UNIVERSITI TEKNOLOGI MARA

EFFECTIVENESS OF SIMPLE TERMINOLOGICAL ONTOLOGY TO SUPPORT DOCUMENT RETRIEVAL IN A SPECIALISED DOMAIN

SEYED ABOLFAZLE MOOSAVIFAR

Thesis submitted in fulfillment
of the requirement for the degree of
Master of Science

Faculty of Computer and Mathematical Sciences

September 2014

AUTHOR'S DECLARATION

I declare that the work in this dissertation 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 indicated or acknowledged as referenced work. This topic 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 Regulation for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student : Seyed Abolfazle Moosavifar

Student I.D. No. : 2010230266

Programme : Master of Science

Faculty : Faculty of Computer and Mathematical Sciences

Thesis Title : Effectiveness of Simple Terminological Ontology to Support Document Retrieval in a Specialised Domain

Signature of Student :.

Date : September 2014

ABSTRACT

The research investigated the proposition that a simple terminological ontology supported by general purpose lexical resources and aided by information retrieval and natural language processing techniques can effectively annotate and retrieve documents in a specialised knowledge domain. This is addressing the evidence from a recent survey, which reported that low satisfaction in the retrieval of documents in a personal collection. A common, but robust approach in this area is keyword-based retrieval. The weakness of keyword-based retrieval is its inability to 'understand' the meaning of the keywords (semantic). Ontology approach is introduced as a way to support semantic retrieval. However, there is a problem with the construction of the ontology by laymen, especially ontologies for specialised domain areas. Therefore, the use of simple terminological ontology (constructed based on intuitive understanding of the domain) is proposed in this research. The research objectives are structured to introduce new algorithms for ontology-based automatic annotation, retrieval and ranking of documents and to check on the reliability of WordNet to provide lexical support for the (simple terminological) ontology-based document retrieval. To achieve these objectives, the Boolean IR model was extended by incorporating four coefficients to adjust the term weights, namely to deal with the word significance and word coherence in multi-word terms, to consider the matching type (exact or synonym) and to factor the category weight when calculating the term weights. To find the retrieval effectiveness, the results of ontology-based retrieval was evaluated against the conventional retrieval, and validated against expert retrieval. The results of the ontology-based automatic annotation were evaluated against expert annotation. In addition, the reliability of using WordNet to provide lexical support was tested during the process of the annotation and retrieval. The research found synonyms from WordNet selected with the correct senses can help to improve the (simple terminological) ontology-based annotation and retrieval of documents in a specialised domain area. The research also found that (simple terminological) ontology-based retrieval that is support by selected synonyms from WordNet can recall all documents that are retrieved using keyword-based retrieval with reasonable precision. The evaluations of the retrieval by get help from expert domain also emphasized this result. The research result also indicated there are few common tags between the automatic and expert annotation. There were issues with the expert annotations; nonetheless, if we regard the expert annotation is paramount, then we suggest semi-automatic annotation of the documents in order to improve the result of ontology-based retrieval. Future researchers can use our research ideas (e.g. annotation and retrieval algorithms; assignment of weights to ontology terms) to make further progress in the field of semantic information retrieval. System designers can base our research findings (e.g. type of lexical support) to decide on methods for improving the retrieval in personal collection.

TABLE OF CONTENTS