Towards a Ranking Model for Semantic Layers over Digital Archives

Pavlos Fafalios, Vaibhav Kasturia, Wolfgang Nejdl
L3S Research Center, Hannover, Germany
{fafalios, kasturia, nejdl}@l3s.de

ABSTRACT

Archived collections of documents (like newspaper archives) serve as important information sources for historians, journalists, sociologists and other interested parties. Semantic Layers over such digital archives allow describing and publishing metadata and semantic information about the archived documents in a standard format (RDF), which in turn can be queried through a structured query language (e.g., SPARQL). This enables to run advanced queries by combining metadata of the documents (like *publication date*) and content-based semantic information (like *entities* mentioned in the documents). However, the results returned by structured queries can be numerous and also they all equally match the query. Thus, there is the need to rank these results in order to promote the most important ones. In this paper, we focus on this problem and propose a ranking model that considers and combines: i) the relativeness of documents to entities, ii) the timeliness of documents, and iii) the relations among the entities.

1. INTRODUCTION

Despite the increasing number of digital archives worldwide (like newspaper and web archives), the absence of efficient and meaningful exploration methods still remains a major obstacle in the way of turning them into a usable source of information [1]. Semantic models try to solve this problem by offering a vocabulary for describing and publishing in the standard RDF format, metadata (e.g., publication date) and semantic (e.g., mentioned entities) information about a collection of archived documents. The produced *Semantic Layers* allow running advanced *entity-centric* queries requesting complex information related to some entities, concepts or events and to some specific metadata values [2]. As an example, we can access a Semantic Layer over a newspaper archive and find articles of a specific time period discussing about a specific category of entities (e.g., *philanthropists*) or about entities sharing some characteristics (e.g., *lawyers born in Germany*). Such advanced information needs can be directly expressed through SPARQL queries (unfriendly for end-users) or through a user-friendly interactive interface which transparently transforms user interactions to SPARQL queries (e.g., a faceted browsing interface [4]). However, the results returned by such queries can be numerous and moreover they all equally match the query. Thus, there arises the need to rank them for discovering and returning to the user the most important ones. An effective ranking method should consider the different factors that affect the importance of a document to the information need, relying at the same time only on the data available in the semantic layer (without accessing documents' full contents).

In this paper, we focus on this problem and propose a model for ranking archived documents returned by a structured query over a semantic layer. The proposed model jointly considers the following aspects: i) the *relativeness* of a document with respect to the entities of interest, ii) the *timeliness* of document's publication date, iii) the temporal *relatedness* of the entities of interest with other entities mentioned in the document. The idea is to promote documents that mention the entities of interest many times, that have been published in important (for the entities of interest) time periods, and that mention many other entities co-occurring frequently with the entities of interest in important time periods. For example, in case we want to rank articles of 1990 discussing about *Nelson Mandela*, we want to favor articles that i) mention *Nelson Mandela* multiple times in their text, ii) have been published in important time periods for *Nelson Mandela* (e.g., February 1990 since during that period he was released from prison), and iii) mention other entities that seem to be important for *Nelson Mandela* during important time periods (e.g., *Frederik Willem de Klerk* who was South Africa's State President in February 1990).

2. RANKING MODEL

2.1 Problem Definition

In our problem, an *entity* is anything with a distinct and meaningful existence that also has an "identity" expressed through a unique ID (e.g., a Wikipedia URI). This does not only include persons, locations, etc., but also concepts (e.g., *democracy*) and events (e.g., *2010 Haiti earthquake*).

Given a collection of archived documents D, a set of entities E_D mentioned in documents of D, and a SPARQL query Q requesting documents from D published within a set of *time periods* P_Q of a fixed granularity Δ (e.g., day, week, etc.) and related to one or more *Entities of Interest*

© 2017 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

(EoI) $E_Q \subseteq E_D$ with logical AND (mentioning all EoI) or OR (mentioning at least one EoI) semantics, the problem is how to rank the documents $D_Q \subseteq D$ that (equally) match Q.

Figure 1 shows an example SPARQL query requesting articles from a newspaper archive, published in 1990 and discussing about the entities *Nelson Mandela* and *Frederik Willem de Klerk* (AND semantics), while the query in Figure 2 requests articles of 1990 mentioning *state presidents of South Africa* (OR semantics). Our objective is to rank the documents returned by such SPARQL queries.

```
SELECT DISTINCT ?article WHERE {
  ?article dc:date ?date FILTER(year(?date) = 1990) .
  ?article oae:mentions ?entity1, ?entity2 .
  ?entity1 oae:hasMatchedURI  dbr:Nelson_Mandela .
  ?entity2 oae:hasMatchedURI  dbr:F._W._de_Klerk }
```

Figure 1: Query requesting articles of 1990 mentioning *Nelson Mandela* and *Frederik Willem de Klerk* (AND semantics).

```
SELECT DISTINCT ?article WHERE {
  ?article dc:date ?date FILTER(year(?date) = 1990) .
  ?article oae:mentions ?entity .
  ?entity oae:hasMatchedURI  ?entityURI .
  ?entityURI dc:subject dbc:State_Presidents_of_South_Africa }
```

Figure 2: Query requesting articles of 1990 discussing about *state presidents of South Africa* (OR semantics).

2.2 Modeling

Relativeness. We consider that if the EoI are mentioned in a document many times, the document should receive a high score since its topic may be about these entities. The term frequency (in our case *entity frequency*) is a classic numerical statistic that is intended to reflect how important a word is to a document [3].

For the case of AND semantics ("$\wedge$"), the *relativeness* score of a document $d \in D_Q$ can be simply defined as:

$$ScoreD_{\wedge}(d) = \frac{\sum_{e \in E_Q} count(e,d)}{\sum_{e' \in E_d} count(e',d)} \tag{1}$$

where $E_d \subseteq E_D$ is the set of entities mentioned in d and $count(e,d)$ is the number of occurrences of an entity e in d.

For the case of OR semantics ("$\vee$"), we can also consider the number of different EoI mentioned in the document (since a document does not probably contain all the EoI as in the case of AND semantics). In that case, the *relativeness* score of a document $d \in D_Q$ can be defined as follows:

$$ScoreD_{\vee}(d) = \frac{\sum_{e \in E_Q} count(e,d)}{\sum_{e' \in E_d} count(e',d)} \cdot \frac{|E_d \cap E_Q|}{|E_Q|} \tag{2}$$

This formula favors documents mentioning multiple times many of the EoI.

Timeliness. A time period of granularity Δ can be considered important for the EoI, if there is a relatively large number of documents mentioning the EoI during that period. For a time period $p \in P_Q$, we consider the following *timeliness* score:

$$ScoreP(p) = \frac{|D_p \cap D_Q|}{|D_Q|} \tag{3}$$

where $D_p \subseteq D$ is the set of documents published during p.

Relatedness. Entities that are co-mentioned frequently with the EoI in important time periods are probably important for the EoI. For example, *Apartheid* was an important concept related to *Nelson Mandela* during 1990. Thus, articles discussing for both *Apartheid* and *Nelson Mandela* should be promoted. However, there may be also some general entities (e.g., *South Africa* in our example) that co-occur with the EoI in almost all documents (independently of the time period). Thus, we should also avoid overemphasizing documents mentioning such "common" entities. First, we consider the following *relatedness* score of an entity $e \in E_D \setminus E_Q$:

$$\begin{aligned} ScoreE(e) &= idf(e) \cdot \sum_{p \in P_Q} \left(ScoreP(p) \cdot \frac{|D_{e,p} \cap D_Q|}{|D_p \cap D_Q|}\right) \\ &= idf(e) \cdot \sum_{p \in P_Q} \frac{|D_{e,p} \cap D_Q|}{|D_Q|} \end{aligned} \tag{4}$$

where $D_{e,p} \subseteq D$ is the set of documents mentioning e and published in the time period p, and $idf(e)$ is the inverse document frequency of e, defined as:

$$idf(e) = 1 - \frac{|D_e \cap (\cup_{e' \in E_Q} D_{e'})|}{|\cup_{e' \in E_Q} D_{e'}|} \tag{5}$$

where $D_e \subseteq D$ denotes the set of documents mentioning e.

This *relatedness* formula considers the percentage of documents in which the entity co-occurs with the EoI in important time periods.

Joining the Models. We can now join the above models and derive a final score for a returned document $d \in D_Q$:

$$S(d) = ScoreP(p_d) \cdot ScoreD(d) + \beta \frac{\sum_{e \in E_d \setminus E_Q} ScoreE(e)}{|E_d|} \tag{6}$$

where $p_d \in P_Q$ is the time period in which d was published, and β is a decay factor for controlling the effect of relatedness.

3. CONCLUSION

We have introduced a model for ranking documents returned by querying a semantic layer over an entity-annotated archived collection of documents. An important characteristic of our approach is that it only exploits the data of the semantic layer (i.e., its RDF triples) and thereby it can be directly applied even at query-execution time. In future, we will extensively evaluate the proposed model and the effect of each of its components. We also plan to investigate how this model can be applied in web archives, where the publication date is not usually available.

Acknowledgements

The work was partially funded by the European Commission for the ERC Advanced Grant ALEXANDRIA (No. 339233).

4. REFERENCES

[1] K. Calhoun. *Exploring digital libraries: foundations, practice, prospects.* Facet Publishing, 2014.

[2] P. Fafalios, H. Holzmann, V. Kasturia, and W. Nejdl. Building and Querying Semantic Layers for Web Archives. In *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'17)*, Toronto, Ontario, Canada, 2017.

[3] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of massive datasets.* Cambridge University Press, 2014.

[4] Y. Tzitzikas, N. Manolis, and P. Papadakos. Faceted exploration of rdf/s datasets: a survey. *Journal of Intelligent Information Systems*, pages 1–36, 2016.