Ontology-Based Semantic Search for National Database of Natural Resources and Environment

Ngoc-Vu Nguyen[1,2(✉)], Hong-Son Bui[3], and Quang-Thuy Ha[1]

[1] University of Engineering and Technology (UET),
Vietnam National University, Hanoi (VNU),
144, Xuan Thuy, Cau Giay, Hanoi, Vietnam
thuyhq@vnu.edu.vn
[2] Department of Information Technology,
Ministry of Natural Resources and Environment,
28 Pham van Dong, Cau Giay, Hanoi, Vietnam
nnvu@monre.gov.vn
[3] Ho Chi Minh City Department of Natural Resources and Environment,
63 Ly Tu Trong, District 1, Ho Chi Minh City, Vietnam
son.ciren@gmail.com

Abstract. Semantic search helps the user queries to be understandable for electric agents searching. In this way, ontology plays the main role to define the semantic and the relations between user queries. The national database of natural resources and environment is a very large database system. Therefore, building a search engine software for the system with high accuracy and fast speed is very important for sharing information in the field of natural resources and environment. However, the existing search engine software in the national database needs to be improved to better meet user's needs. We proposed the architecture of ontology-based semantic search for the national database of natural resources and environment. Based on the proposed architecture, we have built semantic search software (VnNRESS) to demonstrate better results than the existing search software (NRESearch).

Keywords: Ontology · Semantic search · Search engine · Ontology-based semantic search

1 Introduction

In recent decades, the knowledge economy has been chosen as the development strategy in almost all countries. Studies have shown that ICT is the infrastructure to implement most of the activities of the knowledge economy, is an effective means of supporting and facilitating for creating and using knowledge.

The national database of natural resources and environment is a very large database system. Therefore, building a search engine software for the system with high accuracy and fast speed is very important for sharing information in the field of natural resources and environment.

T. Q. Duong et al. (Eds.): INISCOM 2019, LNICST 293, pp. 155–164, 2019.
https://doi.org/10.1007/978-3-030-30149-1_13

Keyword-based searching software was built in the project of building the national database of natural resources and environment. However, the software needed to improve some functions in order to better meet the needs of users: (i) keyword auto-suggestions and autocomplete input box, (ii) pre-processing in the case of Vietnamese without accents, (iii) word tokenizer and extracting a group of keywords from the user's query, (iv) processing with synonyms and related semantic words, (v) removing stop words and (vi) sorting search results. Thus, to reach improvement requirements with the above functions require the new searching software to analyze and "understand" the semantics of user query from which search results are the most relevant to user's demand.

Currently, semantic search and ontology-based semantic search are interested topics of the scientific community. There are works related ontology-based semantic search. Gupta et al. [2] proposed a novel architecture of ontology-based semantic search engine. Aghajani et al. [3] introduced a semantic search engine based on ontology with the name is Semoogle. Moawad et al. [4] proposed an ontology-based architecture for the semantic search engine for the Arabic language. Khan et al. [5] introduced an ontology-based semantic search engine in Holy Quran and an architecture of ontology-driven semantic search proposed by Bonino et al. [6].

In Vietnam, related to the field of research on semantic search, there are two typical research projects. The first is the knowledge-based search for water-related information system for the Mekong Delta, Vietnam by Tran Thai Binh (Universität Bonn University) and the second is classifying questions towards Vietnamese semantic search in the field of health by Nguyen Minh Tuan (University of Technology - VNU).

Therefore, building a semantic search software for the national database of natural resources and environment can solve the above-mentioned problems. In the current solutions to build semantic search software, ontology-based semantic search is a suitable and feasible solution.

This paper presents an ontology-based semantic search architecture for the national database of natural resources and environment. This paper has two main contributions: (i) propose an architecture of ontology-based semantic search for the national database of natural resources and environment, (ii) develop a semantic search software based the proposed architecture to evaluate the effectiveness of the proposed architecture.

The rest of this paper is organized as follows. In the next section, we briefly outline fundamental concepts of ontology building, semantic web and semantic search. Section 3 describes the proposed architecture of ontology-based semantic search for the national database of natural resources and environment. In Sect. 4, we present experimental results and comparing search results and performance of the semantic search software that was developed in our experiments with the existing search engine software. In the last section, we present conclusions and future work.

2 Ontology and Semantic Search

2.1 The Semantic Web

Semantic Web came up in 1998 by Tim Berners-Lee which was published on the road map to the Semantic Web on the home page of the WWW Consortium. According to Tim Berners-Lee, "the Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation". The Semantic Web is considered a future generation of the Web, in which:

- Data could be processed by machine and automated agents.
- The meaning of information and services on the Web is defined for the Web to comprehend and satisfy the requests of people and machines to use its content.
- Explicit semantic information on the Web pages which can be used by intelligent agents to solve complex problems of information collection and Query-reply.

The final objective of the Semantic Web is to be able to keep the accountability of Web content and capability to semantically analyze. It needs a group of structures to model the knowledge and a link between the knowledge and contents. In this way, it depends on two basic components: ontology and semantic annotations. Ontology is used to interpret the textual content of a resource regardless of its format. There have been many fundamental approximations in the Semantic Web in which it is supposed that resources have been semantically annotated (Fig. 1).

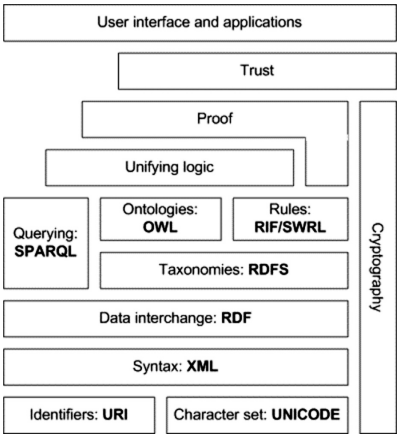

Fig. 1. The semantic web architecture

2.2 Ontology

Ontology term was initially used preferably by AI researchers and now it is one of the bases of the Semantic Web. It is impossible to envision the Semantic Web with no ontology because Semantic Web is the prime research project concerning ontology. The term Ontology is lent from philosophy. There are different definitions for the concept of ontology applied to information systems, each emphasizes a specific aspect. Gruber (1993) defines an ontology as a formal specification of conceptualization or, in other words, declarative representation of knowledge relevant to a particular domain. According to Uschold and Gruninger (1996) ontology is as a shared understanding of some domain of interest. Ontology provides "well-defined meaning" to the information enclosed in the Web also the benefit that different parties over the internet now have "shared" definitions about certain key concepts. The most important characteristic of ontology for the present research effort is their role as a structured form of knowledge representation. Ontology is used for the reason of interoperability among systems based on different schemas and comprehensively describing knowledge about a domain in a structured and sharable way, ideally in a readable format that is processable by a computer.

Noy and McGuinness [7] introduced the process of ontology building with four main steps. Firstly, define classes of the ontology. Secondly, reorganize classes in the class hierarchy of the ontology. Thirdly, define the properties of classes and finally, find instances of classes and value of properties of classes (Fig. 2).

In the TNMT.2015.08.06 project [1], we built an ontology in the field of natural resources and environment and this ontology will be used for the proposed architecture in Sect. 3.

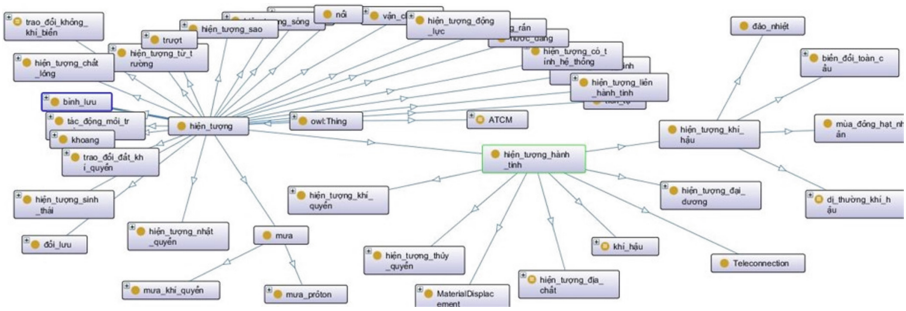

Fig. 2. Example of Vietnam's weather ontology

2.3 Semantic Search

Basically, a semantic search engine has a similar structure to a regular search engine that also includes two main components: the front-end component and the back-end component.

The front-end component has two main functions:

- Query interface: allows users to enter queries.
- Display searching results.

The back-end component plays an important role in the semantic search engine and it has three main functions:

- Analyzing user's queries.
- Searching in databases or document stores for user's queries.
- Document sets, search data/semantic network.

The difference in the structure of semantic search engines compared to a search engine is usually in the internal architecture part, specifically in two components: analyzing questions and searching data sets. Figure 3 shows the basic architecture of the semantic search system:

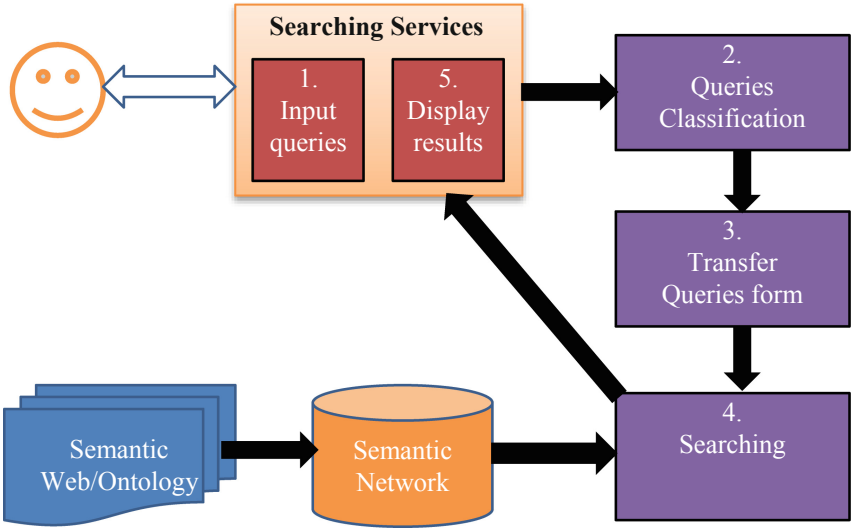

Fig. 3. The basic architecture of a semantic search system

3 The Proposed Architecture of Ontology-Based Semantic Search

Figure 4 shows the proposed architecture of ontology-based semantic search of the national database of natural resources and environment:

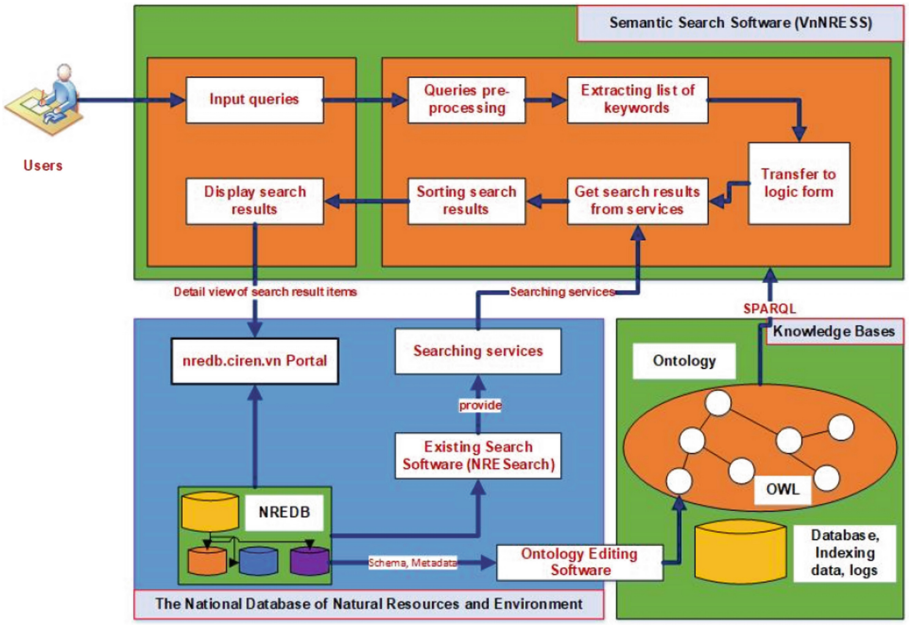

Fig. 4. The proposed architecture of ontology-based semantic search of the national database of natural resources and environment

According to the proposed architecture, semantic search software includes the following main functions:

- Input queries: Allows users to enter a query. The system uses a dictionary and ontology to suggest search keywords.
- Queries prepossessing: analysis and processing the query the user has entered includes the following steps: Firstly, removing spaces and special characters in the query except for the search operator characters. Secondly, removing stop words which are words that have no meaning in the query (using a Vietnamese dictionary of stop words). Thirdly, converting Vietnamese without accents to accented Vietnamese. Finally, separating the query into sentences by search operators.
- Extracting list of keywords: using Vietnamese natural language processing techniques with VietWordNet (a WordNet for the Vietnamese language) and ontology as additional knowledge to extract a list of keywords that match the meaning that the user desires.
- Transfer queries to logic form: building the query in the searchable query format in DBMS.
- Get searching results from the service: using the search service provided by the NRESearch software with input is the processed query to retrieve the search results.
- Sorting search results.
- Display search results.

In functions from 1st to 5th, the ontology of natural resources and environment domain is used as additional knowledge, to improve efficiency and accuracy.

4 Experimental Results

4.1 Datasets

Ontology
The ontology used in this research was built in the TNMT.2015.08.06 project for two domains of natural resources and environment field include geodesy and cartography domain and meteorological hydrological domain. Total classes in this ontology about 111.150 classes.

Databases
Data used for searching include text, tabled data, geospatial data were stored in database systems of the national database of natural resources and environment. The database systems were built in the project of building a national database of natural resources and environment by Ministry of Natural Resources and Environment.

4.2 Evaluation Method

We have created 300 different queries belonging to different types of queries such as Vietnamese-accented queries, Vietnamese without accents queries, and queries have Vietnamese language errors (extra space, contain special characters, contain stop words, etc.) and queries contain a combination of search operators (AND, OR, XOR, NOT).

We used these queries as input to the VnNRESS software and the NRESearch search software. Then for each query, we evaluate the search results of two search softwares based on a list of evaluation criteria such as:

- The semantic relevance of the search results for the user's question.
- The number of search results.
- The sort order of search results.
- Total processing time to produce search results.

4.3 Experiment Results and Analysis

- With most of the queries, the VnNRESS software gave better results than NRESearch software on the following issues:
  - Keyword suggestions: suggestion of search keywords in NRESearch software is not correct. VnNRESS software not only suggests exact search keywords with the Vietnamese-accented query but also exactly with the Vietnamese without accents query. In addition, VnNRESS software also supports a suggestion of search keywords with synonyms and near meanings in the conceptual hierarchy of the ontology.

  - Handling Vietnamese without accents: Since using the ontology has established the equivalent meaning of the Vietnamese word with accents and the Vietnamese word without accents corresponding, so the VnNRESS software handled very well with Vietnamese without accents queries, while the NRESearch software did not yet have this function.
  - Tokenizer and extracting a list of search keywords: it is very simple with the NRESearch software because of separation of the query by space into a list of single words. Meanwhile, the VnNRESS software performs separation by meaningful words using the dictionary and the ontology (e.g. "Ho Chi Minh city" can be understood as three keywords "Ho Chi Minh City", "City" and "Ho Chi Minh". With the separation of keywords into the above meaning, the search results will be more appropriate for the searching needs of users.
  - Elimination of stop words: using a dictionary of Vietnamese stop words inherited from previous studies, eliminating stop words of the VnNRESS software will help to get more suitable search results.
  - Sorting search results: The order of sorting the search results of results display function of the VnNRESS software is more appropriate than the NRESearch software.
  - Handling search operators: The VnNRESS software supports most of the basic search operators and gives more accuracy than the NRESearch software.
  - Accuracy of the search results: Through the above tests, the VnNRESS software has made a list of results that are more suitable for search requests in the user's search content.
- In addition, the VnNRESS software has been added some following functions:
  - Support suggests other search content, more suitable than the content that the user has entered.
  - Having the function of identifying the name in the content of the user's search by using information about place names and information about administrative units: provinces, districts and communes.
  - Supporting the standardization of key words through the relationship between words (for example, geographic names often come together; provincial/district/commune information often goes together,…).
  - Support the identification of information about place names, give correct answers about names, geographical locations, locations on the map of places that appear in the user's search sentence.
  - Supporting search by synonyms, related words (upper level concepts, lower levels in the ontology).
- Performance: Although it takes a lot of time for analyzing users' questions, extracting a list of appropriate keywords but search execution time of the VnNRESS software is acceptable and guaranteed (Figs. 5 and 6).

Fig. 5. The search results of the NRESearch software with input query is "hiển trạng and hà nổi"

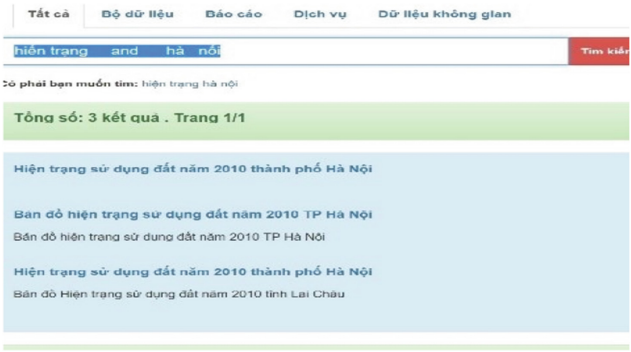

Fig. 6. The search results of the VnNRESS software with input query is "hiển trạng and hà nổi"

5 Conclusions and Future Work

Through the above experimental and assessment results, the VnNRESS software has been proven effective and suitable, more accurate than the NRESearch software. At the same time, the empirical results also confirm the proposed architecture of ontology-based semantic search for the national database of natural resources and environment is appropriate and feasible.

However, in order to achieve greater efficiency, our future works need focus on two main issues:

- Firstly, we need to build a completed ontology for all domains of natural resources and environment field using semi-automatic and automated techniques based on web resources to reduce the effort and cost of ontology building.
- Secondly, we will have to use the complete ontology for all phases of the data search process. Especially in the search indexing phase because in the national

database system of natural resources and environment, indexing is using Apache Lucene software and it separate sentences of document records into a list of single words without taking care of semantics of all words in those sentences.

References

1. Nguyen, N.-V.: Building ontology in the field of natural resources and environment for data integration and semantic search. J. Nat. Res. Environ. 30–32 (2017). (in Vietnamese)
2. Gupta, P.N., Singh, P., Singh, P.P., Singh, P.K., Sinha, D.: A novel architecture of ontology-based semantic search engine. Int. J. Sci. Technol. **1**(12), 650–654 (2012)
3. Aghajani, N.: Semoogle - an ontology based search engine, Master thesis, Norwegian University of Science and Technology (2012)
4. Moawad, I.F., Abdeen, M., Aref, M.M., Shams, A.: Ontology-based architecture for an Arabic semantic search engine. In: The Tenth Conference on Language Engineering (ESOLEC 2010) (2010)
5. Khan, H.U., Saqlain, S.M., Shoaib, M., Sher, M.: Ontology-based semantic search in Holy Quran. Int. J. Future Comput. Commun. **2**(6), 570 (2013)
6. Bonino, D., Corno, F., Farietti, L., Bosca, A.: Ontology driven semantic search. In: World Scientific and Engineering Academy and Society Conferences (2004)
7. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University, Stanford, CA
8. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**(5), 34–43 (2001)