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

© Wan Faridah Hanum Wan Yaacob, April 2005. All rights reserved
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) :  [Signature]

Tarikh
(Date) :  4/4/2027
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 has been proposed for use in navigating and browsing document collections, and discovers hidden similarity and key concepts. It also summarizes 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.
Teknik pengklasifikasian secara automatik berkemampuan untuk menyediakan keperluan informasi untuk sebuah 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 menguruskannya sebuah 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 kesemuaanya 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 Farrahiah 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

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Permission to use</td>
<td>i</td>
</tr>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Abstract (Bahasa Melayu)</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgement</td>
<td>iv</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Figures</td>
<td>viii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
</tbody>
</table>

## CHAPTER 1: INTRODUCTION

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Overview of study</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Document Clustering</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Problem Statement</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Objective of Study</td>
<td>7</td>
</tr>
<tr>
<td>1.4 Scope of Study</td>
<td>7</td>
</tr>
<tr>
<td>1.5 Significance of Study</td>
<td>9</td>
</tr>
<tr>
<td>1.6 Organization of Study</td>
<td>9</td>
</tr>
</tbody>
</table>

## CHAPTER 2: LITERATURE REVIEW

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Document Clustering in Document Management System</td>
<td>11</td>
</tr>
<tr>
<td>2.2 Document Clustering</td>
<td>23</td>
</tr>
<tr>
<td>2.3 Similarity Measures</td>
<td>27</td>
</tr>
</tbody>
</table>
CHAPTER 6: CONCLUSION AND RECOMMENDATION

6.1 Conclusion ......................................................... 77
6.2 Recommendation .................................................. 79

REFERENCES

References ................................................................. 81

APPENDIXES

Appendix A: Snapshot of Document Clustering Prototype
Appendix B: Tables Of Clustering Database
LIST OF FIGURES

Figure 1.1 Architecture for Document Management ........................................ 2
Figure 1.2 Example of Keyword-based Clustering ........................................... 5
Figure 2.1 DocMan’s distribution and replication service ..................................... 13
Figure 2.2 DocMan main window ................................................................. 13
Figure 2.3 Illustration of Scatter/Gather ......................................................... 15
Figure 2.4 The Infobus architecture ............................................................... 15
Figure 2.5 Processing stages in the SONIA system ........................................... 16
Figure 2.6 SenseMaker View Controller ....................................................... 17
Figure 2.7 The e-Cognos Global Architecture ............................................... 18
Figure 2.8 The framework of aiNet ............................................................... 19
Figure 2.9 aiNet algorithm ............................................................................. 19
Figure 2.10 Overview of gCLUTO’s work-flow ............................................... 20
Figure 2.11 Several screen-shots of the clustering dialog .................................. 21
Figure 2.12 Illustration of Clustering ............................................................. 25
Figure 2.13 General clustering process .......................................................... 27
Figure 2.14 Classification of clustering algorithm .......................................... 33
Figure 2.15 Algorithm of HAC ..................................................................... 37
Figure 3.1 Outputs of Design Research ........................................................ 42
| Figure 3.2 | The General Methodology of Design Research ........................................... 43 |
| Figure 3.3 | General architecture of document clustering using IDF .................................... 44 |
| Figure 3.4 | Use case diagram for document clustering ...................................................... 45 |
| Figure 3.5 | Design of document clustering ...................................................................... 46 |
| Figure 3.6 | The interpretations of IDF measurements ...................................................... 49 |
| Figure 3.7 | Illustration for new inserted document ......................................................... 50 |
| Figure 4.1 | Coldfusion Studio 4.5 ................................................................................. 55 |
| Figure 4.2 | Implementation of Coldfusion in World Wide Web .............................................. 57 |
| Figure 4.3 | Main page of document clustering .................................................................. 58 |
| Figure 4.4 | Fill-in form of new document ...................................................................... 59 |
| Figure 4.5 | Notification message for incomplete form ..................................................... 60 |
| Figure 4.6 | Result’s page of document clustering ............................................................. 61 |
| Figure 4.7 | Clustering table ........................................................................................... 62 |
| Figure 4.8 | Statement to compare the keyword and existence description ......................... 63 |
| Figure 4.9 | Statement to match the inserted keyword and description ................................ 64 |
| Figure 4.10 | Statement to calculate IDF .......................................................................... 64 |
| Figure 4.11 | Formula to calculate IDF ............................................................................. 64 |
| Figure 4.12 | Temp table for temporary IDF calculation ....................................................... 65 |
| Figure 4.13 | Computes minimum value of IDF .................................................................. 65 |
| Figure 4.14 | Fruitfully inserted document into Correct Cluster .......................................... 66 |
| Figure 4.15 | Unknown table for unknown cluster ................................................................ 66 |
| Figure 5.1 | Summarization of results with threshold (0) .................................................... 71 |
Figure 5.2  Summarization of results with threshold (1.15) .................. 74
Figure 5.3  Results between two thresholds ................................. 75
Figure 5.4  Comparison between two different thresholds .............. 75
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1.1</td>
<td>Modules of Document Clustering</td>
<td>8</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Experiments of document clustering with threshold (0)</td>
<td>70</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>Summarization of results (threshold 0)</td>
<td>71</td>
</tr>
<tr>
<td>Table 5.3</td>
<td>Experiments of document clustering with threshold (1.15)</td>
<td>72</td>
</tr>
<tr>
<td>Table 5.4</td>
<td>Summarization of results (threshold 1.15)</td>
<td>73</td>
</tr>
</tbody>
</table>
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


81


