Browsing DBpedia Entities with Summaries

Andreas Thalhammer and Achim Rettinger

AIFB, Karlsruhe Institute of Technology
{thalhammer, rettinger}@kit.edu

Abstract. The term "Linked Data" describes online-retrievable formal descriptions of entities and their links to each other. Machines and humans alike can retrieve these descriptions and discover information about links to other entities. However, for human users it becomes difficult to browse descriptions of single entities because, in many cases, they are referenced in more than a thousand statements.
In this demo paper we present SUMMARUM, a system that ranks triples and enables entity summaries for improved navigation within Linked Data. In its current implementation, the system focuses on DBpedia with the summaries being based on the PageRank scores of the involved entities.

Keywords: entity summarization, DBpedia, linked data, statement ranking

1 Introduction

The goal of the Linked Data movement is to enrich the Web with structured data. While the formal nature of these knowledge descriptions targets machines as immediate consumers, the final product is typically consumed by humans. Examples like Wikipedia Infoboxes show that, in many cases, next to textual descriptions users also want to browse structured data in order to get a quick overview about common or main facts of a data object. However, state-of-the-art interfaces like the one of DBpedia deliver all known facts about an entity in a single Web page. Often, the first thing users see when browsing a DBpedia entity are the values of `dbpedia-owl:abstract` in ten different languages. As a first attempt to overcome this issue, we introduce SUMMARUM, a system that ranks triples in accordance to popularity and enables entity summaries for improved navigation within Linked Data. In its current implementation, the system focuses on DBpedia with the summaries being based on the PageRank scores of the involved entities. We also adopted navigation elements from Semantic MediaWiki [3] in order to enable more flexible browsing.

The system is available at `http://km.aifb.kit.edu/services/summa/`.

2 Related Work

The field of browsing Linked Data entities has already been explored thoroughly. For the sake of conciseness, we focus on the most related and/or recent work in this field.

Recent efforts for producing user-friendly interfaces for Linked Data entities include the new *DBpedia interface* (currently available via DBpedia Live)[1] and Magnus Manske's *Reasonator* tool[2] which is based on Wikidata.[3] In the new DBpedia interface, all property-value pairs are ordered in the traditional DBpedia fashion, with values sorted alphabetically in accordance to their labels. In the Reasonator tool, the listings of statements do not seem to implement a particular order.

Similar tools are *aemoo* [4] and *LODPeas* [2]. *aemoo* focuses on schema information: of which class is an entity and to which other classes does the currently browsed entity relate. Further interaction with the related classes enables to detect additional entities of the respective type which can be browsed. *LODPeas* enables to browse further entites that are related to the currently browsed entity. The system makes use of a "concurrence index" which enables to suggest entities that share common property-value pairs. Both systems are focused on presenting entities that are not necessarily directly attached to the currently browsed entity.

Semantic MediaWiki [3] offers search by property-value pairs[4], e.g. by specifying `[[Born In::Hawaii]]`. We adopt this scheme in order to enable users to discover entities which share a specific attribute with the currently browsed one. Thus, browsing `dbpedia:Barack_Obama`, it is possible to discover who else was born in `dbpedia:Hawaii`.

The three major search engines, *Google*, *Yahoo*, and *Bing* also offer summaries of entities. Bing and Google also retrieve lists of entities that are focused on a property-value pairs, e.g. "movies `directed by Quentin Tarantino`". However, this seems to work only in specific domains as querying for "people `born in Hawaii`" does not result in a list of entities.

3 DBpedia PageRank

For our popularity-based approach, we computed the PageRank [1] scores for each DBpedia entity. As a basis for this, we used DBpedia's Wikipedia Pagelinks (en)[5] dataset. This dataset contains triples of the form "Wikipedia page A links to Wikipedia page B". We only use these untyped links, i.e. do not make use of typed links (e.g., `dbpedia-owl:birthplace`) for computation and thus, the computed scores reflect the PageRank of the associated Wikipedia pages. However, we call the dataset "DBpedia PageRank" as the link extraction is performed by the DBpedia framework and the resources are identified with DBpedia URIs.

For the computation of PageRank we used the original formula as described in [1] with a damping factor of 0.85. The number of iterations was set to 40

[1] http://live.dbpedia.org/

[2] http://tools.wmflabs.org/reasonator/

[3] http://wikidata.org

[4] http://semantic-mediawiki.org/wiki/Help:Semantic_search

[5] Wikipedia Pagelinks (en) – http://wiki.dbpedia.org/Downloads39#wikipedia-pagelinks

while the score changes from 20 iterations onwards were marginal and thus, suggest convergence. We publish the computed PageRank scores for the English language DBpedia versions 3.8 and 3.9 at `http://people.aifb.kit.edu/ath/#DBpedia_PageRank`. The dataset is available in tab-separated values and also in Turtle format. For the Turtle representation we used the vRank vocabulary[6] [5].

4 Implementation

The SUMMARUM system is implemented as a Web Service which accepts three query parameters as input:

entity* the URI of a DBpedia entity that the user wants to browse.
k* the maximum number of statements the user wants to retrieve about the entity.
predicate the URI of a DBpedia predicate. If this parameter is present, the system focuses on statements that involve the given entity in combination with the given predicate.

The parameters marked with the star symbol (*) are mandatory. The *predicate* parameter is optional. As an example, it can be used to retrieve a ranked list of statements with `dbpedia-owl:birthPlace` as a predicate combined with the entity `dbpedia:Hawaii`.

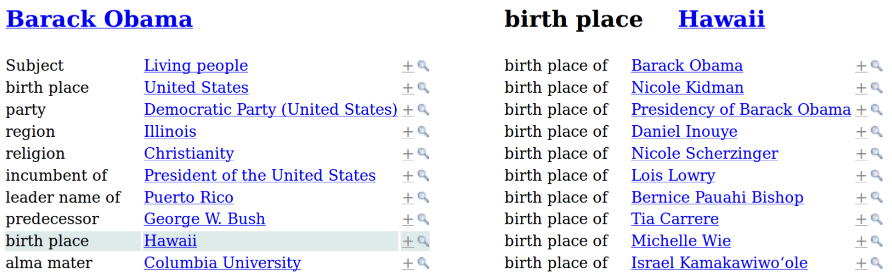
The system currently focuses on statements that involve two DBpedia entities[7] and, as such, does not consider statements with literal values, classes, or external resources. For each entity we use its incoming and outgoing typed links. Thus, the result is a mix of statements where the summarized entity is either in the subject or object position. This also includes results of queries where the *predicate* parameter was given. For example, using `dbpedia-owl:order` in combination with `dbpedia:Apodiformes` will retrieve statements where the entity is in the subject or object position of `dbpedia-owl:order`.

The decision on whether to include a statement in the top-k summary or not depends on the rank position. The score of a statement is the sum of the PageRank scores of the subject and the object. It has to be noted that, with the focus on a specific entity, its own score is not needed for the ranking and appears superfluous as the entity's score influences each ranked statement equally. In fact, we add the score for reasons of consistency as we publish each statement's score in the Turtle output of the service. Using only the subject's (resp. object's) score for ranking the statement would produce the same ranking but two different versions of the statement's score depending on whether the subject or the object is currently in focus.

In many cases, there are more than one statement with the same subject-object pair. Often, this is due to the distinction between DBpedia "property"

[6] vRank – `http://purl.org/voc/vrank`

[7] All DBpedia resources with the prefix `http://dbpedia.org/resource`.



Barack Obama

Subject	Living people
birth place	United States
party	Democratic Party (United States)
region	Illinois
religion	Christianity
incumbent of	President of the United States
leader name of	Puerto Rico
predecessor	George W. Bush
birth place	Hawaii
alma mater	Columbia University

birth place **Hawaii**

birth place of	Barack Obama
birth place of	Nicole Kidman
birth place of	Presidency of Barack Obama
birth place of	Daniel Inouye
birth place of	Nicole Scherzinger
birth place of	Lois Lowry
birth place of	Bernice Pauahi Bishop
birth place of	Tia Carrere
birth place of	Michelle Wie
birth place of	Israel Kamakawiwoʻole

Fig. 1: HTML summary of `dbpedia:Barack_Obama` (left) and the ranked list of statements with `dbpedia-owl:birthPlace` and `dbpedia:Hawaii` (right).

and "ontology" predicates. For these cases, we apply a simple heuristic to decide which statement we present: First, we prefer statements with the entity in the subject role over those with the entity in the object role. Second, we prefer the DBpedia "ontology" over "property" predicates. In all other cases, we select the first statement with the respective subject-object pair.

The SUMMARUM system supports two types of output via content negotiation: HTML (`text/html`) and Turtle (`text/turtle`).

The HTML version is intended for human consumption and thus, includes only a list of ranked statements without scores. The option for browsing entities in combination with predicates resembles the search interface of Semantic MediaWiki. Figure 1 shows two screenshots of the HTML interface.

The Turtle version can be used by machines for further processing or different interfaces and also includes the scores of the statements. For the representation, we use reification of statements in combination with the vRank vocabulary [5]. An example for the output is given in Listing 1.[8]

5 Conclusion and Future Work

Our work adds popularity-based entity summaries to known Linked Data browsing interfaces in order to enhance user experience. We show a live demonstration online and also provide machine-readable output for further reuse of the rankings.

In future versions of SUMMARUM we would like to address the following points:

Predicates In our next major release we plan to focus on the predicate component of the triple.

[8] Query: `http://km.aifb.kit.edu/services/summa/summarum?entity=http://dbpedia.org/resource/Hawaii&predicate=http://dbpedia.org/ontology/birthPlace&k=1`

Listing 1: Example result in Turtle (the namespaces are omitted).

```
[ rdf:type rdf:Statement ;
  rdf:subject <http://dbpedia.org/resource/Barack_Obama> ;
  rdf:predicate <http://dbpedia.org/ontology/birthPlace> ;
  rdf:object <http://dbpedia.org/resource/Hawaii> ;
  vrank:hasRank [ vrank:rankValue "291.5535"^^xsd:float ] ] .
```

Literal values We plan to include literal values as descriptors of the entities. The selection of these values is planned to be based on predicate-statistics about the entity's RDF-type.

i18n and time One of our further contributions will be the exploitation and combination of browsing context for region, language, and timeline-focused summaries.

Data sources We are investigating on how to extend the summarization engine with further data sources such as Freebase and Wikidata.

Visualization and media The HTML output of the system is currently very basic. We plan to put significant effort into the design of a more appealing show case.

Evaluation We plan to extend our previous efforts [6] in designing evaluation scenarios for entity summarization.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 611346.

References

1. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of the 7th intl. conf. on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.
2. Aidan Hogan, Emir Munoz, and Jürgen Umbrich. Lodpeas: Like peas in a lod (cloud). In *Proceedings of the Billion Triple Challenge*, 2012.
3. Markus Krötzsch, Denny Vrandečić, and Max Völkel. Semantic mediawiki. In *The Semantic Web-ISWC 2006*, pages 935–942. Springer, 2006.
4. Alberto Musetti, Andrea Giovanni Nuzzolese, Francesco Draicchio, Valentina Presutti, Eva Blomqvist, Aldo Gangemi, and Paolo Ciancarini. Aemoo: Exploratory search based on knowledge patterns over the semantic web. In *Semantic Web Challenge*, 2012.
5. Antonio Roa-Valverde, Andreas Thalhammer, Ioan Toma, and Miguel-Angel Sicilia. Towards a formal model for sharing and reusing ranking computations. In *Proc. of the 6th Intl. Workshop on Ranking in Databases In conjunction with VLDB 2012*, 2012.
6. Andreas Thalhammer, Magnus Knuth, and Harald Sack. Evaluating entity summarization using a game-based ground truth. In *Intl. Semantic Web Conf. (2)*, volume 7650 of *Lecture Notes in Computer Science*, pages 350–361. Springer, 2012.