ARTICLE IN PRESS

Web Semantics: Science, Services and Agents on the World Wide Web ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

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

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

Effective searching of RDF knowledge graphs

Hiba Arnaout, Shady Elbassuoni *

Computer Science Department, American University of Beirut, Riad El-Solh 1107 2020, Beirut, Lebanon

ARTICLE INFO

Article history:
Received 4 June 2017
Received in revised form 24 October 2017
Accepted 4 December 2017
Available online xxxx

Keywords:
RDF
Ranking
Diversity
Relaxation

ABSTRACT

RDF knowledge graphs are typically searched using triple-pattern queries. Often, triple-pattern queries will return too many or too few results, making it difficult for users to find relevant answers to their information needs. To remedy this, we propose a general framework for effective searching of RDF knowledge graphs. Our framework extends both the searched knowledge graph and triple-pattern queries with keywords to allow users to form a wider range of queries. In addition, it provides result ranking based on statistical machine translation, and performs automatic query relaxation to improve query recall. Finally, we also define a notion of result diversity in the setting of RDF data and provide mechanisms to diversify RDF search results using Maximal Marginal Relevance. We evaluate the effectiveness of our retrieval framework using various carefully-designed user studies on DBpedia, a large and real-world RDF knowledge graph.

1. Introduction

In recent years, the Web has evolved from a network of linked documents to one where both documents and data are linked, resulting in what is commonly known as the Web of Data. Underpinning this evolution is a set of best practices known as Linked Data,[1] which provides mechanisms for publishing and connecting structured data on the Web in a machine-readable form with explicit semantics. The increasing adoption of Linked Data is turning the Web into a global data space that connects data from diverse domains and enables genuinely novel applications.

Recently, Linked Data has grown from an academic endeavor into one that has been embraced by numerous governments and industrial stakeholders, in domains such as business and finance, geography, governance, media, digital libraries, life sciences, in addition to general-purpose and user-generated content datasets. All such data is typically represented as sets of RDF (Resource Description Framework)[2] triples. An RDF triple is a triple with three fields: a *subject*, a *predicate* and an *object* where subjects and predicates are URIs and objects are either URIs or literals. Alternatively, an RDF dataset can be also viewed as a labeled graph, which we refer to as an RDF knowledge graph. In an RDF knowledge graph, node labels are *either* URIs representing resources *or* literals, and edge labels are URIs representing predicates. Fig. 1 shows an excerpt of DBpedia [1], an RDF knowledge graph that is automatically constructed from Wikipedia, and Table 1 shows the corresponding RDF triples. We omit the prefix of all URIs for readability.

The semantic query language for RDF is known as SPARQL.[3] SPARQL allows users to compose structured queries consisting of triple patterns, where a triple pattern is an RDF triple with one or more variables. A variable occurring in one of the triple patterns can be used again in another triple pattern in the query, denoting a join condition. For example, if the user's information need is to find movies that were directed and starred by the same person, the following triple-pattern query can be used:

```
?x director ?y
?x starring ?y
```

where `?x` and `?y` denote variables. Given this 2 triple-pattern query, the results are all the subgraphs of the underlying knowledge graph that are isomorphic to the query graph. For the above example, the result subgraphs should consist of two triples whose subjects and objects are the same, and whose predicates are `director` and `starring`, respectively. For instance, one such subgraph is:

```
Annie_Hall director Woody_Allen
Annie_Hall starring Woody_Allen
```

* Corresponding author.
E-mail addresses: hka22@aub.edu.lb (H. Arnaout), se58@aub.edu.lb (S. Elbassuoni).

[1] http://linkeddata.org/.

[2] http://www.w3.org/RDF/.

[3] https://www.w3.org/TR/rdf-sparql-query/.

https://doi.org/10.1016/j.websem.2017.12.001
1570-8268/

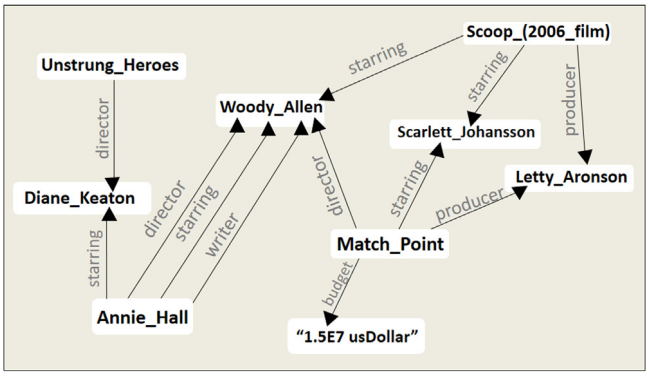

Fig. 1. An excerpt of DBpedia's RDF knowledge graph.

Table 1
The set of RDF triples corresponding to the RDF knowledge graph in Fig. 1.

Subject	Predicate	Object
Annie_Hall	starring	Woody_Allen
Annie_Hall	writer	Woody_Allen
Annie_Hall	director	Woody_Allen
Annie_Hall	starring	Diane_Keaton
Unstrung_Heroes	director	Diane_Keaton
Scoop_(2006_film)	starring	Woody_Allen
Scoop_(2006_film)	starring	Scarlett_Johansson
Scoop_(2006_film)	producer	Letty_Aronson
Match_Point	director	Woody_Allen
Match_Point	starring	Scarlett_Johansson
Match_Point	producer	Letty_Aronson
Match_Point	budget	1.5E7 usDollar

RDF knowledge graphs equipped with triple-pattern querying such as SPARQL querying provides a very powerful tool for knowledge discovery. However, in order to fully utilize RDF Knowledge graphs, the following challenges must be addressed.

1. **Data Incompleteness:** while large RDF knowledge graphs contain a vast amount of information in the form of RDF triples, the majority of information on the Web is available in the form of free text. Thus, extending RDF knowledge graphs with text can increase the scope of such knowledge graphs making them very rich sources of information. For example, the RDF knowledge graph in Fig. 1 contains information about movies, their directors, actors and producers. While this covers a wide range of interesting information, there is still information that cannot be easily captured in the form of RDF triples such as movie plots, taglines, users' comments and so on. Such information naturally appears as free text and by omitting them altogether, we lose a lot of valuable information.
2. **Flexible Querying:** even though triple-pattern queries are highly expressive, they are also very restrictive since they deploy Boolean matching (i.e., a result is either a match to a query or not). It is thus crucial to equip triple-pattern search with flexible querying mechanisms to allow for a more effective searching of RDF data. For example, consider an example query asking for movies that were directed and starred by Woody Allen. This query should be formulated as the following triple-pattern query when run against DBpedia:

```
?x director Woody_Allen
?x starring Woody_Allen
```

   However, if the user fails to express her query this way, say by not using the exact URI referring to Woody Allen or by using the URI `Woody_Allen` as the subject of the triple patterns rather than their object, the query will not yield any result. In such cases, query relaxation, which is similar to query modification in traditional Information Retrieval (IR), is crucial to address queries that cannot be answered as exactly formulated. On the other hand, some information needs cannot be entirely expressed using triple-pattern queries due to the lack of appropriate resources in the knowledge graph. For example, assume that the user is interested in finding movies that are based on a novel, or that have won an Oscar. It is not possible for the user to express such information need using a triple-pattern query only. However, if RDF knowledge graphs were extended with text, and keyword conditions were allowed in queries, this can go a long way in addressing a wider range of information needs such as the ones just mentioned.
3. **Result Ranking:** RDF knowledge graphs may produce too many results for many queries. It will thus be highly beneficial to present users with a ranked list of results rather than a mere list of unranked ones. This is particularly crucial when RDF knowledge graphs and triple-pattern queries are extended with text and when query relaxation is deployed. To this end, some notion of relevance or importance must be defined over the RDF knowledge graph that can be utilized for result ranking. For example, consider the information need of finding movies that were directed and starred by the same person. It might be highly desirable to rank the results based on, say, movie popularity. If the information need was to find those movies that also won an Oscar or were based on a novel, these should constitute additional criteria for result ranking.
4. **Result Diversity:** while ranking ensures that the most relevant results are ranked on top, it is often the case that the top results tend to be homogeneous, making it difficult for users to truly explore the knowledge graph. For example, considering our example query, it might be undesirable to have all the top results about movies directed and starred by one person, say Woody Allen, or all about movies of the same genre. Result diversity can thus play a big role in ensuring that the users get a broad view of the different aspects of the results of their queries, and ensures that, on average, almost all users can find relevant results to their queries in the top ranks.

In this paper, we address all the above challenges and develop a number of novel models and algorithms to effectively search RDF knowledge graphs. Our contributions can be summarized as follows.

- We develop a novel ranking model based on statistical machine translation for triple-pattern queries. Our ranking model operates on top of traditional RDF knowledge graphs as well as on keyword-extended ones that allow keyword conditions in triple-pattern queries.
- We develop a framework for query relaxation that automatically relaxes triple-pattern queries that yield no results in such a way that the original query intention is preserved. We use the constants provided by the user to augment the relaxed queries, by turning them into a set of keywords. This has the advantage of improving the recall of such queries without unduly sacrificing precision. We also develop a principled mechanism to combine the results of relaxed queries using our ranking model.
- We define a notion of result diversity in the RDF setting and develop a general technique based on the Maximal Marginal Relevance [2] in order to provide diverse results to queries over RDF knowledge graphs.

Please cite this article in press as: H. Arnaout, S. Elbassuoni, Effective searching of RDF knowledge graphs, Web Semantics: Science, Services and Agents on the World Wide Web (2017), https://doi.org/10.1016/j.websem.2017.12.001.