UNIVERSITI TEKNOLOGI MARA

THE DOCTORAL RESEARCH ABSTRACTS

Volume: 7, Issue 7 May 2015

SEVENTH ISSUE

INSTITUTE of GRADUATE STUDIES

Leading You To Greater Heights, Degree by Degree

IGS Biannual Publication

Name :
Wael Mohamed Shaher Yafooz

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

Supervisor :
Associate Prof. Dr. Siti Zaleha Zainal Abidin (MS)
Dr. Nasiroh Omar (CS)

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 database's 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, high performance of query processing and time improvement in document clustering with accurate clusters. 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.