DOCUMENT CLUSTERING BASED ON INVERSE DOCUMENT FREQUENCY MEASURE

WAN FARIDAH HANUM WAN YAACOB

UNIVERSITI UTARA MALAYSIA
2005

DOCUMENT CLUSTERING BASED ON INVERSE DOCUMENT FREQUENCY MEASURE

A thesis submitted to the Faculty of Information Technology in partial Fulfillment of the requirements for the degree Master of Science (Intelligent System), Universiti Utara Malaysia

by

Wan Faridah Hanum Wan Yaacob

JABATAN HAL EHWAL AKADEMIK
(Department of Academic Affairs)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certify that)

WAN FARIDAH HANUM WAN YAACOB

calon untuk Ijazah
(candidate for the degree of) **MSc. (Intelligent System)**

telah mengemukakan kertas projek yang bertajuk
(has presented his/ her project paper of the following title)

DOCUMENT CLUSTERING BASED ON INVERSE DOCUMENT FREQUENCY MEASURE

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.
(that the project paper acceptable in form and content, and that a satisfactory knowledge of the field is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor) : **CIK NOORAINI YUSOFF**

Tandatangan
(Signature) :

Tarikh
(Date) : 7/4/2025

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for the postgraduate degree from Universiti Utara Malaysia, I agree that University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in their absence by the Dean of the Graduate School. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Request for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Faculty of Information Technology
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

ABSTRACT (ENGLISH)

Automatic classification techniques are capable of providing the necessary information organization by arranging the retrieved data into groups of documents with common subjects. Recently, document clustering has been put forth as an alternative method of organizing the results of retrieval. It been proposed for use in navigating and browsing document collections, and discovers hidden similarity and key concepts. It also summarize a large amount of document using key or common attributes of cluster and can be used to categorize document databases. This paper describes several narrative clustering techniques such as Porter algorithm, Gusfield algorithm, similarity based on document hierarchy and *Inverse Document Frequency* (IDF), which intersect the documents in a cluster to determine the set of words (or phrases) shared by all the documents in the cluster. This study proposes document clustering based on IDF, where it is assumes that importance of a keyword in calculating similarity measures is inversely proportional to the total number of documents that contain it. IDF is easy to understand, has a geometric interpretation, term weighing shown to help clustering, allow partial matching and returns ranked documents. An important finding in this study, where 30 cases of documents tested with the IDF algorithm, and the results are divided into three category; *correct cluster*, *incorrect cluster*, and *unknown cluster*.

ABSTRACT (BAHASA MELAYU)

Teknik pengklasifikasian secara automatik berkemampuan untuk menyediakan keperluan informasi untuk sesebuah organisasi dengan menyusun atur capaian data ke dalam kumpulan dokumen yang mempunyai subjek yang sama. Kebelakangan ini, pengkelompokan dokumen telah dijadikan sebagai satu proses alternatif untuk menguruskan sesebuah keputusan dalam capaian data. Ia telah dicadangkan untuk menavigasikan koleksi-koleksi dokumen, dan mencari persamaan kata kunci. Ia juga membuat rumusan ke atas bilangan dokumen yang besar dengan menggunakan teknik capaian kekunci dan juga boleh digunakan untuk mengkategorikan pangkalan data dokumen tersebut. Kajian ini menghuraikan beberapa teknik pengkelompokan seperti *Porter Algorithm*, *Gusfield Algorithm*, *similarity based on document hierarchy*, dan juga teknik *Inverse Document Frequency (IDF)*, di mana kesemuanya merupakan penentu untuk menentukan kumpulan perkataan ataupun frasa yang dikongsi bersama oleh semua dokumen yang ada di dalam satu gugusan. Tambahan pula, kajian ini juga mencadangkan pengkelompokan dokumen berasaskan *IDF*, di mana teknik ini mengambil kira nilai kata kunci di dalam pengiraan keseimbangan secara songsang kepada jumlah keseluruhan dokumen yang memiliki kata kunci tersebut. *IDF* adalah satu algoritma yang mudah untuk difahami, mempunyai penterjemahan secara geometrik, membenarkan proses penyesuaian dokumen dan memberikan nilai akhir kedudukan dokumen-dokumen tersebut. Kepentingan yang diperoleh daripada kajian ini berdasarkan 30 kes dokumen yang dikaji, hasilnya boleh dibahagikan kepada tiga kategori iaitu *Correct Cluster*, *Incorrect Cluster*, dan juga kategori *Unknown Cluster*.

ACKNOWLEDGEMENTS

Alhamdulillah, praise to Allah S.W.T, whom granted me the strength, ability and full of guidance to complete this project.

My family has always been there, and writing this project has been made much easier for that. My beloved parents, Rossitah Hashim and Wan Yaacob Mohamed, have always supported me in my study. I love both of you more than I can express. My sweet sister, Wan Farrahiyah and my naughty brother, Wan Mohd Kamarul Hizzaq have done what all good sister and brother do: mercilessly tease at every opportunity, while simultaneously making their love abundantly clear.

Thousands of appreciation and thanks to my supervisor, Miss Nooraini Bt Yusoff for all of her support, patient and also her continuous assistant and guidance during the research, preparations and until the completions of this project. Without her benevolence, I am sure that I cannot complete this project properly.

As ever, my heartfelt thanks go out to my course mates and my wonderful friends in the Intelligent Systems community, who have made academia rewarding and enjoyable. Deserving of a special mention here are Norlia Yusof, Nor Rafidah Mohd, Mohd Zaki Salikon, Wan Hazimah Wan Ismail, Nooraini Omar, Suraya Abd Rahman and all of my friends here: their kindness and hospitality has been astonishing. Far away friends of mine have helped me keep my feet firmly on the ground throughout the writing of this project. Special thanks here to Syazzan Fared Idrus, Idriana and all of their family members.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

This chapter presents the overview of document management systems and document clustering. Problem statement, objectives, scope, and significance of study are also discussed in this chapter.

1.1 Overview Of Study

The main activity of most personal computer users is about creating, managing, deleting and retrieving electronic documents. Existing file management systems is performed using hierarchical structures whereby a document is stored and accessed at a specific location. For example, we would create a file ''Lecture1.ppt'' in the subdirectory ''Lectures'' which is it a subdirectory of the ''Object-Oriented Design'' directory. In fact we are also associating some semantics to the created file.

The contents of the thesis is for internal user only

REFERENCES

Backer, A., Busbach, U. (1996). DocMan: A Document Management system for cooperation support. Proceedings of the Hawaii International Conference on System Sciences, pp.82-91.

Balasubramanian, V., Bashian, A., and Porcher, D. (1997). A large-scale of hyper media application using document management and web technologies. Communications of the ACM.

Baldonado, M. Q., and Winograd, T. (1997). SenseMaker: An Information-Exploration Interface Sup-porting the Contextual Evolution of a User's Interests. In Proceedings of the Conference on Human Factors in Computing Systems, ACM Press, pp. 11–18.

Beeferman, D., and Berger, A. (2000). Agglomerative clustering of a search engine query log. KDD'2000, pp. 407-416.

Castro, L. N., and Zuben, F. J. (2001). AiNet: an Artificial Immune Network for Data Analysis. Idea Group Publishing.

Celentano, A., Pozzi, S., and Toppeta, D. (1992). A multiple presentation document management system. Proceedings of the 10th Annual Conference on Systems Documentation, pp. 63-71.

Church, D. W. (1990). Inverse Document Frequency (IDF): A Measure of Deviations from Poisson. AT&T Bell Laboratories Murray Hill, NJ, USA.

Cutting, D. R., Pedersen, J. O., Karger, D., and Tukey, J. W. (1992). Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 318-329.

Dourish, P., Edwards, W. K., LaMarca, A., and Salisbury, M. (1999). Presto: an experimental architecture for fluid interactive document space. ACM Transactions on Computer-Human Interaction 6(2), pp. 133-161.

Dourish, P., Edwards, W. K., LaMarca, A., Lamping, J., Peterson, K., Salisbury, M., Terry, D. B., and Thornton, J. (2000). Extending document management systems with user-specific active properties. ACM Transaction on Information System 18 (2), pp.140-170.

Dunham, M. H. (2002). Data Mining: Introductory and advanced topics. Prentice Hall.

Dunlop, M. D. (2000). Development and Evaluation of Clustering Techniques for Finding People. Proceedings of the Third International Conference on Practical Aspects of Knowledge Management, Basel, Switzerland. Available:http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-34

Eguchi, K. (1999). Adaptive cluster-based browsing using incrementally expanded queries and its effects. In ACM SIGIR 99.

Griffiths, A., Luckhurst, H., and Willet, P. (1997). Using Interdocument Similarity Information in Document Retrieval Systems. Readings in Information Retrieval, pp. 365-373.

Rajaraman, K. And Pan, H. (1998). Document clustering using 3-tuples. Kent Ridge Digital Labs, Singapore.

Kang, S. (2001). Keyword-based Document Clustering. School of Computer Science, Kookmin University & AITrc, Korea.

Kim, H. R. (2003). Web Personalization. Department of Computer Sciences, Florida Institute of Technology.

Kohonen, T. (1995). Self-Organizing Maps. Springer.

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

Kurtus, R. (2004). Overview of Coldfusion. School for Champions, Kurtus Technologies, Milwaukee, Wisconsin.

Lewis, D.D., and Croft, W. B. (1990). Term clustering of syntactic phrases, ACM-SIGIR, pp. 385-404.

Lin, K., and Kondadadi, R. (2001). A Word-based Soft Clustering Algorithm for Documents. Department of Mathematical Sciences, the University of Memphis, USA.

Lotus. (1993). System Administration Manual, Lotus Notes Release 3.

Meziane, F. and Rezgui, Y. (2004). A Document management methodology based on similarity content. School of Computing, Science and Engineering, Salford University, UK, pp. 15-34.

Purao, S. (2002). Design Research in the Technology of Information Systems: Truth or Dare. GSU Department of CIS Working Paper. Atlanta.

Rasmussen, E. (1992). Clustering algorithms. In W.B. Frakes and R. Baeza-Yates, editors, Information Retrieval, Prentice Hall, pp. 419-442.

Rasmussen, M., and Karypis, G. (2004). gLUTO- An interactive clustering, visualization, and analysis system. University of Minnesota, Department of Computer Science and Engineering.

Rivera, G., Norrie, M. C., and Steiner, A. (2000). IDEOMS: An Integrated Document Environment based on OMS Object-Oriented Database System. Institute for Information System, Swiss Federal Institute of Technology (ETH), Switzerland.

Robertson, S. (2003). Understanding Inverse Document Frequency: On theoretical arguments for IDF. Journal of Documentation 60, no. 5, pp. 503–520.

Sahami, M., Yusufali, S., and Bal-donado, M. (1998). SONIA: A Service for Organizing Networked Information Autonomously. In Proceedings of the Third ACM International Conference on Digital Libraries.

Simon, H. (1996). The Sciences of the Artificial, Third Edition. Cambridge, MA, MIT Press.

Souza, J. M. (2004). Intelligent Document Management. Buildings.Com. Available:http://www.buildingsintegration.com/Articles/detail.asp?articleID=2004

Spark-Jones, P., and Willet, K. (1997). Readings in Information Retrieval. Morgan Kaufman, pp. 305-312.

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

Timo, H., Samuel, K., Krista, L., and Teuvo, K. (1997). WEBSOM-self-organizing maps of document collections. In Proceeding of WSOM '97 Workshop on Self-Organizing Maps, pp. 310-315.

Tang, N., and Vemuri, V. R. (2004). An Artificial Immune System Approach to Document Clustering. Computer Science Department, University of California.

Vaithyanathan, S., and Dom, B. (1999). Model Selection in Unsupervised Learning with Applications to Document Clustering. In Proceedings International Conference on Machine Learning.

Weiss, S. M., White, B. F., and Apte, C. V. (2000). Lightweight Document Clustering. IEEE Intelligent System

Wetherill, M., Rezgui, Y., Lima, C., and Zurli, A. (2002). Knowledge Management for the Construction Industry: The E-Cognos Project. Itcon Vol. 7, pp. 183-196.

Wu, W., and Xiong, H. (2002). Query Clustering in the Web Context. Kluwer Academic Publishers, pp. 1-30.

Xu, X, and Leiss, E. L. (2002). Personal Information Retrieval Visualization (PIRV): Clustering and Visualization of Web Document Search Results. Department of Computer Science University of Houston.

Zamir, O., Etzioni, O., Madani, O., and Karp, R. (1997). Fast and Intuitive Clustering of Web Documents. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining.

Zamir, O. and Etzioni, O. (1998). Web Document Clustering: A Feasibility Demonstration. In ACM SIGIR 98, pp. 46-54.