Information Sciences xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

QueryGen: Semantic interpretation of keyword queries over heterogeneous information systems

Q1 Carlos Bobed*, Eduardo Mena

Department of Computer Science & Systems Engineering, University of Zaragoza, Zaragoza 50018, Spain

ARTICLE INFO

Article history:
Received 26 October 2014
Revised 2 September 2015
Accepted 14 September 2015
Available online xxx

Keywords:
Semantic keyword interpretation
Semantic search
Semantic Web
Ontologies
Knowledge discovery

ABSTRACT

In the last years, users have become used to keyword-based search interfaces due to their ease of use. By matching input keywords against huge amounts of textual information and labeled multimedia files, current search engines satisfy most of users' information needs. However, the principal problem of this kind of search is the semantic gap between the input and the real user need, as keywords are a simplification of the query intended by the user. Moreover, different users could use the same set of keywords to search different information; even the same user could do it at different times. The search system, before accessing any data, should discover first the intended semantics behind the user keywords, in order to return only data fulfilling such semantics. The use of formal query languages is not an option for non-expert users, so a semantic keyword-based search based on *semantic interpretation of keyword queries* could be the solution, i.e., a search that starts discovering the semantics intended for the input user keywords, and then only data relevant to that semantics are returned as answer.

In this paper we present a system that performs semantic keyword interpretation on different data repositories. Our system (1) discovers the meaning of the input keywords by consulting a generic pool of ontologies and applying different disambiguation techniques, (2) once the meaning of each keyword has been established, the system combines them in a formal query that captures the semantics intended by the user, considering different formal query languages and possibilities that could arise, but avoiding inconsistent and semantically equivalent queries, and, finally, (3) after the user has validated the generated query that best fits her/his intended meaning, the system routes the query to the appropriate data repositories that will retrieve data according to the semantics of such a query. Experimental results show the semantic interpretation capabilities and the feasibility of our approach.

1. Introduction

The Web has made a huge and ever-growing amount of information available to its users. To handle and take advantage of this information, users have found in Web search engines their best allies. Most of these search engines have a keyword-based interface to allow users to express their information needs, as this is an easy way for users to define their searches. Thus, the adoption of keyword-based search interfaces has spread widely in the last few years. However, the ease of use of keyword search comes from the simplicity of its query model, whose expressivity is low compared with other more complex query models [31]. In fact, keyword queries are simplifications of the queries that really express the user's information need. On the other hand, the

Q2 * Corresponding author. Tel.: +34 665368692.
E-mail addresses: cbobed@unizar.es (C. Bobed), emena@unizar.es (E. Mena).

http://dx.doi.org/10.1016/j.ins.2015.09.013
0020-0255/© 2015 Published by Elsevier Inc.

use of expressive formal languages (such as SQL or SPARQL) is far from being easy for common users. Moreover, to effectively use formal languages, the user must have previous knowledge of the underlying schema and data s/he is accessing. Thus, the sweet spot would be to mix the expressivity of formal languages with the ease of use of keyword queries, while making the user unaware of the data sources being accessed to solve her/his information needs. Therefore, to deal with these problems, we advocate for a semantic keyword-based search based on *semantic interpretation of keyword queries*, a keyword-based search process in which semantics of both keywords and query languages play a crucial role during the whole search process.

In any search engine which has an unstructured query language as input, the main steps performed are: query construction, data retrieval, and presentation of results. Out of these three steps, the first one is crucial because the more accurate the system is able to capture the user's information need, the more precise results it will retrieve. However, the importance of this first step is usually underestimated by adopting unstructured query models (i.e., bag of keywords), making the quick access to huge amounts of data and the ranking of results the most important steps of the whole process, while leaving the burden of processing non-relevant data to the user. In this way, these approaches might miss what the user really wanted to retrieve as they hide less promoted results, making current search engines useless when looking for certain (non-popular) information.[1] To enhance the search process, we aim at enhancing the capture of the user's information need by combining both the benefits of the structured query models, and the ease of use and spread of the keyword search. When it comes to keyword queries, the process to translate them into a structured query is named *keyword query interpretation* [19]. For this task, several approaches (e.g., [42,46]) advocate starting with the discovery of the meaning of each keyword among the different possible combinations. For instance, the keyword "book" could mean "a kind of publication" or "to reserve a hotel room". These approaches consult a pool of ontologies (which offer a formal, explicit specification of a shared conceptualization [24]) and use disambiguation techniques to discover the intended meaning of each user keyword. So, plain keywords can be mapped to ontological terms (concepts, roles, or instances). However, direct interpretation might not be always possible as users tend to omit information in their keyword searches (the average number of keywords used in keyword-based search engines "is somewhere between 2 and 3" [36][2]), and therefore, relevant underlying knowledge should be used to enhance this interpretation.

In this paper, we delve into that line and present QueryGen, a system that performs semantic interpretation of keyword queries into multiple query language over different data repositories. Our system:

1. Discovers the meaning of the input keywords by consulting a generic pool of ontologies and disambiguates them taking into account their context (the rest of the keywords in the input set); i.e., each keyword in the input has an influence on the rest of the keyword's meanings. In this process, it retrieves and integrates knowledge about the input terms, which will be used in latter stages.
2. Then, as a given set of user keywords (even when their semantics have been properly established) could represent several queries, the system finds all the possible queries using the input keywords in order to precisely express the exact meaning intended by the user. This is done considering different formal query languages (the use of formal languages avoids ambiguities and expresses the user information in a precise way) which are made available to the system by semantically modeling them, and avoiding inconsistent and semantically equivalent queries with the help of a Description Logics (DL) reasoner [4]. During this process, our system considers the addition of *virtual terms*. These virtual terms represent missing keywords that users had in their mind but did not input[3]. This way, our system can explore further meanings when the user has given an incomplete input.
3. Finally, once the user has validated the generated query that best fits her/his intended meaning, our system routes the query to the appropriate structured data repositories that will retrieve data according to the semantics of such a query.

The architecture of our system is flexible enough to deal with different ontologies, formal query languages, and query processing capabilities of underlying data repositories. We aim at achieving the highest expressivity possible taking as starting point a plain keyword-based input as it is the most spread method to request information (using Web search engines). Moreover, our system is robust to incomplete inputs as, using the retrieved background knowledge, even in case that the user input is just a single keyword, our system is able to deal with it by exploring the implicit information description that the user had in mind when that keyword was posed.

In particular, the main contributions of this work are as follows:

- We present our approach to Semantic keyword interpretation, which is completely knowledge-guided. First, our system applies semantic techniques to disambiguate the input keywords and retrieve further knowledge about them; and then, it interprets the semantics of the input keywords structuring them according the semantics of the different formal query languages, obtaining a formal query that is posed to the available underlying data repositories.
- We propose a generalized keyword interpretation process based on semantic models of the query languages used to structure keyword queries (i.e. not tied to any particular query language). The semantic modeling framework proposed allows QueryGen not to be restricted to DL-based query languages, but to use any query language to interpret keyword queries as long as it is correctly modeled.

[1] According to the different criteria adopted by the ranking schema, which can take into account other aspects apart from actual popularity.
[2] This data still hold as for April 2015, http://www.keyworddiscovery.com/keyword-stats.html?date=2015-04-01, last accessed May 20, 2015.
[3] For example, a user looking for movies whose genre is "horror" could enter "horror movie", omitting the keyword "genre".

Please cite this article as: C. Bobed, E. Mena, QueryGen: Semantic interpretation of keyword queries over heterogeneous information systems, Information Sciences (2015), http://dx.doi.org/10.1016/j.ins.2015.09.013