UNIVERSITI TEKNOLOGI MARA

APPLICATION OF CLUSTERING IN MANAGING UNSTRUCTURED TEXTUAL DATA IN RELATIONAL DATABASE

WAEL MOHAMED SHAHER YAFOOZ

Thesis submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Computer and Mathematical Sciences

October 2014

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicates or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that i have been supplied with the academic rules and regulations for postgraduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Student Name : WAEL MOHAMED SHAHER YAFOOZ

Student Id : 2010791005

Program : PhD of Computer Science (CS990)

Faculty : Computer and Mathematical Sciences

Thesis Title : Application of Clustering in Managing Unstructured Textual Data in Relational Database

Signature of student :

Date : October 2014

ABSTRACT

Huge reliance on computer usage in everyday life, leads to a continuous increase of large data applications in textual forms. The data are reposited to a secondary storage for future usage. Therefore, a *relational database* (RDB) is most commonly used as a backbone in most application software for organising such data into structured form. The RDB has robust and powerful structures for managing, organising, and retrieving the data. However, the database structure can still contain large amounts of unstructured textual data. Dealing with unstructured textual data leads to three basic issues; users encounter difficulties to find useful information, inaccurate information retrieval and insufficient performance of query processing. Attempts have been made to resolve all of these issues by using several methods such as: full text searching, text indexing, a database schema management, database data model, and query-based techniques. However, the front-end approach, in the form of software applications, are still needed to organise the unstructured textual information in the RDB. This study proposes a *Textual Virtual Schema Model* (TVSM) as the back-end approach to reorganising textual data inside relational databases, while performing automatic semantic linking and clustering assignments. Upon storing any new unstructured textual data into a database, all words are extracted to uncover the underlying meaning of such data. Their name entities and top most frequent terms are selected for the factors used in a cluster assignment. The model is tested and evaluated by embedding it in a component-based package of a relational databases internal structure. Three experiments have been conducted on textual Reuters corpus, Classic and WAP dataset. The clustering results have been validated using the *F-measure, Entropy* and *Purity* methods of measurement and compared with two common methods, which are information extraction and textual document clustering, for example, *K-means, Frequent Item-Set, Hierarchical Clustering Algorithms* and *Oracle Text*. The results show that there are linkages between structured textual data and unstructured information, quality improvement in textual document clustering with accurate clusters and high performance of query processing. Thus, the proposed technique can increase retrieval performance and produce high accuracy textual data clusters. This model envisages a beneficial and useful approach for various domains that involve big textual data such as document clustering, topic detecting and tracking, information integration, personal data management and information retrieval.

- This research work published in eight international proceeding indexed by ISI and Scopus and two book chapters indexed by ISI and Scopus and one international journal.
- This research work has patent pending under serial number PI2013002636 from MYiPO Malaysia.

TABLE OF CONTENTS

CHAPTER ONE
INTRODUCTION

This chapter will incorporate information that delivers the problem statements, the research questions and objectives, plus the significance of this research work. In addition to this, the introductory chapter will also present the scope of the research and the organisation of the thesis.

1.1 BACKGROUND OF STUDY

The computer was invented to assist humans in various ways. One of the most important purposes of the computer is data storage, management and organization to preserve data for future usage. The processed data can be transformed into information and knowledge. Data are often stored in databases or document files for proper structure, fast retrieval, storage space and greater security as compared to the hard copy (Li, Chung, and Holt, 2008). The significant reliance on computer applications tremendously increases the usage of large textual data (Li, Feng, and Zhou, 2011). Such data can be found on web pages, document files, personal documents, and discussion forums.

The massive volume of textual data in databases, web pages, and document files are usually regarded as unorganised. Such data will be meaningless and unclear for users, because of the difficulty in realizing the relationship between its contents. Textual data can be classified into three forms: structured, semi-structured, or unstructured. Structured data is the best form of information because it facilitates the acquisition and comprehension of knowledge. There are many forms of unstructured or semi-structured data, such as news portals, articles, and discussion forums. Such data is often stored in a *Relational Database Management System* (RDBMS), which are also known as data repositories. Due to RDBMS has a robust structure that manages, organises and retrieves data. This data structure represents data in the form of database tables consisting of attributes (columns) and records (rows). This allows the processing of data manipulation (i.e., insert, update, delete, and retrieve) to be performed efficiently and also allows efficient data retrieval and provides accurate and desireable results. These processes are performed using meta-data such as attribute names. Nonetheless, the RDBMS can contain a huge amount of unstructured textual data, such as textual documents, and these can remain in their original format within the data structure, because of the storing process, which saves the textual data as it is. Figure 1.1 shows the common way of storing textual unstructured data in RDBMS.