Expert Systems with Applications 42 (2015) 5163–5176

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

ONLI: An ontology-based system for querying DBpedia using natural language paradigm

Mario Andrés Paredes-Valverde [a,*], Miguel Ángel Rodríguez-García [a], Antonio Ruiz-Martínez [a], Rafael Valencia-García [a], Giner Alor-Hernández [b]

[a] Departamento de Informática y Sistemas, Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
[b] Division of Research and Postgraduate Studies, Instituto Tecnológico de Orizaba, Mexico

ARTICLE INFO

Article history:
Available online 26 February 2015

Keywords:
Natural language processing
DBpedia
Ontology
Semantic Web

ABSTRACT

The Semantic Web has emerged as an extension of the current Web, in which Web content has well-defined meaning through the addition of logic-based metadata. However, current mechanisms for information retrieval from semantic knowledge bases restrict their use to only experienced users. To address this gap, the natural language processing (NLP) is deemed to be very intuitive from a use point of view, due to it hides the formality of a knowledge base as well as the executable query language. This paper presents a novel ontology-based information retrieval system for DBpedia called ONLI (Ontology-based Natural Language Interface). ONLI proposes the use of an ontology model in order to represent both the syntactic question's structure and the question's context. This model allows inferring the answer type expected by the user through an established question's classification. These features allow reducing the search space thus increasing the probability of providing the correct answer. From this perspective, ONLI was evaluated in terms of their ability to find the correct answer into DBpedia's content, achieving promising results and proving to be very useful to non-experienced users.

1. Introduction

The amount of information available on the Web, intranets or databases has been steadily increased. Most of the information is designed to be read by humans, and not to be meaningfully manipulated by machines. The Semantic Web has emerged as an extension of current Web where the information has well-defined meaning and understandable by not only humans, but also computers achieving these machines can automate, integrate and reuse high-quality information across several applications (Berners-Lee, Hendler, & Lassila, 2001). Some important components of Semantic Web stack are Resource Description Framework (RDF) and ontologies. RDF is a standard for encoding metadata and other knowledge on the Semantic Web. On the other hand, ontologies are the paramount technology of the Semantic Web. In this work, an ontology is viewed as "a formal and explicit specification of a shared conceptualization".

Nowadays, the information stored in ontology-based knowledge bases has significantly grown and are being applied to different domains such as recommendation (Amini, Ibrahim, Othman, & Nematbakhsh, 2015; Carrer-Neto, Hernández-Alcaraz, Valencia-García, & García-Sánchez, 2012; Colombo-Mendoza, Valencia-García, Rodríguez-González, Alor-Hernández, & Samper-Zapater, 2015), information retrieval (Rodríguez-García, Valencia-García, García-Sánchez, & Samper-Zapater, 2014b), innovation management (Rodríguez-García, Valencia-García, Alcaraz-Mármol, & Carralero, 2014a), sentiment analysis (Peñalver-Martinez et al., 2014) , medicine (Ivanović & Budimac, 2014; Ruiz-Martínez, Valencia-García, Martínez-Béjar, & Hoffmann, 2012), human perception (Prieto-González, Stantchev, & Colomo-Palacios, 2014), and Customer Relationship Management (CRM) (García-Sánchez, Valencia-García, Martínez-Béjar, & Fernández-Breis, 2009) to mention but a few.

One of the most well-known ontology-based knowledge base is DBpedia (Lehmann et al., 2014), which represent a crowd-sourced community effort to structured information from Wikipedia and make this information available on the Web. DBpedia provides a description of 4.0 million of things such as people, places, organizations, species, and diseases. Several research efforts to provide support mechanisms for information retrieval from semantic knowledge bases have been reported (Pan, Thomas, & Sleeman,

* Corresponding author. Tel.: +34 868 88 85 22.
E-mail addresses: marioandres.paredes@um.es (M.A. Paredes-Valverde), miguelangel.rodriguez@um.es (M.Ángel Rodríguez-García), arm@um.es (A. Ruiz-Martínez), valencia@um.es (R. Valencia-García), galor@itorizaba.edu.mx (G. Alor-Hernández).

http://dx.doi.org/10.1016/j.eswa.2015.02.034
0957-4174/

2006) (Hartig, 2013). Also, there are formal query languages for RDF data such as RQL (Karvounarakis, Alexaki, Christophides, Plexousakis, & Scholl, 2002), SeRQL (Sesame RDF Query Language) (Broekstra & Kampman, 2006), and SPARQL (Simple Protocol and RDF Query Language). However, the use of these query languages demands a high level of knowledge and expertise about technologies such as RDF and query language expressions. The need to make the contents of the Semantic Web accessible to all kind of users entails the development of new support mechanisms for information retrieval mainly focused to end users. To address the gap between knowledge bases systems and end users, the natural language paradigm is generally deemed to be very intuitive from a use point of view (Cimiano, Haase, Heizmann, Mantel, & Studer, 2008). A Natural Language Interface (NLI) is a system that allows user to access information stored in some repository by formulating requests in natural language (Kaufmann & Bernstein, 2010). NLI enables hiding the formality of a knowledge base as well as the executable query language from end-users, allowing them to use all communicative power of language they already possess instead of being forced into an unnatural and limited mode of communication. Natural language has been widely used as support of database query, question answering, command and control, and other types of information systems (Zhou, 2007).

In this work, a semantic-based approach is presented for retrieving information from DBpedia through question expressed on natural languages. The approach considers the textual elements that constitute a question which are processed by means of natural languages techniques such as POS-tagging, lemmatizing, chunking, and search of synonyms. The obtained elements are queried in knowledge base in order to establish the question context. This information is organized in an ontological model from which the possible ambiguities are managed; also the answer type expected is inferred thus reducing the search space. Then different queries against the knowledge base are performed with the aim of obtaining possible answers. Once possible answers are obtained, these ones are organized by relevance and they are shown back to the user. This approach aims to provide a system which does not require users to learn specialized vocabularies, or to know the structure of the knowledge base.

The remainder of the paper is structured as follows. Some related works about natural languages interfaces development are presented in Section 2. The architecture design of the proposed approach, components and interrelationships are described in Section 3. In Section 4, a use case scenario is shown in order to describe the functionality of the proposed approach. In Section 5, the effectiveness of this proposal is evaluated. Finally, conclusions and future work are presented.

2. Related works

A great number of researchers have focused their works on NLI development applied to multiple contexts and data sources. In this section, some of the most prominent research works on this subject from 2013 to 2015 were reviewed.

2.1. Natural Language Interfaces to Databases (NLIDB)

Nowadays, there are research efforts focused on the development of natural language interfaces applied to relational databases. An example of this approach is NaLIR (Li & Jagadish, 2014), which represents the linguistic understanding of the input sentence by a parse tree, then the system maps proximity of the patterns in the parse tree to proximity of corresponding database concepts, the result is shown back to the user in order to ensure that the sentence was correctly understood. The understanding is then translated into an SQL query, which is executed by the DBMS against the relational database. Other works are focused on developing domain-independent NLIDB, such as AskMe (Llopis & Ferrández, 2013a) which uses a template-based approach for the dynamic generation of the *lexer*, syntactic and semantic parser. First, AskMe searches for the ontology for a particular domain in the ontology repository, if this ontology does not exist; then the system builds the ontology that captures the overall characteristics of the database domain. Next, it builds the parsers which help understanding user's queries and translate them into SQL-based queries to be executed against the database. Other approach is presented in (Shah, Pareek, Patel, & Panchal, 2013), this work presents an interesting approach called NLKIDB (Natural Language and Keyword Base Interface to Database). This approach tokenizes input query by a lexical analyzer and lexicons are parsed to the syntax analyzer. If the input query is syntactically valid, lexicons are analyzed by the semantic analyzer through a domain ontology and then a SQL-based query is built. If the input query is invalid then keyword agent converts it into SQL-based query by using rule base knowledge.

2.2. Natural language interfaces to knowledge bases (NLIKB)

In the context of natural language interfaces to knowledge bases there are prominent efforts to provide interfaces for a specific context such as (Doğmuş, Dogmus, Patoğlu, Patoglu, & Erdem, 2014) which introduces a controlled natural language interface (CNLI) for rehabilitation robotics ontology. This interface translates controlled queries to SPARQL-based queries by using query description trees and description logic concepts. In the context of CNLI, (Damljanović, Agatonović, Cunningham, & Bontcheva, 2013) presented a system for querying ontologies which generates a set of ranked query interpretation from user's question , and the answer is shown to the best ranked option. Also, this system provides a feedback to the user in order to deal with empty result set, and providing a clear idea to the user whether a query reformulation might or might not help in answering the question.

On the other hand, some efforts are focused on developing portable or configurable NLIKB, such as Quelo NLI (Franconi, Gardent, Juarez-Castro, & Perez-Beltrachini, 2014) which guides users in formulating queries by means of suggestions which allow adding, replacing or deleting snippets of English text at different points thereof. For this purpose, Quelo NLI links text spans of the natural language query with elements of the underlying formal query, i.e. concepts and relationships. Other approaches, such as SWSNL (Habernal & Konopík, 2013) (Semantic Web Search Using Natural Language) uses ontologies as a main mechanism for storing the domain structure and all the data, and to describe the meaning of user queries expressed in natural language. In SWSNL, the input query is analyzed by using NLP techniques such as Named Entity Recognition (NER) and Semantic analysis. Then, a KB-independent semantic representation of input query is built. Finally, this representation is transformed to a query language in order to be executed against the knowledge base.

Table 1 shows a comparative analysis of some of the most recent research efforts focused on NLI development. This analysis was carried out considering the following factors: (1) domain, (2) NL interface approach, (3) NLP techniques and (4) representation of the question.

Other works have used different approaches for the design and development of the NLI, some of them (Doğmuş et al., 2014) are based on a CNLI approach which limits the user's expressiveness due to the vocabulary restrictions imposed by the system. Other works (Doğmuş et al., 2014; Franconi et al., 2014; Llopis &