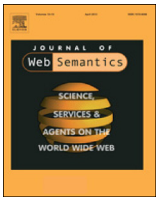
Web Semantics: Science, Services and Agents on the World Wide Web 37–38 (2016) 170–183

Contents lists available at ScienceDirect

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

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



Contextualized ranking of entity types based on knowledge graphs

Alberto Tonon [a,*], Michele Catasta [b], Roman Prokofyev [a], Gianluca Demartini [c], Karl Aberer [b], Philippe Cudré-Mauroux [a]

[a] *eXascale Infolab, University of Fribourg, Switzerland*
[b] *EPFL, Lausanne, Switzerland*
[c] *Information School, University of Sheffield, UK*

ARTICLE INFO

Article history:
Received 1 April 2015
Received in revised form
6 October 2015
Accepted 22 December 2015
Available online 6 January 2016

Keywords:
Entity typing
Ranking
Context
Crowdsourcing
Knowledge graphs

ABSTRACT

A large fraction of online queries targets entities. For this reason, Search Engine Result Pages (SERPs) increasingly contain information about the searched entities such as pictures, short summaries, related entities, and factual information. A key facet that is often displayed on the SERPs and that is instrumental for many applications is the entity *type*. However, an entity is usually not associated to a single generic type in the background knowledge graph but rather to a set of more specific types, which may be relevant or not given the document context. For example, one can find on the Linked Open Data cloud the fact that Tom Hanks is a person, an actor, and a person from Concord, California. All these types are correct but some may be too general to be interesting (e.g., person), while other may be interesting but already known to the user (e.g., actor), or may be irrelevant given the current browsing context (e.g., person from Concord, California). In this paper, we define the new task of ranking entity types given an entity and its context. We propose and evaluate new methods to find the most relevant entity type based on collection statistics and on the knowledge graph structure interconnecting entities and types. An extensive experimental evaluation over several document collections at different levels of granularity (e.g., sentences, paragraphs) and different type hierarchies (including DBpedia, Freebase, and schema.org) shows that hierarchy-based approaches provide more accurate results when picking entity types to be displayed to the end-user.

1. Introduction

A large fraction of online queries targets entities [1]. Commercial search engines are increasingly returning rich Search Engine Result Pages (SERPs) that contain not just ten blue links but also images, videos, news, etc. When searching for a specific entity, users may be presented in the SERP with a summary of the entity itself taken from a background knowledge graph. This search task is known as ad-hoc object retrieval [2,3], that is, finding an entity described by a keyword query in a structured knowledge graph. After correctly identifying the entity described by the user query, the subsequent task is that of deciding what entity information to present on the SERP among all potential pieces of information available in the knowledge graph. It is possible, for example, to display pictures, a short textual description, and related entities.

One interesting entity facet which can be displayed in the SERP is its *type*. In public knowledge graphs such as Freebase, entities are associated with several types. For example, the entity 'Peter Jackson' in Freebase[1] has 17 types, among which 'Person', 'Ontology Instance', 'Film director', and 'Chivalric Order Member' can be found. When deciding what to show on the SERP, it is important to select the few types the user would find relevant only. Some types are in most cases not compelling (e.g., 'Ontology Instance') while other types (e.g., 'Film director') may be interesting for a user who does not know much about the entity. Users who already know the entity but are looking for some of its specific facets might be interested in less obvious types (e.g., 'Chivalric Order Member', and its associated search results).

More than just for search, entity types can be displayed to Web users while browsing and reading Web pages. In such a case, pop-ups displaying contextual entity summaries (similar to the ones displayed on SERPs like the Google Knowledge Panel) can be shown

* Corresponding author.
E-mail addresses: alberto.tonon@unifr.ch (A. Tonon), michele.catasta@epfl.ch (M. Catasta), roman.prokofyev@unifr.ch (R. Prokofyev), g.demartini@sheffield.ac.uk (G. Demartini), karl.aberer@epfl.ch (K. Aberer), philippe.cudre-mauroux@unifr.ch (P. Cudré-Mauroux).

http://dx.doi.org/10.1016/j.websem.2015.12.005
1570-8268/© 2016 Elsevier B.V. All rights reserved.

[1] http://www.freebase.com/edit/topic/en/peter_jackson.

to the users who want to know more about a given entity she is reading about. In this case again, picking the types that are relevant is critical and highly context-dependent.

A third example scenario is to use selected entity types to summarize the content of Web pages or online articles. For example, one might build a summary for a given news article by extracting the most important entities in the article and listing their most relevant types (e.g., 'this article is about two actors and the president of Kenya').

In this paper, we focus on the novel task of ranking available entity types based on their relevance given a context. We propose several methods exploiting the entity type hierarchy (i.e., types and their subtypes like 'person' and 'politician'), collection statistics such as the popularity of the types or their co-occurrences, and the graph structure connecting semantically related entities (potentially through the type hierarchy).

We experimentally evaluate our different approaches using crowdsourced judgements on real data and extracting different contexts (e.g., word only, sentence, paragraph) for the entities. Our experimental results show that approaches based on the type hierarchy perform more effectively in selecting the entity types to be displayed to the user. The combination of the proposed ranking functions by means of learning to rank models yields to the best effectiveness. We also assess the scalability of our approach by designing and evaluating a Map/Reduce version of our ranking process over a large sample of the CommonCrawl dataset[2] exploiting existing `schema.org` annotations.

In summary, the main contributions of this paper are:

- The definition of the new task of entity type ranking, whose goal is to select the most relevant types for an entity given some context.
- Several type-hierarchy and graph-based approaches that exploit both schema and instance relations to select the most relevant entity types based on a query entity and the user browsing context.
- An extensive experimental evaluation of the proposed entity type ranking techniques over a Web collection and over different entity type hierarchies including YAGO [4] and Freebase by means of crowdsourced relevance judgements.
- A scalable version of our type ranking approach evaluated over a large annotated Web crawl.
- The proposed techniques are available as an open-source library[3] as well as an online web service.[4]

The present work is based on our previous contribution on type ranking [5]. However, we extend our previous article in several different ways: We propose a new context-aware approach to rank entity types that extends the notion of context in which an entity appears to exploit the text surrounding it in addition to other co-occurring entities (Section 6.4), and a new method that mixes different features coming from both the knowledge base, including entity popularity, and the type hierarchy (Section 6.5). We report additional details on the methods we used to build our text collection, including a pilot study we did in order to evaluate the best task design to collect relevance judgements (Section 7). We add a discussion on the relation among the features we take into consideration and on the comparison between hierarchy based and context-aware methods for ranking entity types (Section 9).

The rest of the paper is structured as follows. We start below by describing related work from entity-search and ad-hoc object retrieval. Then, we introduce the most important concepts in the Web of Data (Section 3) to formally define our new type ranking task in Section 4. In Section 5 we present the architecture of our system, and in Section 6 propose a series of approaches to solve it based on collection statistics, type hierarchies, and entity graphs. Section 8 presents experimental results comparing the effectiveness of our various entity ranking approaches over different document collections and type hierarchies as well as a scalability validation of our Map/Reduce implementation over a large corpus. Finally, we conclude in Section 10.

2. Related work

Entity-centric data management is a re-emerging area of research at the overlap of several fields including Databases, Information Retrieval, and the Semantic Web. In this paper we target the specific problem of assigning types to entities that have been extracted from a Web page and correctly identified in a pre-existing knowledge graph.

Named Entity Recognition and Ranking. Classic approaches to Named Entity Recognition (NER) typically provide as output some type information about the identified entities; In most cases, such types consist of a very limited set of entities including Person, Location, and Organization (see e.g., [6,7]). While this is useful for applications that need to focus on one of those generic types, for other applications such as entity-based faceted search it would be much more valuable to provide specific types that are also relevant to the user's browsing context.

In the field of Information Retrieval, entity ranking has been studied for a few years. Historically, the first entity-oriented task being addressed was expert finding [8] where the focus is on one specific entity type, that is, people. The goal is to find people who are knowledgeable about the requested topic. After this, works have looked at how to search for multiple entity types. Early approaches on entity ranking focused on entities of different types which are present in Wikipedia [9,10]. In the IR context, TREC[5] organized an Entity Track where different entity-centric search tasks have been studied: Four entity types were considered in that context, i.e., people, products, organizations, and locations. Type information can also be used for entity search tasks, e.g., by matching the types of the entities in the query to the types of the retrieved entities (see for instance [11]). Moreover, the IR community has organized workshops on entity search topics at the major research venue [12,13].

More recently, we have proposed a hybrid approach to rank entity identifiers as answer to a Web search query [3]. We used both standard IR methods based on inverted indexes as well as a structured search approach that exploits the graph structure (also including type information) connecting entities among each other to improve search effectiveness. In [14] the authors show how the number of entities used for the graph-based search step influences search effectiveness. Related to this is the aggregation of all data available about a specific entity [15], also including its types.

In the NLP field, entity extraction methods are continuously being developed. Here also, the types that are considered are typically rather limited. For example, in the method proposed in [16] 18 types are considered. In [17,18], authors propose a NER system to recognize 100 entity types using a supervised approach. The starting point to define the 100 entity types is the BBN linguistic collection[6] which includes 12 top types and 64 subtypes.

Entity types. The Semantic Web community has been creating large-scale knowledge graphs defining a multitude of entity types.

[2] http://commoncrawl.org/.

[3] https://github.com/MEM0R1ES/TRank.

[4] http://trank.exascale.info.

[5] http://trec.nist.gov.

[6] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T33.