Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 1-5, 2012.

A Novel Concept-based Search for the Web of Data

Melike Sah and Vincent Wade
Knowledge and Data Engineering Group, Trinity College Dublin, Dublin, Ireland
{Melike.Sah, Vincent.Wade}@scss.tcd.ie

Abstract. With the increasing volumes of data, access to the Linked Open Data (LOD) becomes a challenging task. Current LOD search engines provide flat result lists, which is not an efficient access method to the Web of Data (WoD). In this demo, we introduce a novel and scalable concept-based search mechanism on the WoD, which allows searching based on meaning of objects. In particular, the retrieved resources are dynamically categorized into UMBEL vocabulary concepts (topics) using a novel fuzzy retrieval model and resources with the same concepts are grouped together to form categories, which we call *concept lenses*. In addition, search results are presented with hierarchy of categories and concept lenses for easy access to the LOD. Such categorization enables concept-based browsing of the retrieved results aligned to users' intent or interests. Results categorization can also be used to support more effective personalized presentation of search results.
Keywords: Categorization, concept-based search, semantic indexing, fuzzy retrieval model, linked open data, UMBEL, scalability, demo.

1 Introduction

Linked Open Data (LOD) or the Web of Data (WoD) is becoming a de-facto for publishing structured and interlinked data according to a set of Linked Data principles and practices. The main promise of the LOD is providing rich Web-scale interlinked metadata, which can be consumed by Web applications in more innovative ways that was not possible before. However, as the number of datasets and data on the LOD is increasing, the challenge turn into finding and accessing the relevant datasets and data. Thus, LOD search engines are becoming more important to enable exploration and browsing of LOD data and search engines are crucial for the uptake of the WoD.

On the other hand, current WoD search engines and mechanisms, such as Sindice [1] and Watson [2], display the search results as ranked lists. In particular, they present the resource title or example triples about the resource in the search results. However, presentation of resource titles is not an efficient presentation method for the WoD since users cannot understand "what the resource is about" without opening and investigating the LOD resource itself. Sig.ma service or Sig.ma end-user application, attempts to solve this problem with a data mash-up based presentation paradigm by using querying, rules, machine learning and user interaction [3]. The user can query the WoD and Sig.ma presents rich aggregated mashup information about the retrieved

resources. Sig.ma's focus is on data aggregation and it is not for search results presentation. Another search paradigm for the LOD is faceted search/browsing, which provide facets (categories) for interactive searching and browsing [4]. The main limitation of the faceted search mechanisms is that facet generation depends on specific data/schema properties of underlying metadata. Thus it can be challenging to generate useful facets to large and heterogeneous WoD [5]. It is evident that more efficient WoD search mechanisms are needed for the uptake of LOD by a wider community.

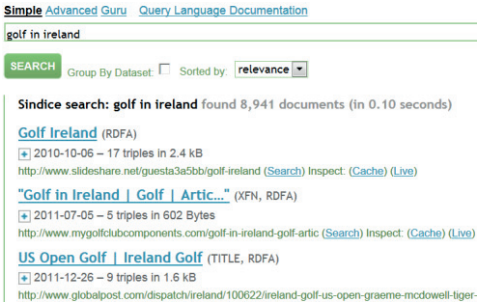
To overcome this issue, we introduce and demonstrate a novel concept-based search for the WoD using UMBEL concept hierarchy (http://umbel.org) and a novel fuzzy retrieval model [6][8]. In particular, the WoD is searched and the retrieved results are categorized into concepts based on their meaning. Then, LOD resources with the same concepts are grouped to form categories, which we call *concept lenses*. In this way, search results are presented using a hierarchy of categories and concept lenses, which can support more efficient access and browsing of results rather than flat result lists. Moreover, categories can be used for efficient personalization.

There are three unique contributions of our approach [6]: (1) For the first time, UMBEL is used for concept-based Information Retrieval (IR). (2) A second contribution is in novel semantic indexing and fuzzy retrieval model, which provides efficient categorization of search results in UMBEL concepts. (3) A minor contribution is the realization of a concept-based search realm to WoD exploration. Concept-based search has occurred in traditional Web. In this paper, we improve our previous work [6]: (1) With a more scalable system architecture, where the system performance can scale by using an indexing service at the server-side for dynamic categorizations. (2) In our previous work, a flat list of concepts was presented. In the current version, we improved the presentation by organizing concepts into a hierarchy, where users can locate relevant lenses using hierarchical organization. In this demo, we discuss the benefits of our approach compared to traditional WoD search engines using scenarios. In addition, we will discuss how the two main challenges, categorization accuracy and system performance, are resolved.

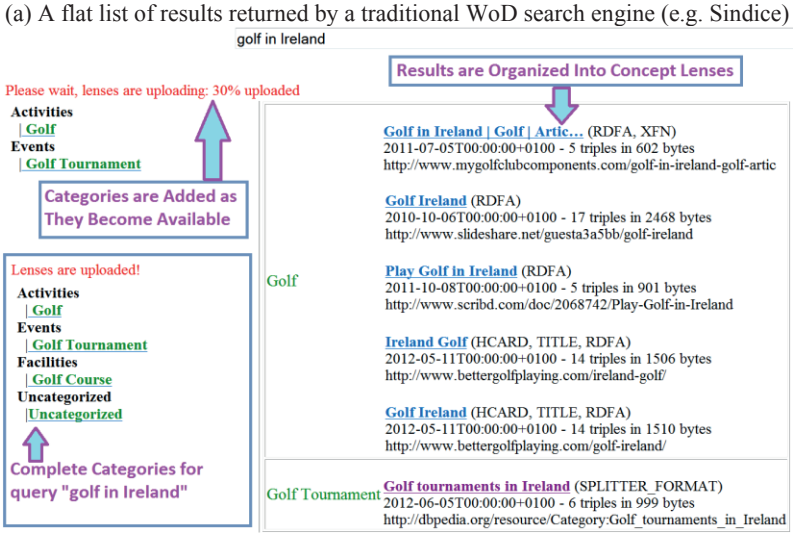
2 A Search Scenario on the Web of Data

To better illustrate the benefits of the proposed concept-based search, we describe a real life search scenario on the WoD (see our demo [8]). Assume Sue is knowledge engineer designing a website for "tourism in Ireland". She is designing an ontology to structure the site content and wants to populate the ontology with metadata and instances. Assume this ontology contains activities in Ireland, such as "golf". First, she searches for existing information on the WoD using "golf in Ireland" query. As shown in Figure 1(a), such a query may return many diverse results by a traditional WoD search engine (e.g. Sindice in this example). In this case, she needs to open and investigate large number of results for its suitability to her investigation. On contrary, when the same query is searched on our concept-based search, the results are automatically categorized and presented with hierarchy of categories and concept lenses as shown in Figure 1(b). In this case, Sue can discard irrelevant matches easily and can locate matching resources based on their concepts. In this example, Sue may no-

tice that she can include classes and metadata about golf courses and golf tournaments in her ontology. In general, hierarchical search results clustering have the advantage of providing shortcuts to the items that have similar meaning. It also allows better topic understanding and favours systematic exploration of search results [7]. Our concept-based search is unique on the LOD to support such search results exploration on the WoD. Moreover, as seen in Figure 1, most of the results are Web pages that contain embedded metadata. Thanks to robust categorization, our approach is applicable to categorization of Web pages on the Web (in most cases we only use URL labels) as well as categorization of LOD resources on the Semantic Web.



(a) A flat list of results returned by a traditional WoD search engine (e.g. Sindice)



(b) The same results are presented with categories and concept lenses by our approach

Fig. 1. Comparison of (a) traditional and (b) concept-based search for query "golf in Ireland"

3 Proposed Concept-based Search

System Architecture (Figure 2). Client-side is developed with Javascript and AJAX (parallel processing and incremental presentation for performance). Java Servlets are utilized at the server side where we use Jena for processing RDF and Lucene IR framework for indexing and implementation of categorization. Sindice Search and Sindice Cache APIs are used for searching the WoD and accessing RDF descriptions

of LOD resources. In our approach, results that are retrieved by the Sindice Search are further processed to categorize into categories. For this purpose, features are extracted from LOD resources and matched to UMBEL concept descriptions using a fuzzy retrieval model [6]. Categorized LOD resources are cached to a local index for system performance and sent to client for presentations with categories and concept lenses. Our search mechanism can work with any query and on any dataset because of a proposed robust categorization method and broad concepts provided by UMBEL.

UMBEL Concept Vocabulary. UMBEL is sub-set of OpenCyc. It provides broad topics (~28,000 concepts) with useful relations and properties drawn from OpenCyc (i.e. broader/narrower classes, preferred/alternative/hidden labels). UMBEL concepts are also organized into 32 supertype classes (e.g. Event, Activities, Places, etc.), which make it easier to reason, search and browse. In traditional concept-based IR systems, the concept descriptions are indexed using a vector space model (i.e. term frequency, inverse document frequency – *tf*×*idf*). For more efficient representations, we applied a novel semantic indexing model; associated weight of the term to the concept depends on where the term appears in a structured concept description (i.e. in URI label, preferred/alternative labels, sub/super-concepts labels).

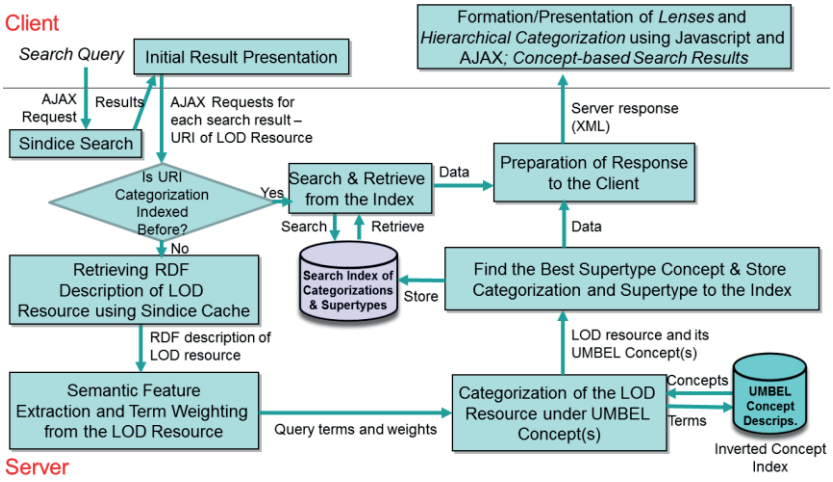


Fig. 2. System Architecture

Feature Extraction from the Context of LOD Resources. In order to categorize LOD resources under UMBEL concepts, lexical information is mined from the common features of LOD resources, such as *URI*, *label*, *type*, *subject* and *property names*. Moreover, a semantic enrichment technique is applied to gather more lexical information from the LOD graph by traversing owl:sameAs links. From the extracted terms, stop words are removed and the terms are stemmed into their roots. Then, the terms are weighted according to their term frequency and where they appear in the LOD resources; i.e. terms that are appear in subject and type fields may provide more contextual information about the resource. Thus, they are weighted higher.

Categorization of LOD Resources. The extracted terms from the LOD resource is matched against UMBEL concept descriptions using a novel fuzzy-based retrieval model. Proposed fuzzy retrieval model generates a fuzzy relevancy score according to

relevancy of a term to semantic elements (structure) of concept(s) ([6] for details): For example, UMBEL concepts are organized into a hierarchy of concepts. A concept may have relevant terms in concept, more specific terms in sub-concepts and more general terms in super-concepts. Instead of combining all the terms from the concept, sub-concepts and super-concepts, we weight term importance based on where they appear. Then, a fuzzy retrieval model combines term weights and a voting algorithm is applied to decide the final categorization of the LOD resource. Moreover, supertype class of the UMBEL concept needs to be found for hierarchical presentation of categories. In UMBEL, a concept might belong to more than one supertype class. We apply a voting algorithm, i.e. supertype class with the highest *tf×idf* rank of all LOD terms will be selected as the best representing supertype for that UMBEL concept.

Client-Side. At the client-side, a script (Javascript functions) processes the server responses and incrementally generates/updates hierarchical categories as well as concept lenses using AJAX. In this way, we prevent long delays in server responses.

Indexing for a Scalable Performance. In our approach, search results are processed in parallel for a scalable performance. In this paper, the system performance is enhanced further by adding a search index at the server-side. After the categorization, UMBEL and supertype concepts of the LOD URI are indexed. Since the index size affects search performance and the required disk space, we only index concept names without the base namespace. When a URI is requested, first the indices are searched; if URI has not been processed before, we apply dynamic categorization. Thus, we achieve significant decrease in network traffic and supply on-time categorizations.

Evaluations. Extensive evaluations are carried out to test the performance of our system on a benchmark of ~10,000 DBpedia mappings (see [6]). Evaluations showed that the proposed fuzzy retrieval model achieves very promising results ~90% precision, which is crucial for the correct formation of categories and the uptake of the proposed concept-based search. Moreover, the system performance can scale thanks to parallel processing and the use of search indices with minimum disk space.

4 References

1. Delbru, R., S., Campinas, G., Tummarello: Searching Web Data: an Entity Retrieval and High-Performance Indexing Model. Journal of Web Semantics, vol. 10, pp. 33-58, (2012)
2. D'Aquin, M., E., Motta, M., Sabou, S., Angeletou, L., Gridinoc, V., Lopez and D., Guidi: Toward a New Generation of Semantic Web Applications. IEEE Intelligent Systems (2008)
3. Tummarello, G., R., Cyganiak, M., Catasta, S., Danielczyk, R., Delbru and S., Decker: Sig.ma: live views on the Web of Data, Journal of Web Semantics, 8(4), pp. 355-364 (2010)
4. Heim, P., T., Ertl and J., Ziegler: Facet Graphs: Complex Semantic Querying Made Easy, Extended Semantic Web Conference (ESWC), LNCS, vol. 6088, pp. 288-302, (2010)
5. Teevan, J., S. T., Dumais and Z. Gutt.: Challenges for Supporting Faceted Search in Large, Heterogeneous Corpora like the Web. Workshop on HCIR, (2008)
6. Sah, M., and V., Wade. A Novel Concept-based Search for the Web of Data using UMBEL and a Fuzzy Retrieval Model. Extended Semantic Web Conference (ESWC), (2012)
7. Carpineto, C., S., Osinski, G., Romano, and D. Weiss. A Survey of Web Clustering Engines. ACM Computing Surveys, 41(3), 2009.
8. A demo is available online at https://www.scss.tcd.ie/melike.sah/golf_demo.swf