On the Use of Semantic Knowledge Bases for Temporally-aware Entity Retrieval

Krisztian Balog
Krisztian.Balog@idi.ntnu.no

Kjetil Nørvåg
Kjetil.Norvag@idi.ntnu.no

Department of Computer and Information Science
Norwegian University of Science and Technology
Sem Sælands vei 7-9, Trondheim, Norway

ABSTRACT

In this paper, we propose the development of entity retrieval models that are *temporally-aware*. Using a classification scheme of entity-oriented information needs proposed in prior work, we discuss how the addition of temporal expressions would affect retrieval results for each of these query types. We describe the design of a retrieval framework that is capable of catering for a range of these information needs and identify main challenges at the component level.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

Keywords: Entity search, temporal information retrieval, semantic search.

1. INTRODUCTION

Entity search and temporal information retrieval have emerged in recent years as important research topics. Both are representatives of efforts to move from traditional bag-of-words representation of documents to semantically more informed information access systems. They are orthogonal to each other, as they rest on fundamentally different conceptions of purpose, but they are not unrelated. In fact, the two are complementary and essential parts of intelligent information access systems. Our research interests lie in the intersection of the above two topics; specifically, we investigate how entity retrieval could be made temporally-aware, using semantic knowledge bases (KB) enriched with temporal information.

2. BACKGROUND

Existing research in Information Retrieval has mostly viewed entities as static objects and abstracted away from the temporal dimension. While this is a reasonable simplification for a number of application scenarios, i.e., when the user is only interested in the current state of things, it was also, in fact, necessitated by two reasons: (i) making entities findable by means of keyword search was challenging enough in itself, let alone incorporating time, and (ii) temporal information was not readily available and extracting it would have increased the complexity of retrieval systems to a level that was prohibitive at that time. Recent years have witnessed significant progress on both aspects. Entity search has moved from bag-of-words models built from text usage around entity mentions to semantically more informed representations that incorporate structured data sources [7]. Also, increasingly complex information needs are being considered, as can be witnessed by the tasks featured at various benchmarking evaluation campaigns. As for (ii), a very important development was the introduction of YAGO2, a knowledge base created from Wikipedia, GeoNames, and WordNet (containing 447 million facts about 9.8 million entities), in which entities and facts are anchored in time and space [4].

Temporal IR so far has mostly been focused on providing access to time-dependent documents, where time is either publication time or content time (temporal expressions mentioned in documents) [3, 5]. There is limited work in the intersection of entity search and temporal IR, and most of it is focused on systems that offer entity-oriented access to news collections, see, e.g., [1, 2].

3. QUERY TYPES

The task we address is *ad-hoc entity retrieval* (sometimes referred to as *semantic search*): "answering arbitrary information needs related to particular aspects of objects [entities], expressed in unconstrained natural language and resolved using a collection of structured data" [8]. Following the classification scheme introduced in [8], we consider four query types. Additionally, we introduce a fifth category, which was not considered as a separate class in [8] (presumably because these type of queries did not have a strong presence in the query logs). For each query class, we discuss its extension along the temporal dimension.

- **Entity query**: The aim is to find a particular entity. A temporal expression might be added to the query with the intent to narrow or filter the results. E.g., "08 toyota tundra," where 08 refers to the model year 2008. Nearly all types of entities can be annotated with some temporal attributes, yet, using these as a means of disambiguation would be unnatural for many of them; consider, for example "windsor hotel philadelphia."
- **Type query**: The user wants to retrieve entities of a particular type or class. Again, temporal expressions could be used as a means of filtering (e.g., "composers of the 18th century," or "new olympic sports introduced in 2012").
- **Attribute query**: The intention of the query is to find the value of a particular attribute of some entity or type. Many attributes can change over time. Unless explicitly stated, the user is most likely interested in the current information (e.g., "population of New York") and would explicitly state otherwise (e.g., by adding "in 2010" to the previous query).
- **Relation query**: The goal is to find out and describe the relationship between two or more entities. Relationships, indeed, have a strong temporal dependence. For example, the relation of "Tom Cruise" to "Nicole Kidman" was husband between 1990 and 2001 and since 2001 it is ex-husband. Additionally, they co-starred in a number of movies.

ESAIR'12, November 2, 2012, Maui, Hawaii, USA.
ACM 978-1-4503-1717-7/12/11.

- **Complex query**: The target is a list of entities that stand in some required relation with other entities. We consider them "complex," as today's web search engines do not yet provide sufficient support for these type of requests; typically, the underlying information need cannot be answered with a single one-off query (e.g., "albums released by Leonard Cohen after he wrote Suzanne").[1]

To remain focused, in the remainder of this paper we restrict ourselves to queries where the desired unit of retrieval is entities and results are presented as a ranked list; this entails entity queries, (a subset of) type queries, and complex queries.

4. APPROACH

Our goal is to answer information needs, expressed as unconstrained natural language queries, using a collection of temporally enhanced structured data. This will involve dealing with three main questions concerning (i) the representation of entities, (ii) the representation of information needs, and (iii) the development of a retrieval model that matches these two and computes a relevance score. We briefly examine these three issues in turn.

Representing entities. We assume that entities are RDF resources, described in the form of subject-predicate-object (SPO) triples. Entities have a unique identifier (URI, i.e., the subject), attributes (where the object is a literal), and typed relationships with other entities (where the object is a URI). This defines a graph where nodes are either entities (URIs) or attributes, and edges represent typed relationships between them. We add a temporal dimension on top of this basic layer of representation, following the principles laid out in YAGO2 [4]. Temporal existence t is defined in terms of start and end times: $t = [t_b, t_e]$. (Note that t_b or t_e might not be set if there is no corresponding temporal information available in the knowledge base.) Each entity identifier is associated with a timestamp: (ID, t). In YAGO2, timestamps are defined for four main entity types: people, groups, artifacts, events. For other types of entities existence time would not be meaningful, therefore t is not set. Relationships too might have a temporal existence, defined in terms of start and end times. We capture this by extending SPO triples with temporal information: (s, p, o, t). Again, t is not necessarily set for all relations; for example, the type of the entity is permanent and do not change over time.

The main challenges within this component concern the completeness and correctness of the underlying knowledge base. YAGO2 has a human confirmed accuracy of 95%. Completeness cannot be measured, but given that no KB can ever be complete, it is safe to say that automatic means of populating the KB are desired. Kuzey and Weikum [6] provide an example of efforts in that direction, where temporal facts and events are harvested from the textual contents of Wikipedia articles (as opposed to infoboxes).

Representing information needs. Our goal with this component is to understand and create an explicit representation of the information need that is expressed in the query. We represent the acquired interpretation of the query as a set of constraints on nodes and edges of the knowledge graph. Existing text-based approaches to entity retrieval consider two types of constraints: (i) terms or phrases matching one or more attributes of the entity (e.g., the `dbbprop:name` predicate should contain "New York"), where one might consider different types of string matching (exact, loose, numeric, etc.), and (ii) relationship with another entity (e.g., the `dbpedia-owl:birthPlace` predicate should match `dbpedia:New_York`). The temporal dimension gives rise to two additional constraints: (i) existence of the entity, and (ii) existence of the relation. We can consider various operators on timestamps and time-spans, such as IN, $BEFORE$, or $AFTER$.

Deriving the interpretation of the query is a challenging research problem that is far from being solved even for regular document search. Matters are further complicated when temporal expressions come into play. The first difficulty arises in relation to the detection and resolution of temporal expressions; there exist tools, such as TARSQI [9], for annotating documents, but it remains to be seen how well these would perform on short queries that lack proper grammar and context. Second, we need to perform reasoning, i.e., associating temporal expressions with entities or relationships. Third, we might want to include temporal constraints, even if not explicitly stated in the query. E.g., for "formula 1 team principals," the user is most likely interested in people who currently hold this position. For other queries, like "olympic sports" it is not that obvious whether such constraint needs to be imposed or not.

Matching entities and information needs. Because of our desire for graded relevance, as opposed to binary matches, we advocate the use of IR-style ranking as opposed to SPARQL-like querying. Language modeling techniques provide a common ground and a theoretically sound framework to combine existing research on entity retrieval in RDF data [7] with models on temporal IR that consider the inherent uncertainty of temporal expressions [3].

5. CONCLUSIONS AND FUTURE WORK

We addressed the problem of extending ad-hoc entity retrieval with a temporal dimension. We considered five types of entity-oriented queries and discussed what the addition of temporal expressions would entail for each of these. In doing so, we followed a rational and obvious line of thinking; however, an in-depth analysis of search and usage logs of an actual web search engine would be needed in order to identify the frequency of these cases (and perhaps of additional ones) occurring. Further, we presented a high-level design of a temporally-aware entity retrieval framework and identified the main challenges concerning each of its components. Our next step is to head towards the realization of this system.

References

[1] O. Alonso, K. Berberich, S. Bedathur, and G. Weikum. Time-based exploration of news archives. In *Proceedings of HCIR'10*, 2010.

[2] K. Balog, M. de Rijke, R. Franz, H. Peetz, B. Brinkman, I. Johgi, and M. Hirschel. Sahara: Discovering entity-topic associations in online news. In *Proceedings of ISWC'09*, 2009.

[3] K. Berberich, S. Bedathur, O. Alonso, and G. Weikum. A language modeling approach for temporal information needs. In *Proceedings of ECIR'10*, 2010.

[4] J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum. YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of WWW'11*, 2011.

[5] N. Kanhabua and K. Nørvåg. Determining time of queries for re-ranking search results. In *Proceedings of ECDL'10*, 2010.

[6] E. Kuzey and G. Weikum. Extraction of temporal facts and events from wikipedia. In *Proceedings of TempWeb'12*, 2012.

[7] R. Neumayer, K. Balog, and K. Nørvåg. On the modeling of entities for ad-hoc entity search in the web of data. In *ECIR'12*, 2012.

[8] J. Pound, P. Mika, and H. Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of WWW'10*, 2010.

[9] M. Verhagen, I. Mani, R. Sauri, R. Knippen, S. B. Jang, J. Littman, A. Rumshisky, J. Phillips, and J. Pustejovsky. Automating temporal annotation with TARSQI. In *Proceedings of ACL'05*, 2005.

[1] The types of queries studied within the List Completion task at the INEX Entity Ranking track, the Related Entity Finding task at the TREC Entity track, and the List Search task at the Semantic Search Challenge are all variations of some sort on this problem.