Events Retrieval Using Enhanced Semantic Web Knowledge

Pierre-Yves Vandenbussche, Charles Teissèdre

Events Retrieval Using Enhanced Semantic Web Knowledge

Pierre-Yves Vandenbussche[1,2] and Charles Teissèdre[1,3]

[1] Mondeca, 3, cité Nollez, 75018, Paris, France
[2] INSERM UMRS 872 ÉQ.20, Ingénierie des Connaissances en Santé, Paris, France
[3] MoDyCo – UMR 7114 CNRS / Université Paris Ouest Nanterre La Défense, 200 av. De la République, 92001 Nanterre, France
{firstname.lastname}@mondeca.com

Abstract. In this article, we present an experimental end user application to query DeRiVE 2011 challenge dataset in an innovative and intuitive manner. After enriching the dataset with external sources of information, it is indexed in a way that enables users to submit queries combining keywords, location and temporal anchor, in a single search field. The goal is to ease event retrieval providing a simple user interface to query and visualize events over time.

Keywords: Events Retrieval, Semantic Web, Data Mashup

1 Introduction

While geolocation services have enjoyed strong progress, few initiatives take into consideration chrono-localization and temporal query processing for Information Retrieval over the Web [1]. As Linked Open Data grows, things are changing, since more and more temporally anchored data is available. However, processing temporal data remains a challenge from (i) a modeling point of view, (ii) for data acquisition, (iii) as well as in terms of querying and navigating through it.

In this article, we address the last issue of querying and navigating through temporal data. We describe a system using the RDF data provided along with the DeRiVE 2011 challenge[1]. The dataset describes entertainment events related to music, such as advertisements for concerts or festivals. It also provides some information about agents involved in these events and about their location. The main objective of the system we present here is to hide data complexity and make it simple to query, providing a single search field as a first step in events retrieval. The goal is to make DeRiVE dataset temporally browsable. The considered use case consists in finding events occurring at a given period of time at a specific location.

After a brief overview of how temporal information is handled in the context of Information Retrieval over the Web, we will describe the way we processed the dataset to enrich it with external sources of information and to index it. We will then describe the final application to query and browse the dataset.

[1] Dataset is available at: http://semanticweb.cs.vu.nl/derive2011/Challenge.html

2 Temporal Information Retrieval over the Web

Retrieving temporal information over the Web of Content (*i.e.* HTML-based Web) and in the Web of Data (aka the Semantic Web) are two different issues, though they may converge on some points.

Temporal Search within the Web of Content. Major search engines currently offer few temporal search services. One such service is Google timeline feature[2], which offers a way to visualize keywords frequency at different periods of time and to browse sentences where these keywords are associated with a date. However, temporal expressions are reduced to point in time with no duration extent, hence there is an important loss of information. Processing temporal information expressed in Web documents is a challenge from at least three different points of view: (i) modeling temporal references (models should be able to represent dates and intervals, but may also need to cope with approximate information (*e.g.* "*by the end of the 13th century*"), iterative occurrences (*e.g.* "*every day from 10am to 8pm*"), as well as deictics (*e.g.* "*yesterday*", "*two months ago*") and anaphorics (*e.g.* "*the day before*")); (ii) document annotation (it requires processing huge amount of documents with NLP techniques that necessarily have to deal with imperfect precision and recall rate) and (iii) relevancy ordering of the results from the temporal perspective (how to rank documents by relevance from the temporal perspective?).

Temporal Search within the Web of Data. While the modeling issue remains a difficulty, the acquisition process in this context is quite different, since the data to process is structured. Data acquisition however can be an issue as well. As for the querying process, the main querying language, SPARQL, allows filtering results in a timespan (*i.e.* intervals of well defined dates). This approach explains why generally only well defined temporal properties are effectively employed in LOD[3].

The three Web sites that provided data for the challenge relies on this process: Upcoming Yahoo!, Last.fm and Eventful all propose similar approach to event retrieval. The main search scenario, with little variation depending on the Web site, follows this path: user has to provide a location, then a type of event (concert/festival), then eventually a musical genre, a date filter, etc. Such rich faceted search scenario is not made possible, though, with the DeRiVE challenge dataset, since no information on the type of event is provided.

3 Processing DeRiVE 2011 Dataset

The application we present here is an experimental retrieval engine with the goal to query and browse events temporally in the simplest way possible. It can be used both

[2] URL for the query "revolution": http://bit.ly/relfGV

[3] Despite Time Ontology [2] capability to describe complex time knowledge representation, it is generally not used in all its' complexity.

in the context of the Web of Content [3] and the Web of Data, as it relies on indexing process and NLP resources for temporal references extraction which can analyze either a query or Web documents. For the DeRiVE challenge, in order to get enough information to enable users to submit queries combining keywords, location and temporal information, we first had to enrich the dataset. The DeRiVE 2011 dataset is composed of 107.874 events and related knowledge. Knowledge is originating from Upcoming Yahoo! (12.15%), LastFm (53.04%) and Eventful (34.81%). It has been transformed by EventMedia [4]. The dataset is made of more than 1.800.000 statements. Temporal information consists in either single dates or intervals of dates.

Event geo-location augmentation. 98.794 events (91.58%) have latitude and longitude information. The first knowledge augmentation process concerns events' geolocation. It tries to fetch city, country and address information from coordinates, using Google and Yahoo! reverse geocoding API. In our application this geolocation information is used during query processing to cope with countries or cities. It is also used to propose a map visualization using Google maps API.

Event Image augmentation. Images provide a simple way to ensure a pleasant way to experience event browsing. To associate images to events, we set up a strategy based on images information in the Semantic Web (via SPARQL queries on EventMedia and DBpedia) and on the Web (via Flickr API). As a result, at least one image was associated to 95.01% of events. SPARQL query example on EventMedia endpoint[4] using event URI:

```
SELECT distinct ?imageURI ?image
WHERE{
      ?imageURI                    <http://linkedevents.org/ontology/illustrate>
      <http://data.linkedevents.org/event/dba9e034-fea0-4d01-ba4c-
      fb0515b89051>.
      {?imageURI <http://www.w3.org/ns/ma-ontlocator> ?image. }
      UNION{?imageURI <http://www.w3.org/ns/ma-ont#locator> ?image. }
}
```

Agent Information augmentation. Information about agents involved in an event is valuable for our application users. By enriching the dataset with Wikipedia links that point toward articles concerning these agents, users can further their search. We collected these links thanks to SPARQL queries on DBpedia endpoint. We have been able to find Wikipedia links for 25.22% of the agents. SPARQL query example on DBpedia endpoint[5] using agent label:

```
SELECT distinct ?wikiLink
WHERE {
      {?s <rdfs:label> "Bob Dylan"@en.}
      UNION{?s <http://xmlns.com/foaf/0.1/name>
      "Bob Dylan"@en.}
      {?s a <http://dbpedia.org/ontology/Person>.}
      UNION{?s a <http://dbpedia.org/ontology/Band>.}
      ?s <http://xmlns.com/foaf/0.1/page> ?wikiLink.
}
```

[4] URL: http://semantics.eurecom.fr/sparql
[5] URL: http://dbpedia.org/sparql

4 An Experimental Temporal Search Engine to Retrieve Events

The system we have implemented is both a search engine and a tool to visualize and browse events[6]. Temporal query relies on the search engine developed by [3]. The search engine is able to process queries with approximate temporal conditions like "*around May 2007"*, even if this temporal expression does not exist in DeRiVE data. From the temporal perspective, event retrieval is based on an algorithm that calculates similarity scores between the temporal reference of the query and those that are associated to events. Based on Lucene and several modules to compute the dataset (see fig 1), the system can handle queries that may combine keywords, location and temporal information, such as "*rock in London in August 2008*" or "*Bonn by the end of 2007*". Temporal information, location information and event or agent description are indexed as different fields once the dataset is fully preprocessed.

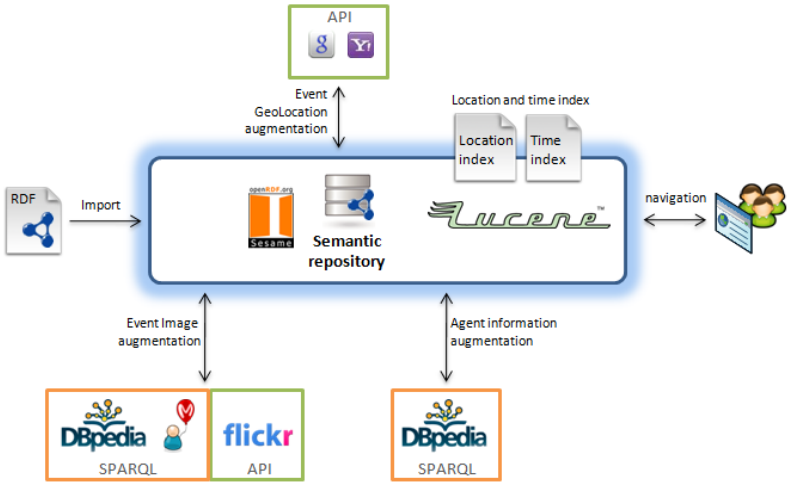

Fig. 1. System's architecture.

Queries are analyzed in such way that keywords, dates and location information are separated. Temporal data recognition in queries is performed thanks to an NLP module described in [5]. The location extraction is performed thanks to a dictionary built during the indexing process: the dictionary contains cities and countries entities collected during the event geolocation enrichment process. Any other information that may appear in a query is considered as simple keywords, on which no semantic analysis is performed.

The events returned by the system are presented on a SIMILE timeline[7] (see fig 2). The timeline on which results are displayed is fully browsable, which means that users can move over time: the system generates new queries on the fly as users move forward or backward in time.

[6] The system can be tested at the following address:
http://labs.mondeca.com/ChallengingTime/?locale=en&demo=eventMedia
[7] URL: http://www.simile-widgets.org/timeline/

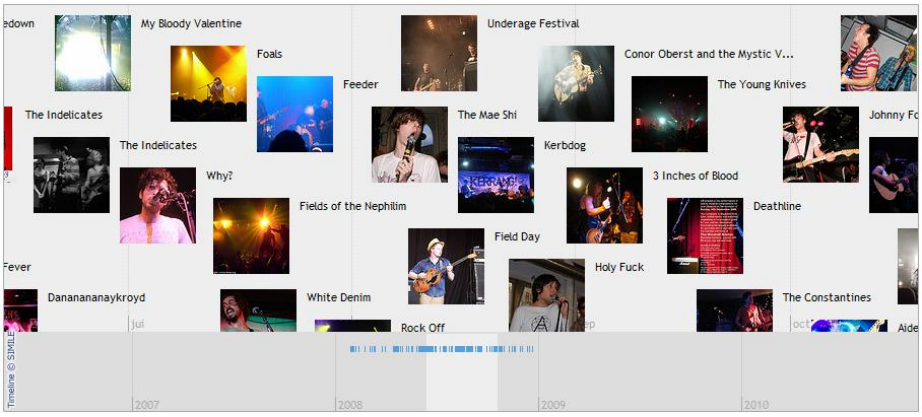

Fig. 2. Screenshot of the UI for the query "*rock in London in August 2008*".

5 Conclusion and Perspectives

The experimental application presented could be used as first step in events retrieval. Since the approach is generic and not tightly bound to DeRiVE dataset, it can be used in any other use case scenario where data is temporally anchored. If the dataset had contained information about musical genres, it could have been interesting to introduce faceted search with SolR tool, so that users could refine the results and eventually disambiguate query. Another interesting feature for possible improvement would be to synchronize a map for geolocation with the timeline, so as to present the results both in their temporal and geographic context.

Acknowledgments. This project is partially granted by Datalift ANR project (ANR-10-CORD-009) and Chronolines ANR project (ANR-10-CORD-010).

References

1. Alonso, O.; Gertz, M. & Baeza-Yates, R.: On the Value of Temporal Information in Information Retrieval. Proc. of ACM SIGIR Forum 41, no. 2 (December), 35-41 (2007)
2. Hobbs, JR & Pan, F.: An ontology of time for the semantic web.Proc. of ACM Transactions on Asian Language 3, no. 1 (March), 66-85 (2004)
3. Teissèdre, C; Battistelli, D. & Minel, J.-L.: Recherche d'information et temps linguistique : une heuristique pour calculer la pertinence des expressions calendaires. Proc. of TALN 2011, Montpellier (2011)
4. Troncy, R.; Malocha, B. & Fialho, A. Linking events with media Proceedings of the 6th International Conference on Semantic Systems, 1-4 (2010)
5. Teissèdre, C.; Battistelli, D. & Minel, J.-L.: Resources for Calendar Expressions Semantic Tagging and Temporal Navigation through Texts. Proc. of LREC 10, Malta (2010)