Web Semantics: Science, Services and Agents on the World Wide Web 36 (2016) 32–57

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

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

Personalized concept-based search on the Linked Open Data

Melike Sah [a,*], Vincent Wade [b]

[a] Department of Computer Engineering, Near East University, Nicosia, North Cyprus, via Mersin 10, Turkey

[b] Centre for Global Intelligent Content, Knowledge and Data Engineering Group, Trinity College Dublin, Dublin, Ireland

ARTICLE INFO

Article history:
Received 16 January 2014
Received in revised form
31 August 2015
Accepted 26 November 2015
Available online 17 December 2015

Keywords:
Categorization
Concept-based search
Fuzzy retrieval model
Personalized search/exploration
Linked Open Data
UMBEL

ABSTRACT

In this paper, we present a novel personalized concept-based search mechanism for the Web of Data based on results categorization. The innovation of the paper comes from combining novel categorization and personalization techniques, and using categorization for providing personalization. In our approach, search results (Linked Open Data resources) are dynamically categorized into Upper Mapping and Binding Exchange Layer (UMBEL) concepts using a novel fuzzy retrieval model. Then, results with the same concepts are grouped together to form categories, which we call *concept lenses*. Such categorization enables concept-based browsing of the retrieved results aligned to users' intent or interests. When the user selects a concept lens for exploration, results are immediately personalized. In particular, all concept lenses are personally re-organized according to their similarity to the selected lens. Within the selected concept lens; more relevant results are included using results re-ranking and query expansion, as well as relevant concept lenses are suggested to support results exploration. This allows dynamic adaptation of results to the user's local choices. We also support interactive personalization; when the user clicks on a result, within the interacted lens, relevant lenses and results are included using results re-ranking and query expansion.

Extensive evaluations were performed to assess our approach: (i) Performance of our fuzzy-based categorization approach was evaluated on a particular benchmark (∼10,000 mappings). The evaluations showed that we can achieve highly acceptable categorization accuracy and perform better than the vector space model. (ii) Personalized search efficacy was assessed using a user study with 32 participants in a tourist domain. The results revealed that our approach performed significantly better than a non-adaptive baseline search. (iii) Dynamic personalization performance was evaluated, which illustrated that our personalization approach is scalable. (iv) Finally, we compared our system with the existing LOD search engines, which showed that our approach is unique.

1. Introduction

1.1. Motivations

With the adoption of the Linked Open Data (LOD) by a wider Web community, large volumes of semantic data are being generated. The challenge now is finding and exploring relevant information on the Web of Data (WoD). This is crucial for the uptake of the LOD by applications in order to support both ordinary Web and Semantic Web users. Thus, LOD search engines play a vital role for providing efficient access mechanisms. However current approaches such as Swoogle [1], Sindice [2], Watson [3], OKKAM [4] use keyword-based search and ranked result lists presentation of traditional information retrieval (IR), which is not very effective with the growing LOD [5]. Like the traditional Web, keyword queries perform well if the user has specific information needs. If the user's aim is to gather or explore information in unfamiliar domains, then the existing approaches provide little support. Although some approaches allow results filtering by class, predicate, etc., users are required to know technical details (i.e. URIs, SPARQL queries). This means these features are only available to Semantic Web experts. Some form of categorization can also be useful and OKKAM organizes the results into seven concepts (e.g. location, organization, artifact instance, artifact type, event and other). However, the provided concepts are too broad to be real value. There is little research investigating search problems on the WoD.

* Corresponding author.
E-mail addresses: melike.sah@neu.edu.tr (M. Sah), Vincent.Wade@scss.tcd.ie (V. Wade).

http://dx.doi.org/10.1016/j.websem.2015.11.004
1570-8268/

Another search paradigm for the LOD is faceted search/browsing systems, which provide facets (concepts) for interactive search and browsing [6]. Facets assist results filtering and exploration. However, the main limitation of faceted search is that facet creation depends on specific data and schema properties of underlying metadata and it can be difficult to generate useful facets to large and heterogeneous WoD [7–12].

Traditional IR has been investigating efficient search mechanisms for decades; results clustering and personalized search are two popular methods for enhancing search effectiveness. In clustering search, results are organized into categories for assisting users in results exploration and in disambiguation of the query (Snaket [13], Vivisimo.com, carrot2.org). Results categorization is also widely used, such as Google categories, Yahoo Directories and Open Directory Project (ODP). Although clustering search and faceted search seem similar, the latter filters results based on schema/metadata, whereas the former clusters results based on their meaning (language model).

Alternatively, personalized search aims to improve retrieval efficiency by adapting results to context/interests of individual users; thus the user can explore personally relevant results [14]. It is a popular research and commercial interest (i.e. Google). However, personalized search gained very little focus on the Semantic Web [15]. This could be because of isolated and low volumes of metadata created in early linked data initiatives. As the size of LOD increases, personalized search and interactions become more crucial.

Our objective is to improve current search mechanisms on the LOD with a novel personalized concept-based search. In order to achieve this, we address the following key challenges: (1) Categorization of LOD resources under a conceptual structure. (2) Providing a scalable system architecture for dynamic categorization of LOD resources. (3) Supporting a novel search interface based on concept-based results exploration. (4) Introducing a model such that results categorization can be used as a tool for adaptive results presentation. In addition, personalization should be non-intrusive and scalable. (5) Representing users' search intents using the available semantic and syntactic information. (6) Introducing a scalable similarity score which includes both semantic and syntactic similarity measures for results adaptation. All of the mentioned features were implemented and an online demo is publicly available as we discuss in Section 2.

1.2. Contributions

We innovatively combined results categorization and personalized IR to introduce a novel personalized concept-based search mechanism for the WoD. In our approach, users can use keyword or URI queries and results are organized into concepts using UMBEL conceptual vocabulary [16]. Here, the concept can be thought as a class, topic or category. To avoid any confusion, we will use the term concept in the rest of the paper. For categorization of results, we introduce a novel retrieval model, which works on any linked dataset, scalable and reasonably accurate (F-Measure of ~87% on 10,000 mappings). On the client-side, results with the most confident UMBEL concept categorization are grouped to form *concept lenses*. To clearly define, a concept lens corresponds to a UMBEL concept and the search results belonging to this concept. Similarly, when we say concept lenses, we mean a set of UMBEL concepts and the search results belonging to these concepts. Therefore, concept lenses depend on the actual search query. The aim of concept lenses is to aid results exploration. Then, based on user interactions with the results, we apply personalization in two cases: (i) When a user selects a concept lens. All concept lenses are reorganized based on conceptual similarity. In addition, within the selected lens, immediately more relevant results are included using results re-ranking and query expansion as well as similar lenses are suggested for exploration. (ii) When a user clicks onto a result, then we support interactive personalization. Last N clicks of the user within the search session are used to add relevant results to the interacted lens and relevant concept lenses are suggested for exploration. In our approach, users' search intents are represented as a vector of concepts (semantic) and terms (syntactic). Thus a combined similarity measure can be used for results adaptation.

Please note that our personalization approach is non-intrusive, privacy preserving and scalable, since it does not require an explicit user login and the personalization is implemented at the client-side. In addition, our approach is adaptable and can be plugged on top of any linked data search engine; in this paper, we use Sindice [2]. It only requires categorizations of resources using UMBEL, which can be achieved by number of methods such as our retrieval model [17].

Our contributions can be summarized as follows:

- We introduce a novel personalized concept-based search and exploration mechanism for the WoD. To the best of our knowledge, no such previous work exists.
- We introduce a novel and robust categorization algorithm using a novel fuzzy retrieval model. In our fuzzy retrieval model, first we extract terms from search results (LOD resources). Then, relevancy of the extracted terms on UMBEL concepts is calculated using a fuzzy function. In our fuzzy function, we calculate membership scores for each semantic part of a UMBEL concept using an extended $tf \times idf$ model. Then, these scores are combined and an ontological relationships driven voting algorithm is applied in order to categorize the search result into a specific UMBEL concept.
- Categorizations are utilized for concept-based results presentation. Results with the same most confident categorization are grouped together and presented, which we call concept lenses. In this way, users can explore results by clicking on concepts.
- We suggest the use of results categorization as a tool for personalized concept lenses re-ranking, results re-ranking, query expansion and lenses suggestion. To the best of our knowledge, we are the first to combine these four personalization techniques based on concept-based search results presentation and user click histories. In related works [13,14], ontology is utilized for results re-ranking and query expansion. However, results are presented as traditional ranked lists. On the other hand, in our concept-based personalization approach, a vocabulary (UMBEL in this case) is the key element of categorization, concept-based presentation and personalization of results. In particular, we track user clicks on both concept lenses and individual search results. We develop personalization strategies and personalized ranking models to rank both concept lenses and results based on user actions. In addition, we predict relevant concept lenses based on ontological relationships between concepts. Finally, we use user clicks and UMBEL concept labels for query expansion. The evaluations have indicated that the use of our personalization strategies and lenses approach provides significantly better results than a non-adaptive baseline search systems.

A part of this work has been presented in [17–19]. In this paper, we extend the explanations and evaluations of the framework. In particular, we introduce two new evaluations: Compare our approach with the existing LOD search engines and present a new user study using 32 participants. The user study consisted of a task-based evaluations based on real life user information needs regarding to a tourist domain. In the user study, we compared our personalized concept-based search with a non-adaptive baseline search system in order to assess the added value of categorization and personalization.