AN ANALYSIS OF FUZZY CLUSTERING ALGORTIHMS FOR SUGGESTION OF SUPERVISOR AND EXAMINER OF THESIS TITLE

AZRINA BINTI SUHAIMI

A project report submitted in partial fulfilment 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

To my beloved parents Suhaimi Mohd Noor & Jariah Salim..
Thanks for your love..

To my siblings, Ajak & Anor..
I love you guys...

To my dear Dzulhelmi..
Thanks for being there..

To my late grandma..
You are always in my mind..

ACKNOWLEDGEMENTS

Alhamdulillah, it is with Allah that I get to complete Project II in time given. Here, I would like to take this opportunity to express my sincere thanks to my supervisor, Associate Professor Dr. Naomie binti Salim for her attention and guidance throughout the project. Without her help, I would be lost and making out nothing.

Special thanks to all MSC Project II students, Kak Adel, Kak Hawa, Kak Siti, and Kak Nisa for all the helps and advises. Not forgetting to all the course mate of MSC for cheering me up when I feel down, especially to Kembar, Fadh, Mas and Hud.

And lastly, thanks to my parents for your love and support. Your encouragement gives me the strength to complete my master studies. Not forgetting to Dzulhelmi for being very supportive.

ABSTRACT

Document clustering has been investigated for use in a number of different areas of information retrieval. In this project, the use of Fuzzy clustering techniques for suggestion of supervisors and examiners of thesis in School of Postgraduate Studies at Faculty of Computer Science and Information Technology are studied. The aim of this project is to assist the administration in assigning supervisors and examiners to each post graduate student for their project. Preprocessing tasks for document clustering that are applied in this project are commonly used in the Information Retrieval field, which are stemming, stopword removal, and indexing. Document is represented using the Vector Space Model. The index terms are then clustered using Fuzzy clustering algorithms based on similarity. The selected algorithms for Fuzzy are Fuzzy C-means and Gustafson Kessel. The clustering results are evaluated in terms of classification accuracy to predict the thesis supervisor(s) or examiner(s). Experiments show that Fuzzy C-means gives better result compared to Gustafson Kessel. However, the performances of both techniques are not at the top level. Hence, these techniques are not suitable for use in suggestion of supervisors and examiners. Nevertheless, to get a better performance, a larger dataset, thorough experiments and detailed evaluation has to be carried out and this will take longer time.

ABSTRAK

Kaedah pengelompokan dokumen telah digunakan secara efektif dan meluas di dalam bidang Capaian Maklumat. Tesis ini membandingkan teknik-teknik yang terdapat dalam pengelompokan dokumen yang mana fokus utama adalah terhadap teknik pengelompokan bagi algoritma *Fuzzy*. Teknik pembandingan ini akan digunakan untuk memilih penyelia dan penilai bagi tesis di Sekolah Pengajian Siswazah, Fakulti Sains Komputer dan Sistem Maklumat. Sistem yang dibangunkan diharap dapat membantu pihak pengurusan dalam menentukan penyelia dan penilai bagi tajuk cadangan pelajar. Pra pemprosesan yang digunakan untuk teknik pengelompokan adalah yang biasa digunakan dalam bidang Capaian Maklumat seperti kaedah mewakilkan perkataan dengan kata dasarnya, pembuangan senarai *stopword* dan pengindeksan. Model ruangan vektor akan digunakan untuk mewakilkan dokumen. Terma-terma indeks akan dikelompokkan menggunakan teknik algoritma *Fuzzy* berdasarkan keserupaan antara dokumen-dokumen yang terlibat. Teknik algoritma *Fuzzy* yang akan digunakan ialah *Fuzzy C-means* dan *Gustafson Kessel*. Eksperimen yang telah dijalankan menunjukkan bahawa pencapaian algoritma *Fuzzy C-means* adalah lebih baik jika dibandingkan dengan algoritma *Gustafson Kessel*. Namun demikian, pencapaian kedua-dua teknik ini tidak berada pada kedudukan yang memuaskan. Maka, teknik-teknik ini tidak sesuai untuk mencadangkan pemilihan penyelia dan penilai. Namun demikian, keputusan pencapaian yang baik boleh dihasilkan sekiranya menggunakan set data yang lebih besar, menjalankan eksperimen yang menyeluruh dan melakukan penilaian yang terperinci terhadap keputusan teknik pengelompokan dan ini akan mengambil masa yang agak lama.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

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. Therefore, document clustering arises as a problem in information retrieval which can be defined as grouping documents into clusters according to their topics or main contents in an unsupervised manner. Document clustering has always been used as a means to improve the performance of retrieval and navigating large data.

Various methods have been developed and applied successfully as a solution in document clustering (Frakes, 1992). A large variety of algorithms have been suggested which can be categorized as hierarchical clustering algorithms, partitional clustering algorithms, spectral clustering algorithms and matrix factorization algorithm (Zheng *et al.*, 2004). These algorithms help in effectively identifying document clusters with different topics.

1.1 Background of Problem

Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clustering was investigated for improving the precision and recall in information retrieval systems (van Rijsbergen, 1989; Kowalsky, 1997) and as efficient way of finding the nearest neighbors of a document (Buckley and Lewit, 1985). More recently, clustering has been proposed for use in browsing a collection of documents (Cutting *et al.*, 1992) or in organizing the results returned by a search engine in response to a user's query (Zamir *et al.*, 1997). Document clustering has also been used to automatically generate hierarchical clusters of documents (Koller and Sahami, 1997). A somewhat different approach (Charu *et al.*, 1999) finds the natural clusters in document taxonomy, and then uses these clusters to produce an effective document classifier for new documents.

Clustering methods can be broadly divided into two types: hierarchic methods and non hierarchic or partitioning methods. The hierarchic methods use an $N \times N$ similarity matrix, containing the pairwise similarities in a dataset of size N objects, to create a nested set of clusters. The non hierarchic methods divide a dataset into a single level partition, with or without overlap between the clusters. The same method can then be applied to the resulting partition to produce a hierarchy of partitions. Both types of methods have been used for document collections. The extensive clustering experiments carried out in the SMART Project (Salton, 1971) used a variety of non hierarchic methods. More recent work has used the hierarchic methods.

Agglomerative hierarchical clustering and K-means are two clustering techniques that are commonly used for document clustering. Agglomerative hierarchical clustering is often portrayed as "better" than K-means, although slower, while K-means is used because of its efficiency (Steinbach *et al.*, 2000). These have been shown to be substantially more effective for retrieval than the non hierarchic methods. However, they are substantially less efficient in operation since the non hierarchic methods do not involve the calculation and processing of the inter document similarity matrix (which has a time requirement of at least order $O(N^2)$ for a collection of N documents).

In recent times, researches try to improve the efficiency and effectiveness of retrieval by applying techniques belonging to Artificial Intelligence such as fuzzy clustering. Topics that characterize a given knowledge domain are somehow associated with each other. Those topics may also be related to topics of other domains. Hence, documents may contain information that is relevant to different domains to some degree. With Fuzzy clustering methods, documents are attributed to several clusters simultaneously and thus, useful relationships between domains may be uncovered, which would otherwise be neglected by hard clustering methods. Fuzzy is used in IR to support the need to develop intelligent information retrieval systems in this information age, when the users are faced with the increasingly difficult task of searching through huge amounts of data for useful information (Kraft *et al.*, 2000). Fuzzy is applied for document classification to find natural clusters in documents. From the Fuzzy clusters, fuzzy logic rules are constructed in an attempt to capture semantic connections between index terms.

The other application of document operations that has been commonly applied in IR is ranking algorithms which were investigated since 25 years ago. This type of retrieval system takes as input a natural language query without Boolean syntax and produces a list of records that "answer" the query, with the records ranked in order of likely relevance. Ranking retrieval systems are particularly appropriate for end users. This type of retrieval systems has also been closely associated with clustering. Early efforts to improve the efficiency of ranking systems for use in large data sets proposed the use of clustering techniques to avoid dealing with ranking the entire collection (Salton, 1971). It was also suggested that clustering could improve the performance retrieval by pre grouping like documents (Jardin and van Rijsbergen, 1971).

From all the studies that have been done by researches in the past, hierarchical clustering techniques has been favored in the area of document clustering because of its efficiency and non hierarchical techniques is commonly used for its shorter computation time. Recently, fuzzy clustering has been applied in solving problems of IR. In this project, a comparison of fuzzy algorithms; Fuzzy C-means and Gustafson Kessel will be performed to discover which of these techniques can prove the effectiveness in producing good cluster in the domain problem. It is essential to compare between the algorithms in different domain problem rather than employing only one method to discover the best result. The study of comparing the various techniques in document clustering will definitely give a benefit towards the research in this area.

In Universiti Teknologi Malaysia, each student who performs project is assigned to one or more supervisors to be supervised throughout the semester. Students are assigned to supervisors after proposing a topic for their project. By the end of the semester, students are again assigned to one or more examiners for project presentation. Supervisors and examiners play an important role for a student to accomplish his or her project and obtain good result. Thus, the most appropriate supervisors and examiners must be rightly assigned to each student in order to assist them for completing the project.

In FSKSM alone, the current approach of assigning supervisors and examiners for each postgraduate student is carried out manually. Every semester, the task of assigning supervisors and examiners is first determined during a meeting with lecturers of the faculty and with the help of few of administration staffs. The problem occurs when sometimes there is confusion to determine the right supervisors and examiners for the right students. There are also some students who take a combination of research areas for their project. For example, a student may propose a topic regarding database field and has some approaches in computer graphics. Student might want co-supervisors in assisting him or her in the research of the combined area. It might take a long period to generate a list of suitable supervisors and examiners according to their expertise and experiences.

Since the current system of allocating supervisor and examiner is managed by human, which means human intervention may sometimes prove to have inaccurate result for creating the list of supervisors and examiners. It may affect the quality of students' administration and also the time spent to choose the right supervisors and examiners. The problem arises when students do not get supervision from the right supervisor and this may affect the thesis performance and can cause problems during project execution.

Therefore in this thesis, clustering is mainly used to group the projects into separate cluster based on their similarity. Document clustering techniques are used to suggest the most appropriate supervisors and examiners for the proposed project. Fuzzy clustering algorithms will be used for document clustering analysis and they are to be compared to find out which is the best techniques to produce good result in suggestion of supervisor and examiner of thesis title. In addition, the implementation of mathematical algorithms makes the system more concrete for imprecise situation.

1.2 Statement of Problem

- Can Fuzzy C-means be used for determining supervisor(s) and examiner(s) of thesis effectively?
- Can Gustafson Kessel algorithm based document clustering give better result than Fuzzy C-means algorithm and vice versa for determining supervisor(s) and examiner(s) of thesis effectively?

1.3 Objectives of the Thesis

The objectives of the thesis have been identified and outlined as below:

- To apply clustering techniques which are Fuzzy C-means and Gustafson Kessel for suggestion of supervisor(s) and examiner(s) of thesis title.
- To make a comparison between Fuzzy C-means and Gustafson Kessel clustering in terms of effectiveness for predicting supervisor(s) and examiner(s) of thesis title.

1.4 Scope

Below defines the scope of the study, which involves several areas:

- Titles and abstracts of 210 theses of master project in FSKSM will be stored and used for IR processes.
- Porter stemming and stopword removal will be executed to get only relevant words.
- Applying Fuzzy C-means and Gustafson Kessel algorithm for comparison and evaluation analysis.

1.5 Project Plan

This project is carried out in two semesters. The first part of the project focuses on understanding the general view of document clustering problem in IR field and the past approaches that have been applied by other researches as well as methodology to be used in this project. Most of the time in the first semester is used to explore and gather relevant information from the text books and published journals. The total understanding in document clustering and artificial intelligence methods is important in order to know the different methods that can be used in solving clustering problems. At the end of first semester, a better understanding of Fuzzy algorithms clustering techniques is achieved before executing them. The report for the first semester includes Introduction, Literature View, Methodology and Initial Findings.

The second part of the project involves implementing Fuzzy C-means and Gustafson Kessel for suggestion of supervisor(s) and examiner(s) for each post graduate student in FSKSM based on their proposed thesis title. The implementation begins with preprocessing includes stemming, stopword removal, and Vector Space Model. Next, Fuzzy algorithms are applied to group the index terms into separate cluster based on their similarity that relate to the research field for deciding the most suitable supervisor and examiner. The evaluation towards Fuzzy prediction is performed to see the accuracy of selected supervisor or examiner. Comparison is carried out based on clustering analysis of Fuzzy C-means and Gustafson Kessel clustering. The second part of the report includes Experimental Design, Experimental Result and Conclusion. Appendix A shows the Gantt chart of the project as guidance throughout the project.

1.6 Thesis Contribution

This project gives better insights in the use of document clustering for determination of supervisor(s) and examiner(s) of thesis title. The study also can suggest which Fuzzy clustering algorithm to use to achieve effectiveness; Fuzzy C-means or Gustafson Kessel.

1.7 Outline of Thesis

The contents of the thesis are arranged by chapter. The contents of each chapter are as follows:

- Chapter 1 gives a general introduction of this thesis, which includes the background problem, problem of statement, objectives, goal, scope, project plan as well as project expected contribution.
- Chapter 2 presents a review of relevant and related literature on automated IR systems. It also gives an introduction on overview of document clustering approaches in IR field such as hierarchical and non hierarchical clustering as well as artificial intelligence approaches such as fuzzy clustering. This chapter also presents how these approaches being applied in automated IR systems. Besides that, preprocessing approaches applied in IR domain are also elaborated.
- Chapter 3 discusses about the experimental design which describes deeply about the methodology framework.
- Chapter 4 presents the result from the clustering and also the accuracy of the prediction of supervisor or examiner.
- Chapter 5 presents the conclusion as well as suggestion for further research. There is also a discussion regarding the results that reflects the performance of the clustering techniques.

REFERENCES

Barnard, J. M. and Downs, G.M., (2002). Clustering Methods and Their Uses in Computational Chemistry. In: Lipkowitz, K.B. and Donald B. Boyd D.B. (Ed).*Reviews in Computational Chemistry*. 18. 1-40.

Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function*. Plenum Press. New York.

Borgelt. C, and Nurnberger. (2003). A. Experiments in Document Clustering Using Cluster Specific Term Weights. Department of Knowledge Processing and Language Engineering, University Of Magdeberg, Germany.

Buckley, C. and Lewit A.F. (1985). Optimizations of Inverted Vector Searches. *SIGIR'85*. 97-110.

Charu, C.A. Gates, S.C. and Yu, P.H. (1999). On the Merits of Building Categorization Systems by Supervised Clutsering. *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 352 – 356.

Chowdhury GG (2001). Introduction to Modern Information Retrieval. *Library Association Publishing London*.

Cui, X. and Potok T.E. Document Clustering Using Particle Swarm Optimization. Applied Software Engineering Research. Oak Ridge National Labarotary.

Cutting, D.R., Karger, D.R., Pederson, J.O. and Tukey, J.W. (1992). A Cluster-Based Approach to Browsing Large Document Collection. *SIGIR '92*. 318-329.

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

Downs G.M. and Barnard J.M. (1995). *Hierarchical and Non-Hierarchical Clustering.*Barnard Chemical Information Ltd.

Dunn, J. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact, Well-Separated Cluster. *Journal of Cybernatics*. 3(3). 32-57.

Frakes W.B. (1992). Information Retrieval: Data Structures and Algorithms. *Prentice Hall*, New Jersey.

Fujii, H. and Croft W.B. (1993). A Comparison of Indexing Techniques for Japanese Text Retriaval. Computer Science Department, University of Massachusetts.

Fung, G. (2001). A Comprehensive Overview of Basic Clustering Algorithms.

Gath, I. and Geva, A.B. (1989). Unsupervised Optimal Fuzzy Clsuetring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 11(7). 773-781.

Gustafson E.E. and Kessel W.C. (1979). Fuzzy Clustering with a Fuzzy Covariance Matrix. *Proc. 18th IEEE Conference on Decision and Control.* 761–766.

Holliday, J.D., Rodgers, S.L and Willet, P. (2004). Clustering Files of Chemical Structures Using the Fuzzy k-Means Clustering Method. *J. Chem. Inf. Compt. Sci*. 44.894-902.

Hull, D. (1996). Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70-84.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.

Jain, A.K, Murty, M.N. and Flynn, P.J.(1999). *Data Clustering: Review*. ACM Computing Surveys.

Jardine, N. and van Rijsbergen, C.J. (1971). The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7.217-240.

Kaymak, U. and Setnes, M. (2000). Extended Fuzzy Clustering Algorithm. *ERIM Report Series Research in Management*. 1-23.

Koller, D. and Sahami, M. (1997). Hierarchically Classifying Documents Using Very Few Words. *Proceedings of the 14th International Conference on Machine Learning (ML), Nashville, Tennessee*. 170-178.

Korenius, T., Laurikkala, J., Jarvelin, K. and Juhola, M. (2004). Stemming and Lemmatization in the Clustering of Finnish Text Documents. *CIKM'04*.

Kowalsky ,G. (1997). Information Retrieval Systems – Theory and Implementation *Kluwer Academic Publisher.*

Kraft, D.H., Chen, J. and Mikulcic, A. (2000). Combining Fuzzy Custering and Fuzzy Inferencing in Information Retrieval. *IEEE.*

Krovetz, R. (1993).Viewing morphology as an inference process. *Proc. 16th ACM SIGIR Conference*. 191-202.

Kurita, T. (1991). An efficient agglomerative clustering algorithm using a heap. *Pattern Recogn.* 24(3). 205–209.

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

Mao, J. and Jain, A. K. (1992). Texture Classification and Segmentation Using Multiresolution Simultaneous Autoregressive Models. *Pattern Recogn.* 173–188.

Martin, E. (2003). Pap-Smear Classification. Technical University of Denmark.

Orengo, V.M. and Huye, C. (2001). A Stemming Algorithm for the Portuguese Language. School of Computing Science, Middlesex University, The Burroughs, London.

Ross, G. J. S. (1968). Classification Techniques for Large Sets of Data. Academic Press, Inc., New York, NY.

Shah, J.Z. and Salim, N. (2004) FCM and GK Clustering of Chemical Datasets using Topological Indices. Faculty of Computer Science & Information System, UTM, Malaysia.

Salton, G.(1971). The SMART Retrieval System – Experiments in Automatic Document Retrieval. *Englewood Cliffs. NJ: Prentice-Hall.*

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

Salton, G. (1991). Developments in automatic text retrieval. *Science 253.* 974–980.

Steinbach, M., Karypis, G. and Kumar, V. A Comparison of Document Clustering Techniques. Department of Computer Science and Engineering, University of Minnesota.

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

van Rijsbergen, C.J (1989). Information Retrieval. Buttersworth, London, second edition.

Zamir, O., Etzioni, O., Madani, O. and Karp, R.M. (1997). Fast and Intuitive Clustering of Web Documets. *KDD'97*. 287-290.

Zheng, Y.N., Dong, H.J. and Chew, L.T. (2004). Document Clustering Based on Cluster Validation. *CIKM'04*.

Zhengxin, C, (2001). Data Mining and Uncertain Reasoning. *A Wiley-Interscience Publication*.