THESIS INFORMATION

- The Doctoral Dissertation: **ONTOLOGY BASED INFORMATION RETRIEVAL**
- Field: **Computer Science**
- Field code: **62.48.01.01**
- PhD student: **Ngo Minh Vuong**
- Supervisor: **Assoc. Prof. Dr. Cao Hoang Tru**
- Institution: **Ho Chi Minh City University of Technology – VNUHCM**

1. Abstract of the dissertation

Current text document retrieval systems are facing to many challenges in discovering and representing the semantics of queries and documents. Document retrieval based on lexical matching of keywords has many drawbacks because it only considers the surface forms of words appearing in a text rather than the meaning of the words. Meanwhile, the content of a text is mostly determined by concepts such as named entities and WordNet words. On the other hand, the meaning of a query could express more clearly user intention if it is expanded with suitable latent concepts. The objective of this thesis is to exploit ontologies of named entities, WordNet words and entity relationship facts to improve the performance of document retrieval in terms of the precision and recall measures.

In a text, concepts are expressed by their surface forms like entity names or word labels. Those concepts contain hidden ontological features under their surface forms, such as aliases/synonyms, super-classes/hypernyms, sub-classes/hyponyms and identifiers/senses. Besides, each query also implies those entities that are related to entities explicitly appearing in the query.

This thesis consists of three main parts. First, the thesis explores ontological features of named entities, different combinations of them and keywords, and evaluates their impact to document retrieval performance, in which name-class pairs and identifies of named entities have not been exploited in previous works. Second, the thesis proposes usage of form-sense pairs of WordNet words in addition to other basic ontological features that have been used previously. Third, the thesis exploits an ontology of facts to expand a query by latent entities that have explicit relations with other entities in the query.

The proposed models are implemented by extending the basic vector space model and experimented on benchmark datasets and standard performance measures. Experiment results show that the proposed models give better retrieval performance than the models of related works and the traditional keyword-based document retrieval model. Especially, this thesis uses statistical significance tests to confirm the actual improvement in performance of the proposed models.

2. Contributions of the dissertation

The thesis proposed the document retrieval models which exploited ontological features of named entities, WordNet words and facts, relatively completely and comprehensively, to improve the performance of document retrieval, including:

1. The model exploited ontological features of named entities and combined them with keywords.
2. The model exploited ontological features of WordNet words combined with keywords.
3. The model expanded a query with named entities by spreading on explicit relations in the query.
4. The model combined the methods in the above proposed models.

The performances of the proposed models were tested by experiments and a statistical significance testing.

3. Suggestions for future work

From the researches and results of this thesis, we propose a number of directions for further researches as following:

1. Exploiting latent entities bridging relations with the entities in the query through explicit relations in it.
2. Combining the R+CSA algorithm of the thesis with a pseudo feedback algorithm to the entities added into the query would be actually appropriate with the contents of the query more.
3. Exploiting ontological features of named entities and WordNet words on the information retrieval models being different from the vector space model.
4. Exploiting a topic modeling to represent topics by ontological concepts such as named entities and WordNet words, and use these models to represent documents and queries.

Supervisor	PhD Student
Assoc. Prof. Dr. Cao Hoang Tru	**Ngo Minh Vuong**