Proceedings of the 16th Annual ACM SIGIR Conference*, pp. 116-125. Pittsburgh, PA.

Chris D. Paice. (1994). An Evaluation Method for Stemming Algorithms. *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference*

Research and Development in Information Retrieval, pages 42-50, 3-6 July 1994

Cormack, R.M. (1971). A review of classification. *Journal of the Royal Statistical Society*, Series A, 134:321-353.

Croft, W.B. (1978). Organizing and searching large files of document descriptions. *Ph.D. Thesis*, Churchill College, University of Cambridge.

Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W. (1992). Scatter/Gather: A cluster based approach to browsing large document collections. In *Proceedings of the 15th Annual ACM SIGIR Conference*, pp. 126-135. Copenhagen, Denmark.

Defays, D. (1977). An efficient algorithm for a complete link method. *Computer Journal*, 20:93-95.

Dawson, J.L. (1974): "Suffix removal for word conflation," *Bulletin of the Association for Literary & Linguistic Computing*, 2 (3), 33-46.

Doszkocs, T. E., Reggia, J., & Lin, X. (1990). Connectionist models and information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 25, 209-260.

Dunlop, M. D. Development and evaluation of clustering techniques for finding people, *Proc. of the Third Int. Conf. on Practical Aspects of Knowledge Management (PAKM2000) Basel, Switzerland*, 30-31 Oct. 2000,

El-Hamdouchi, A. and Willett, P. (1989). Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal*, 32(3):220-227.

Griffiths, A., Robinson, L.A., Willett, P. (1984). Hierarchic agglomerative clustering

methods for automatic document classification. *Journal of Documentation*, 40(3):175-205.

Griffiths, A., Luckhurst, C., and Willett, P., "Using interdocument similarity information in document retrieval systems. *Journal of the American Society for Information Science*, 37:3-11.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science, 42*, 1(1991), 7-15.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: Wiley.

Hearst, M.A. and Pedersen, J.O. (1996). Re-examining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In *Proceedings of the 19th Annual ACM SIGIR Conference*, pp. 76-84. Zurich, Switzerland.

Hideo Fuji, W. Bruce Croft, *A Comparison of Indexing Techniques for Japanese Text Retrieval*,

Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1996) "Newsgroup exploration with WEBSOM method and browsing interface". Report A32, Faculty of Information Technology, Helsinki University of Technology (Rakentajanaukio 2 C, SF-02150 Espoo, Finland).

Jardine, N. and Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications. *Computer Journal*, 11(2):177-184.

Jardine, N. and van Rijsbergen, C.J. (1971). The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7:217-240.

Kandel, A., Schenker, A., Last, M. and H. Bunke, (2003). A Comparison of Two Novel Algorithms for Clustering Web Documents, *Proceedings of the 2nd International Workshop on Web Document Analysis (WDA 2003)*, pp. 71-74, Edinburgh, Scotland.

Kohonen, T. (1997). *Self-Organizing Maps*. 2nd ed., Springer-Verlag, Berlin.

Korenius, T., Laurikkala, J., Jarvelin, K., Juhola, M. (2004). Stemming and Lemmatization Finnish document. *ACM Conference on Information and Knowledge Management (CIKM)*

Krovetz, R., et. al., Viewing morphology as an inference process, *Proc. 16th ACM SIGIR Conference*, Pittsburgh, June 27-July 1, 1993; pp. 191-202.

Lancaster, F. W. (1979). *Information retrieval systems : characteristics, testing and evaluation*, 2nd ed., New York, John Wiley.

Lance, G.N. and Williams, W.T. (1967). A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer Journal*, 9:373-380.

Leuski, A. (2001). Interactive information organization: techniques and evaluation. *Ph.D. Thesis*, University of Massachusetts, Amherst.

Leuski, A. (2001). Evaluating document clustering for interactive information retrieval. *Proceedings of Tenth International Conference on Information and Knowledge Management (CIKM'01)*, pages 41-48, Atlanta, Georgia, USA,

Lin, X., Soergel. D.. & Marchionini. G. (1991. October). A self-organizing semantic map for information retrieval. In Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research und Development in I&rmation Retrieval (pp. 262-269). Chicago, IL.

Lovins J.B., 1968: "Development of a stemming algorithm," *Mechanical Translation and Computational Linguistics* **11**, 22-31.

Macskassy, S.A., Banerjee, A., Davidson, B.D., Hirsh, H. (1998). Human

performance on clustering web pages: a preliminary study. In *Proceedings of The 4th International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pp. 264-268. New York, NY.

Maron, M. E., & Kuhns. J. L. (1960). On relevance, probabilistic indexing and information retrieval. Journal of the ACM, 7, 2 16-243.

Milligan, G. W. and Cper, M.C. (1985). An Examinatin of Procedures for Determining the Number f Clusters in a Data Set. *Psychometrika*. 50: 159-179

Milligan, G.W., Soon, S.C., Sokol, L.M. (1983). The effect of cluster size, dimensionality, and the number of cluster on recovery of true cluster structure. *IEEE Transactions on Patter Recognition and Machine Intelligence*, 5(1):40-47.

Mirkin, B. (1996). *Mathematical Classification and clustering*. Kluwer

Mock, K. (1998). A Comparison of Three Document Clustering Algorithms: TreeCluster, Word Intersection GQF, and Word Intersection Hierarchical Agglomerative Clustering. *Intel Technical Report*

Mojena, R. (1977). Hierarchical Grouping Methods and Stopping Rules: An Evaluation. *The Computer Journal*, 20: 359-363

Murtagh, F. (1985). Multidimensional Clustering Algorithm, *COMPSTAT Lectues 4, Physic-Verlag*, Vienna

Na Tang and Rao Vemuri, V. (2004). Web-based Knowledge Acquisition to Impute Missing Values for Classification, *IEEE/WIC/ACM International Joint Conference on Web Intelligence*, Beijing, China

Pirkola, A. (2001). Morphological typology of languages for information retrieval. *Journal of Documentation, 57*, 3 (2001), 330-348.

Porter, M.F. (1980). An Algorithm for Suffix Stripping. *Program - Automated Library and Information Systems*, 14(3): 130-137.

Qin Ding et al., *Data Mining Survey*, North Dakota State University.

Qin He, (1999). Neural Network and Its Application, *Spring*, University of Illinois at Urbana-Champaign

Reggia, J. A.; & Sutton, G. G., III. (1988). Self-processing networks and their biomedical implications. Processings of the IEEE, 76, 680-692.

Salton, G. and Wong, A. (1978). Generation and search of clustered files. *ACM Transactions on Database Systems*, 3(4):321-346.

Salton, G. and McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24:513-523.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single link cluster method. *Computer Journal*, 16:30-34.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical taxonomy: the principles and practice of numerical classification*. San Francisco: W.H. Freeman.

Sparck Jones, K., (1972). "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, v28, pp 11-21, 1972.

Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques. *KDD Workshop on Text Mining*. D.C., USA

Sudipto, G., Rajeev, R., and Kyuseok, S. (1998). CURE: An efficient clustering algorithm for large databases. In *Proc. of 1998 ACM SIGMOD Int. Conf. on Management of Data*, 1998.

Sudipto, G,, Rajeev, R., and Kyuseok, S. (1999). ROCK: a robust clustering algorithm for categorical attributes. In *Proc. of the 15th Int'l Conf. on Data Eng.*, 1999.

Turtle, H.R. and Croft, W.B. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.***3** (1991), 187-222.

van Rijsbergen, C.J. (1971). An algorithm for information structuring and retrieval. *Computer Journal*, 14:407-412.

van Rijsbergen, C.J. and Sparck Jones, K. (1973). A test for the separation of relevant and non relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251-257.

van Rijsbergen, C.J. and Croft, W.B. (1975). Document clustering: An evaluation of some experiments with the Cranfield 1400 Collection. *Information Processing & Management*, 11:171-182.

van Rijsbergen, C.J. (1979). *Information Retrieval*. London: Butterworths, 2nd Edition.

Voorhees, E.M. (1985a). The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval. Ph.D. Thesis, Technical Report TR 85-705 of the Department of Computing Science, Cornell University.

Voorhees, E.M. (1986). Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval. *Information Processing & Management*, 22(6): 465-476.

Ward, J.H. (1963). Hierarchical grouping to minimize an objective function. *Journal of the American Statistical Association*, 58:236-244.

Weiss, R., Velez, B., Sheldon, M. (1996). HyPursuit: A hierarchical network search

engine that exploits content-link hypertext clustering. In *Proceedings of Hypertext '96*, pp. 180-193. Washington, DC.

Williams, W.T., Clifford, H.T., Lance, G.T. (1971a). Group size dependence: a rationale for choice between numerical classifications. *Computer Journal*, 14:157-162.

Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24(5):577-597.

Wishart, D. (1969). An algorithm for hierarchical classification. *Biometrics*, 25:165-170.

Zamir, O. and Etzioni, O. (1988). Web document clustering: A feasibility demonstration. In *Proceedings of the 21st Annual ACM SIGIR Conference*, pp. 46-54. Melbourne, Australia.