# THESIS INFOMATION

Title:

**A SEMANTIC - BASED DESIGN METHOD FOR DOMAIN-SPECIFIC DOCUMENT MANAGEMENT AND RETRIEVAL SYSTEMS**

Major: Computer Science
Major Code: 62.48.01.01
PhD Student: Huynh Thi Thanh Thuong
Advisor: Assoc. Prof. Do Van Nhon
University: University of Information Technology, Vietnam National University - Ho Chi Minh City

## 1. ABSTRACT

This dissertation proposes a novel method for the problem of Ad-hoc Document Retrieval in a specific domain. This work follows a keypharse graph-based approach for text representation, takes advantage of a fine-grained, well-controlled domain ontology as the underlying semantic resource, and utilizes advanced search techniques, thereby significantly improving retrieval performance. In addition, the study also focuses on a comprehensive framework for designing domain-specific document management and retrieval systems, called "Semantic Document Base Systems". These such systems are capable of managing semantic information associated with document content and facilitating semantic processing in search. The implementation of several specific applications illustrates the effectiveness and feasibility of the proposed ideas. Besides, the benefits of graph-based document representation models and related techniques are also demonstrated through the Document Similarity Measurement task. The utilization of large knowledge bases (e.g. DBpedia, Wikipedia) makes available fine-grained information about concepts, entities, and their semantic relations, thus resulting in a knowledge-rich interpretation.

All research results of the dissertation are published in international conference proceedings and journals indexed in Web of Science, Scopus, EI Compendex, Inspec, DBPL, ACM Digital Library, etc.

## 2. THE MAIN CONTRIBUTIONS OF THE THESIS

The main contributions are listed as follows:

1) A novel semantic document retrieval method based on domain ontology and Keyphrase Graphs for document representation, including: The Classed Keyphrase based Ontology model (CK-ONTO for short) captures domain knowledge and semantics, is employed for comprehending queries and documents, as well as for assessing semantic similarity; Keyphrase Graph-based document models, along with a method for generating structured representations of texts; A graph matching technique for measuring semantic relevance in search.
2) A framework for building a kind of domain-specific document management and retrieval system, called Semantic Document Base Systems (SDBS).

3) Three practical application systems are implemented: The learning resource repository management system in Computer Science domain; The Information Technology Job-posting retrieval system; The Vietnamese online news aggregating system in Labor and Employment domain alongside Public Investment and Foreign Investment domain, served Binh Duong Department of Information and Communications, Viet Nam.
4) A keyphrase graph-based method for document similarity measurement
5) Three prototype knowledge bases (KB) in CK-ONTO model: Computer Science KB, Labor & Employment KB, and IT-Jobs. The designing process of such systems was presented in depth along side with experimental setup and dataset as a benchmark to evaluate search efficiency.

## 3. FUTURE WORKS

The dissertation's methods should be studied and improved further by:

- Take advantage of heuristics to reduce computational complexity even further and optimize the performance of search algorithms.
- Develop concept-oriented document representation and knowledge integration methods for various relevant domains.
- Design a mechanism to automatically update the ontology as well as other components affected by the change (e.g., keyphrase graphs of documents); Enhance the inference ability of ontological knowledge bases.
- Expand and develop more appropriate and effective methods for Vietnamese document retrieval.
- Develop automatic indexing methods for document warehouses, explore the utilization of distributed databases, graph databases, and specialized computational models in processing extremely big graph data. These efforts aim to optimize the information/document retrieval process in the context of big data.
- Conduct research on a comprehensive design solution for integrated systems that combine knowledge querying with document retrieval.

| ADVISOR | PHD STUDENT |
| --- | --- |
| DO VAN NHON | HUYNH THI THANH THUONG |