INFORMATION OF DOCTORAL DISSERTATION

PhD student name: **TẠ DUY CÔNG CHIẾN**

Dissertation title: **BUILDING A TOPIC-ORIENTED DOCUMENT-BASED INFORMATION EXTRACTION MODEL FOR A SPECIFIC DOMAIN (COMPUTING DOMAIN)**

Major: **COMPUTER SCIENCE**

Major code: **62.48.01.01**

Scientific Advisor: **Professor, Doctor. PHAN THỊ TƯƠI**

School: **HCMC University of Technology, Vietnam**

DISSERTATION SUMMARY:

Nowadays, besides Information Retrieval and Question Answering, Information Extraction has become an emerging trend in the modern area of digital information processing. Especially, in some certain domains such as Medicine, Biology and Education, Information Extraction has made significant contributions for the improvement of human living condition. However, there are also a lot of challenges arising when the information extraction is exploited in a specific domain, e.g. processing data from different resources, the quality and correctness of the extracted information and system performance when dealing with large-scale datasets. Therefore, there is much research, both in international and national scales, has been conducted on the issue of building an efficient information extraction system on a specific domain. It prompts the general objective of this dissertation, which is to build a topic-oriented document-based information extraction system for a specific domain (chosen as Computing within the dissertation scope).

To develop such a proposed system, the dissertation proposes a novel methodology which combines ontology engineering, natural language processing and statistics algorithms. In order to fulfil its objective, the dissertation has made the three key problems as follows:

* **Problem 1–Constructing and enriching Computing Domain Ontology (CDO)**: Identifying and extracting the objects and the semantic relations based on the different sources in order to build and enrich CDO.
* **Problem 2–Ontology-based Identifying the topic of the queries:** the analysis and identification the topic of the queries is an one of the problems that the dissertation must solve to build the Information extraction system based on CDO for answering the queries.
* **Problem 3-Information Extraction for answering the queries**: For answering the queries the dissertation will extract information based on CDO after identifying the topic of the queries in the problem 2.

The experiments of the dissertation were conducted from the following materials: (i) a dataset of text documents collected from the ACM Digital Library; (ii) Wikipedia; (iii) WordNet and (iv) a collection of the user's queries inputted directly into the system. The

experiment results show that the algorithms and the proposed models are feasible and introduce significant improvement compared to similar existing works.

NEW CONTRIBUTIONS / RESULTS OF THE DISSERTATION:

The dissertation proposed a theoretical model and methodology of the topic-oriented document-based information extraction system for a specific domain (Computing domain) and archived the new key results as follows:

* **The first-Building and enriching Computing Domain Ontology**. The dissertation proposes a topic-based theoretical model of the information extraction system on textual documents.
* **The second-Identifying the topic the queries**. The dissertation proposes the structure, classes and instances of a specific ontology, known as Computting Domain Ontology (CDO), which renders more semantic relations than other typical existing ontological models.
* **The third-Extraction the semantic relations.** The dissertation proposes and improves some algorithms of information extraction for building and enriching the ontology.
* **Lastly-Information extraction system for answering queries**. The dissertation develops an ontology-based information extraction system for answering user's queries.

Above results were also presented in published papers of the author.

APPLICATIONS OF DISSERTATION'S RESULT IN PRACTICE:

The achieved results of the dissertation lay the foundation of further research of the author. They also support to develop applications in research projects of the author and other external information retrieval system in practice.

FURTHER RESEARCH OF THE DISSERTATION:

There are several matters for further research of the author as follows.

* **Problem 1:** Dissertation will enrich CDO from the dataset of text documents which do not know their topic and automatically upgrade instances of CDO from the papers or the web sites related to Computing domain.
* **Problem 2:** Dissertation will upgrade the new semantic relations for applying the different applications.
* **Problem 3:** Dissertation will optimize algorithms' implementation in extraction information from the different sources and the CDO for answering the queries.

Scientific Advisor	PhD student
Prof. Dr. PHAN THỊ TƯƠI	TẠ DUY CÔNG CHIẾN