GRAPH BASED TEXT REPRESENTATION FOR
DOCUMENT CLUSTERING

ASMA KHAZAAL ABDULSAHIB

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)
SCHOOL OF COMPUTING
COLLEGE OF ARTS AND SCIENCES
UNIVERSITI UTARA MALAYSIA
2015
PERMISSION TO USE

In presenting this dissertation in fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this dissertation in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this dissertation 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 dissertation.

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

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok
ABSTRACT

Advances in digital technology and the World Wide Web has led to the increase of digital documents that are used for various purposes such as publishing and digital library. This phenomenon raises awareness for the requirement of effective techniques that can help during the search and retrieval of text. One of the most needed tasks is clustering, which categorizes documents automatically into meaningful groups. Clustering is an important task in data mining and machine learning. The accuracy of clustering depends tightly on the selection of the text representation method. Traditional methods of text representation model documents as bags of words using term-frequency index document frequency (TFIDF). This method ignores the relationship and meanings of words in the document. As a result the sparsity and semantic problem that is prevalent in textual document are not resolved. In this study, the problem of sparsity and semantic is reduced by proposing a graph based text representation method, namely dependency graph with the aim of improving the accuracy of document clustering. The dependency graph representation scheme is created through an accumulation of syntactic and semantic analysis. A sample of 20 news group, dataset was used in this study. The text documents undergo pre-processing and syntactic parsing in order to identify the sentence structure. Then the semantic of words are modeled using dependency graph. The produced dependency graph is then used in the process of cluster analysis. K-means clustering technique was used in this study. The dependency graph based clustering result were compared with the popular text representation method, i.e. TFIDF and Ontology based text representation. The result shows that the dependency graph outperforms both TFIDF and Ontology based text representation. The findings proved that the proposed text representation method leads to more accurate document clustering results.

KEYWORDS

Text Representation scheme, Dependency Graph, Document Clustering
ACKNOWLEDGEMENT

In the name of Allah the most gracious the most merciful First and foremost, Praise to Allah, Lord of the Worlds and prayers and peace are upon the master of messengers the Prophet Mohammed. Our leader in this life until the closing.

I would like to convey my deepest gratitude to my supervisor, Dr. SITI SAKIRA KAMARUDDIN for all continuous guidance and advices given to me in writing up of this dissertation.

Next I would like to thank University Utara Malaysia (UUM) staff. Especially, School of Computing staff for their cooperation with me.

Especial thanks to my husband Amjad Majed and my kids (Noor, Mohammed and Haidr) for their constant support and encouraged me and sacrifice during the production of this dissertation. I would like to thank my father and mother for Permanent their prayer for me to finish this work. Lastly, I want to thank my country, Iraq for the material and moral support for getting the Master Certificate.
DEDICATIONS

To the Most Merciful…

In the name of Allah, Most Gracious, Most Merciful.

TABLE OF CONTENTS

PERMISSION TO USE i

ABSTRAK ii

ABSTRACT iii

ACKNOWLEDGEMENT iv

DEDICATION v

TABLE OF CONTENTS vi

LIST OF FIGURES x

LIST OF TABLES xii

LIST OF APPENDICES xiii

CHAPTER ONE: INTRODUCTION 1

1.1 DOCUMENT CLUSTERING 1

1.2 TEXT REPRESENTATION SCHEME 2

1.3 DOCUMENT CLUSTERING TECHNIQUES 5

1.4 PROBLEM STATEMENT 8

1.5 RESEARCH QUESTATION 9

1.6 RESEARCH OBJECTIVES 9

1.7 SCOPE OF THE STUDY 10

1.8 SIGNIFICANCE OF THE STUDY 10
CHAPTER TWO: LITERATURE REVIEW

2.1 INTRODUCTION

2.2 TEXT REPRESENTATION METHOD
   2.2.1 TF-IDF
      A. TF-IDF ALGORITHM PRINCIPLE
      B. TF-IDF WEIGHTING METHODS
   2.2.2 N-GRAM
   2.2.3 ONTOLOGY
   2.2.4 GRAPHBASEDTEXT REPRESENTATION
      A. CONCEPTUAL GRAPHS
      B. FORMALCONCEPT ANALYSIS
      C. CONCEPT FRAMEGRAPHS
      D. DEPENDENCYGRAPH

2.3 CLUSTERING METHODS
   2.3.1 PARTITIONAL CLUSTERING
      A. K-MEANS
      B. SETTING UP
   2.3.2 HIERARCHICAL CLUSTERING
   2.3.3 DENSITY BASED CLUSTERING
   2.3.4 GRID-BASED CLUSTERING

2.4 EVALUATION METHODS

2.5 SUMMARY
# Chapter Three: Research Methodology

3.1 Theoretical Study 54

3.2 Research Design 54

## 3.2.1 Document Collection 55

## 3.2.2 Document Pre-Processing 56

- **A. Split Documents** 57
- **B. Stemming** 57
- **C. Tokenization** 58
- **D. Part of Speech Tagging** 59

## 3.2.3 Text Representation Scheme 60

- **A. Parsing** 60
- **B. Construct Dependency Graph** 61

## 3.2.4 Apply Clustering Algorithm 62

## 3.2.5 Evaluation and Result Analysis 62

# Chapter Four: Dependency Graph Based Text Representation

4.1 Introduction 64

4.2 Document Pre-Processing 65

4.3 Parsing 70

4.4 Generating Dependency Graphs 71

4.5 Ontology Based Approach for Semantic Analysis 77

4.6 Summary 84
LIST OF FIGURES

Figure 1.1: Clustering Algorithm 6

Figure 2.1: K- means Algorithm 40

Figure 2.2, A: Determining the Number of Clusters 40

Figure 2.2, B: Randomly Guessing the K-means Center Location 40

Figure 2.2, C: Determining Which Center Each Data Point is Closest 41

Figure 2.2, D: Finding the Centroid of The Points Owned by a Center 41

Figure 2.2, E: Repeat Until Terminated 42

Figure 2.3: The Steps of K-means Algorithm 43

Figure 3.1: Research Design 55

Figure 3.2: The 20 Newsgroup 56

Figure 3.3: Dependency Graph Generated for Beijing is a big city. The city is very beautiful 62

Figure 4.1: Steps Building The Dependency Graph 64

Figure 4.2: Single Document Called alt.atheism 65

Figure 4.3: Split Single Document in Python 66

Figure 4.4: The Example Document After Splitting into Sentences 66

Figure 4.5: Porter Stemming Algorithm 67

Figure 4.6: Stemming Process 67

Figure 4.7: Tokenization Code in Python 68

Figure 4.8: Example Document After Tokenization Process 68
LIST OF TABLES

Table 2.1: Summary of Text Representation Scheme 33
Table 2.2: Summary of Clustering Methods 50
Table 4.1: Size Documents Before and After Constructing DG 81
Table 5.1: The Value of Confusion Matrix 90
Table 5.2: Average of Experimental Results of The DG 93
Table 5.3: TP, FP, FN and TN of Dependency Graph 94
Table 5.4: Precision, Recall, F-measure and Accuracy of TFIDF Representation 96
Table 5.5: Precision, Recall, F-measure and Accuracy of DG Representation 97
Table 5.6: Precision, Recall, F-measure and Accuracy of Ontology Representation 98
List of Appendices

Appendix A: Size Documents before and after Construct Dependency Graph 120
CHAPTER ONE

INTRODUCTION

1.1 DOCUMENT CLUSTERING

Document clustering is considered a vital technology in the era of internet. It’s an essential technique in mining underlying structures in text document data sets. Furthermore, this is a very interesting research topic that has influenced a number of researchers and practitioners from a number of fields, including data mining, machine learning, and information retrieval due to its fundamental role in many of real-world applications (Andrews & Fox, 2007). Text clustering means finding the groups that are related to each other. These groups are collected together in an unstructured formal document. In fact, clustering becomes very famous for its ability to offer an exceptional way of digesting in addition to generalize a good quantity of information. The extracting appropriate feature is considered the basis of clustering. Clustering text documents into category groups is a necessary step in the mining of abundance text data on the Web, indexed and retrieval or incorporate information systems and extract proper feature (concept) of a problem area. Text documents are often represented as high-dimensional, sparse vectors and complex semantics (Dhillon, et al., 2001& Jing, et al., 2005).

In existing clustering methods, a document is often represented as “bag of words” (in BOW model), N-grams (in suffix tree document model), or TF-IDF without considering the natural language relationships between the words (Wang et al., 2011).
The contents of the thesis is for internal user only
REFERENCES


Chakrabarti, S. (2003). Mining the Web: Discovering knowledge from hypertext data: Morgan Kaufmann.


Huang, J., Sun, H., Song, Q., Deng, H., & Han, J. (2013). Revealing Density-Based Clustering Structure from the Core-Connected Tree of a Network. Knowledge and Data Engineering, IEEE Transactions on, 25(8), 1876-1889.


Mitchell, B. S., & Mancoridis, S. Clustering module dependency graphs of software systems using the bunch tool.


Stumme, G. (2002). Formal concept analysis on its way from mathematics to computer science Conceptual Structures: Integration and Interfaces (pp. 2-19): Springer.


