Knowledge-Based Systems 21 (2008) 655–665

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Combining weights with fuzziness for intelligent semantic web search

Hai Jin [a,*], Xiaomin Ning [a], Weijia Jia [b], Hao Wu [a], Guilin Lu [c]

[a] Services Computing Technology and System Laboratory, Cluster and Grid Computing Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
[b] Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
[c] Department of Mathematics and Statistics, Shanghai LiXin University of Commerce, Shanghai 201620, China

ARTICLE INFO

Article history:
Received 12 July 2007
Accepted 24 March 2008
Available online 4 April 2008

Keywords:
Fuzzy description logic
Intelligent search
Rank
Semantic web
User preference
Weighting

ABSTRACT

Intelligent retrieval for best satisfying users search intensions still remains a challenging problem due to the inherent complexity of real-world semantic web applications. Usually, a search request contains not only vagueness or imprecision, but also personalized information goals. This paper presents a novel approach which formulates one's search request through tightly combining fuzziness together with the user's subjective weighting importance over multiple search properties. A special ranking mechanism based on the weighed fuzzy query representation is proposed. The ranking method generates a set of "degree of relevance" – an overall score which reflects both fuzzy predicates and the user's personalized preferences in the search request. Moreover, the ranking method is general and unique rather than arbitrary. Hence, search results shall be properly ordered in terms of their relevance with respect to best matching the search intension. The experimental results show that our approach can effectively capture users information goals and produce much better search results than existing approaches.

1. Introduction

The semantic web is the vision of having data on the web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications [4]. Extensive efforts have been made for the realization of the semantic web layered architecture. At present, the ontology layer in which data is encoded in ontology languages such as RDF/RDFS and OWL DL [3] rooted in Description Logics (DLs) [2], has reached sufficient maturity. Accordingly, some query languages which usually adopt an expressive SQL-like declarative syntax form including RQL, RDQL, SeRQL and SPARQL [37], have also been developed to accurately access the semantic web data. Hence, many recent research interests have focused on rules and proofs, such as rule languages of Description Logic Program (DLP) [14], SWRL [17] and their counterpart inference engines (e.g., FaCT++ [10], KAON2 [21], RacerPro [33]) which allow computers to conduct automated reasoning to deduce new information from knowledge bases (KBs).

Due to the intrinsical complexity of real-world knowledge, however, intelligent retrieval of semantic web data to best capture users search intensions still remains an open problem [42]. Firstly, the information intension of a search request usually involves vagueness or imprecision which pervasively other than merely an exception exists in many real-world applications of the semantic web such as multimedia processing and retrieval [12,27], medical diagnosis [13] and geospatial [7]. Consider, for instance, a knowledge base on scientific literatures. A Ph.D. candidate may raise search requests like "find *recent* publications *written by* Zadeh" or "find *highly* cited publications talking *about* description logic". In this case, "*recent*", "*highly*" and even "*about*" are so-called fuzzy or vague predicates [43] since they can hardly have a precise definition as the predicate "*written by*" does have. Current standard semantic web ontology languages as well as rule inferences are based on bivalent logic (i.e., either *true* or *false*) and hence fail to represent the above vague knowledge. Secondly, one may have his/her own quality measurement or favoritism among multiple criteria for interesting search results. That is, the user might inequally values the multiple criteria. For instance, the Ph.D. candidate issues the following request over the literatures KB: finding *recent* publications *highly* cited *written by* Zadeh, but *prefer* more weight on recentness *over* citation. Consider another example in a multimedia semantic web application. A user might want to retrieve movie clips which have a *predominantly* green scene with a *loud* noise in the sound track, but *favor* twice as much about the color of the picture *over* as about the loudness of the sound. Here, "*prefer* A *over* B" or "*favor* A *over* B" indicates the user's own importance measurements on two different criteria (i.e., A and B). Finally, unlike classical semantic web query languages

* Corresponding author. Tel.: +86 27 8754 3529; fax: +86 27 8755 7354.
E-mail addresses: hjin@hust.edu.cn (H. Jin), ningxm@hust.edu.cn (X. Ning), itjia@cityu.edu.hk (W. Jia), haowu@hust.edu.cn (H. Wu), luguilin@gmail.com (G. Lu).

0950-7051/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2008.03.040

which return results in disorder that only meet query conditions with an exact boolean matching, it is more appropriate to formulate the above desired search results as a set of "degree of relevance" (e.g., in the interval [0, 1]). The "degree of relevance" is an overall score which should reflect both the multiple search properties and the subjective measurements thus able to support ranked queries. Moreover, as the number of ontologies available in the semantic web increases rapidly, it is reasonable to say that specialized ranking strategies will be indispensable for better satisfying user requirements to prevent providing users with many search results in disorder.

This paper proposes a novel approach which enables intelligent semantic web search for best satisfying users search intensions. The approach combines the user's subjective weighting importance over multiple search properties together with fuzziness to represent search requirements. A special ranking mechanism based on the above weighed fuzzy query is also presented. The ranking method considers not only fuzzy predicates in the query, but also the user's personalized interests or preferences. Therefore, the search results shall be properly ordered in terms of their "degree of relevance" with respect to the query request. In addition, the ranking method is general and unique rather than arbitrary. To verify its effectiveness, we evaluate the approach over real-world data from CiteSeer metadata (http://citeseer.ist.psu.edu/oai.html) which covers literatures in the field of computer science and computer technology.

The rest of this paper is organized as follows. Section 2 reviews related work. We detail our proposed weighted fuzzy approach in Section 3. Section 4 reports experimental results. Finally, we conclude the paper and point out future research directions in Section 5.

2. Related work

Efficient and intelligent semantic web search plays an important role in the vision of the semantic web [15,29,34,41]. Rocha et al. [34] present a semantic search [15] approach which combines traditional search engine techniques together with ontology based information retrieval. The approach differs from traditional search engines in that it seeks to find important related entities to a given set of keywords using a spreading activation algorithm [8] with the domain knowledge provided by experts. But the expressivity of its keywords based search model is very limited since it is unable to answer more complex structured queries. Additionally, their approach could not handle multiple vague search requirements and personalized preferences as our work does. In our previous work [29], we present a framework RSS which implements ranked semantic search on the semantic web. In this framework, search results can be greatly expanded with entities which are most semantically related to the query. Although the RSS query model is much more expressive than [34], it is unable to deal with more complex structured fuzzy queries. [41] implements a personalized content management system which is to improve the retrieval process by taking into account user preferences on the semantic context of ongoing user activities. In the context of personalized retrieval, the personal relevance measure is combined with query-dependent and user-neutral search result rank values to produce the final rank score for a document. Although the search process takes user preferences into account, the query model is still based on crisp logic and hence fails to represent vagueness.

In recent years, many impressive works have been carried out for tackling with imprecise and vague queries in the semantic web [38,44]. These approaches usually apply a so-called fuzzy description logic [40] which is an integration of Zadeh's fuzzy logic [43] with classical DLs [2]. For example, Zhang et al. [44] proposes a model which integrates information retrieval with formal fuzzy DL based query and reasoning to utilize textual and semantic information for searching in semantic portals. Similar to ours, the model also supports IR related formal queries like "find publications *written by* Tim Berners-Lee talking *about* semantic web". However, their model could not deal with multiple fuzzy predicates since they only focused on textual information. It also lacks the important ranking capability as provided in our approach. Stoilos et al. [38] presents a fuzzy extension to OWL called Fuzzy OWL which can capture imprecise and vague knowledge. A prototype of fuzzy reasoning engine (FiRE) is also implemented to capture and reason about such fuzzy knowledge. FiRE does not support fuzzy datatype expressions [30] like "*recent* publications" or "*highly* cited publications" which are addressed by our work. Supporting flexible fuzzy queries has also been actively investigated in the databases area [6,23,27]. SQLf [6] allows imprecise querying of regular relational databases through extending the SQL language to satisfy user needs more closely. RankSQL [23] is a systematic and principled framework to support efficient evaluations of ranking (top-K) queries in relational database systems, by extending relational algebra and query optimization. In the system, ranking is taken as a first-class database construct which differs from SQLf. Meghini et al. [27] gives a fuzzy DL based multimedia information retrieval model which implements semantic based retrieval and offers a framework for the integration of the multimedia (including text and images) and multidimensional aspects (including form, content and structure). For personalized query in database system, Koutrika and Ioannidis [22] present a personalization framework for database queries based on information stored in structured user profiles that keep single, unconditional preferences. In this framework, a preference model assigns to each atomic query condition a personal degree of interest and then computes the degree of interest in complex query condition based on the degrees of interest in the constituent atomic ones.

To the best of our knowledge, the work that is most similar to ours is given in Ref. [36]. Siberski et al. [36] propose a comprehensive extension of SPARQL [37] which adds preference-based querying capabilities to SPARQL languages and hence directly supports the expression of preferences. They extend SPARQL with a new preferring solution sequence modifier. Therefore, users can specify which kind of answers they prefer by adding the preferring clauses to their queries like "FILTER (A or B) PREFERRING A". But the search expression is still restricted to crisp logic. In addition, the effectiveness of this method is also unknown, since no evaluation results are given. Our work also relates to f-SWRL [31] which is a fuzzy extension of the expressive rule language SWRL [17] to include fuzzy assertions and fuzzy rules. In f-SWRL, atoms can include a "weight" (a truth-value between 0 and 1) that represents the "importance" of the atom in a rule. For example, the f-SWRL rule: $\mathrm{Tall}(?x) * 0.7 \wedge \mathrm{Light}(?x) * 0.8 \rightarrow \mathrm{Thin}(?x)$, indicates that if one is Tall (with importance factor 0.7) and Light (with importance factor 0.8) then one is Thin. The semantics of weights are based on fuzzy aggregation functions, such as linear aggregation or weighted sum. Like [38], it also could not handle fuzzy datatypes, as we instead do.

3. Combining weights with fuzziness: our approach

This section details our approach for intelligent semantic web search. To make this paper self-explanatory, in Section 3.1 we first introduce fuzzy $\mathscr{ALC}(\mathbf{D})$ [39]. Then, Section 3.2 gives an overall architecture employing our approach, Section 3.3 describes the underlying fuzzy knowledge base, Section 3.4 presents user queries formulation and finally Section 3.5 depicts our ranking mechanism and retrieval.