Multimedia information retrieval as a practical application for interlinking approaches

Antonio J. Roa-Valverde
playence KG
Innsbruck, Austria
antonio.roa@playence.com

ABSTRACT

This work introduces playence media, a tool for the annotation, search and navigation of media assets, and describes how interlinking approaches can be applied to perform multimedia information retrieval.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Algorithms

Keywords

Multimedia information retrieval, LOD, interlinking

1. INTRODUCTION

There are several works in the literature considering different approaches for interlinking datasets in the Linked Open Data (LOD) cloud [3]. In [1], authors state the advantages of having a completely interlinked network in order to better exploit the knowledge available. Even though the benefits of having a full interlinked data network are not arguable, there is still a lack of direct applications for interlinking techniques that help the end user to understand the benefits of applying them.

In this paper, we focus on the application of LOD principles to multimedia information retrieval. Multimedia content opens new perspectives regarding to integration and usage requirements of the information. In this way, dealing with multimedia content is still an open problem. For instance, main search engines are still blind when looking for information inside video or audio content. The main problem when dealing with media assets is the access to the information represented by the media content. Common tasks like searching, navigating and filtering information can be very hard when it is provided as image, video or audio format. Another problem to be tackled is that media content is not used in isolation, but together with the information available in other heterogeneous data sources.

Following this topic, in section 2 we present playence media, a suite for the annotation, search and navigation of media assets. We briefly introduce the tool and show how LOD is exploited in two different use cases. In section 3 we describe the interlinking approaches that guide the functionality mentioned in the use cases. Finally, section 4 summarizes the contribution of this work and states further directions for its improvement.

ISemantics 2011 7th Int. Conf. on Semantic Systems, Sept. 7-9, 2011, Graz, Austria

2. PLAYENCE MEDIA AND THE ROLE OF INTERLINKING

playence Media [1] is a suite for the management, annotation and retrieval of multimedia assets. It gathers approaches from audio and video analysis, image processing, text engineering and knowledge modeling to provide a solution for multimedia content annotation. The annotations are created using different methodologies that range from manual to semi-automatic mechanisms. Independently of the method, all these techniques have in common the use of a domain ontology that models every specific use case and helps to formalize the content available within the assets.

The main functionality provided by playence media focuses on searching for the relevant part of a media asset. E.g., *find the video sequences where Pau Gasol appears scoring when playing with Los Angeles Lakers.* For doing so, search relies on previous annotations of media assets. The annotation is a compulsory step before any search can be performed. Therefore, the quality of annotations determines the accuracy of search results.

Behind the big scene of annotation and search, interlinking approaches can be put in practice. In this scope, the aim of using interlinking is twofold. First, it helps to relate the knowledge represented in the domain ontology with external sources, complementing the existent annotations and facilitating the retrieval of the asset on query time, for instance using query expansion. Second, the new links created can be useful to retrieve more information related with the content depicted in the asset. This is known as content augmenta-

[1]A demo is available at http://demos.playence.com:8080/media/ . User: isem11. Password:isem11. A complete guideline can be found at www.playence.com after registration.

tion and it is used during the presentation of results. In the following, we describe both applications in more detail.

In the next sections we consider the DBpedia[2] dataset and the sport ontology developed by playence. Although the ontology is not very complex, it contains more than 10000 instances. Those instances have been extracted from freebase[3], allowing us to have an extensive knowledge base without putting so much effort on its construction.

2.1 Knowledge Expansion

The benefit of using LOD is on the fact of browsing heterogeneous data sets as a whole. For example, we could start by getting the list of players of LA Lakers, then continue by getting specific information about Pau Gasol and finally, obtaining information about the city of birthplace of the same player.

By interlinking our domain ontology to external LOD sets we are expanding the knowledge available in playence media, so that it can be incorporated directly into the search process. This is represented in figure 1. The figure depicts an excerpt of the sports ontology, more precisely the properties related to the concept *Player*. As it is observed, interlinking this concept with the homologue concept in the DBpedia ontology allows the use of properties that previously were not available in the sports ontology.

A possible way for exploiting the internal knowledge is through query expansion mechanisms. Query expansion can be defined as the process of enriching the original query with information that helps to increase the amount of hits which are still interesting for the user. The aim is to give support to the users to express their willingness in a transparent way, while keeping the trade-off between precision and recall stable.

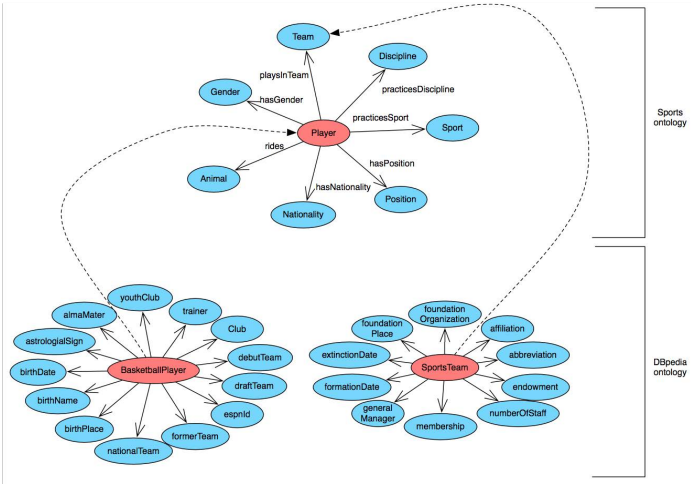

Figure 1: Knowledge expansion by data interlinking

In playence media, we reuse the knowledge of the domain ontology to complement the queries introduced by the user, leading to more precise results than simple queries relying uniquely on syntactic approaches. As stated previously, this knowledge is used during the annotation of media assets and it is exploited on query time. Figure 2 shows an overview of this process. In this example, the user introduces the query *pau gasol*. The system analyzes the user's input and finds that this text makes reference to the concept *Pau Gasol* modeled in the domain ontology (despite not being represented in this example, it is possible to find multiple correspondences of the domain ontology in the user query). When the concept is identified the system uses the ontology to extend the query with related information to the concept, so that media assets annotated with this extra information can be retrieved. In the figure, note that the media asset is not annotated with the concept *Pau Gasol*, however it is retrieved because previously was annotated with the concept *Los Angeles Lakers*, which is directly related with the concept *Pau Gasol*. The algorithm that performs the query extension can be tuned to consider different neighborhood distances in the domain ontology. In this example, we have used a distance of one hop for simplicity.

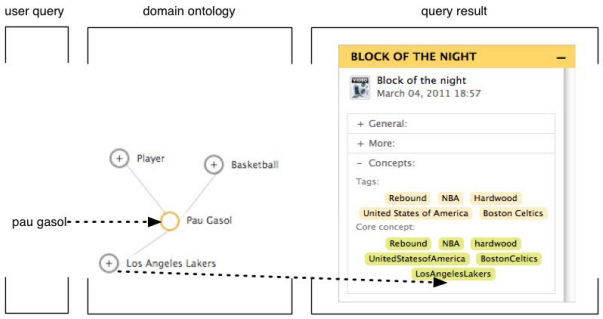

Figure 2: Semantic query expansion

2.2 Content Augmentation

Current approaches to annotate media or textual content rely on the manual copy/paste of information available across different data sources. This is not efficient because users can only slowly navigate the large corpus of knowledge. Furthermore, copy/pasting information is not desirable because it introduces outdated and redundant information. In this way, the application of interlinking approaches to multimedia content is a recent idea that tries to exploit the semantic relationships among objects or sequences within multimedia resources relying on the LOD principles. The main target is to enhance media annotation by reusing the knowledge from the Linked Data cloud. The way how the information is reused to complete the annotations depends on the application aim. This means that the same information can be combined in several manners to offer different functionality to the user.

In playence media, the assets are complemented with information extracted from related sources previously calculated through interlinking. Starting from an annotation seed the system is able to find related sources that can be fused to present valuable information to the user. The way the different data sources are combined corresponds to the area of data integration, which is out of the scope of this document. Figure 3 is an example depicting a video about Pau Gasol. This video has an annotation where the concept *Pau Gasol* has been identified. This concept has been previously interlinked with the equivalent concept in DBpedia, so that new information can be dereferenced and added to the initial

[2]http://dbpedia.org/About
[3]http://www.freebase.com

description of the video. In this case, the figure shows how the abstract of the DBpedia page is added to the original description.

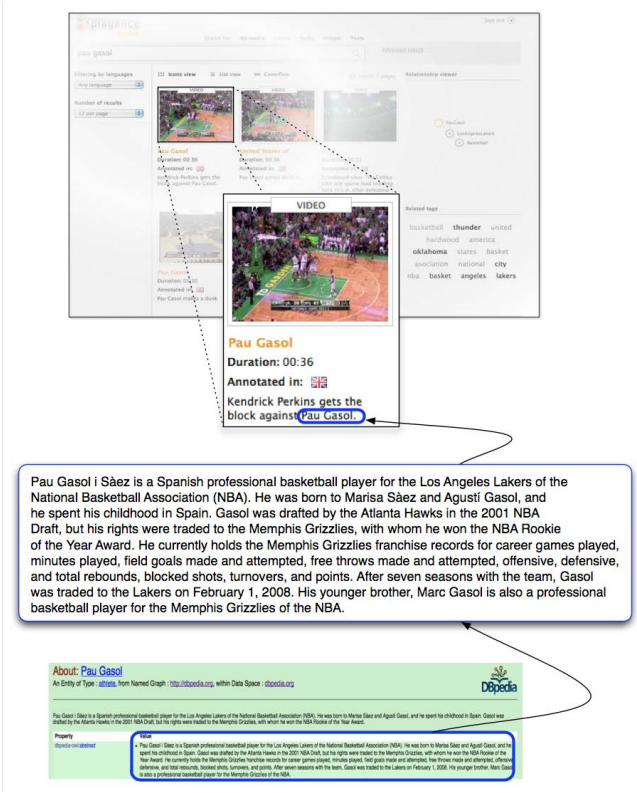

Figure 3: Content augmentation in media assets

3. INTERLINKING MEDIA ASSETS AND LOD

In this section we describe the interlinking approaches that are currently implemented in playence media and allow the functionality mentioned in the previous section. Note we just give an overview of each approach and exclude any evaluation or comparison to other techniques.

3.1 NER based interlinking

NER makes reference to Named Entity Recognition, a technique within the scope of text engineering whose target is identifying occurrences of entities in unstructured texts. We have applied this method for interlinking our sports ontology with DBpedia.

The idea relies on applying a NER process to the string extracted from every URI in the sports ontology. The entities used for recognition are the entities available on DBpedia. In this way, as a previous step to the recognition we have created a dictionary containing the DBpedia entities that will be used for the lookup. New links are created when there is a match, i.e, an entity is recognized in the URI. In order to increase the amount of hits the different strings have been canonized. For doing so, we have eliminated stopwords and applied processes of lemmatization and stemmization.

The main problem of this method is that its accuracy relies only on the little amount of information available on the URI, which is not enough for disambiguating, i.e., it is not possible to decide whether a URI refers to a certain entity when there are different named entities sharing the same name. E.g., Cordoba can make reference to a city on the South of Spain and at the same time to a city in Argentina. Furthermore, this method is not working when URIs use any special kind of representation that has nothing to do with the contents being referenced. A small improvement to this approach contemplates related labels available in the RDF dataset, usually modeled through predicates like *rdfs:label* and *skos:altLabel*. While this modification has shown a better behavior to recognize entities represented by non-descriptive URIs, the problem of ambiguity still remains in most of the cases.

In order to overcome these problems, in the following section we present another method taking into account the context of the entities represented by the URIs.

3.2 Context based interlinking

Assuming an RDF data set D (a simple ontology or a full data set published as LOD) and a given identifier id (modeled as a URI) for an entity e in D, we define the context as the set of RDF triples modeling any statement about e. The context of an identifier contains detailed information that can help to understand what kind of entity the identifier refers to. This information can be used to perform entity disambiguation in a more robust way than using only the URI identifier or associated labels.

Our method inspires on the ideas presented in [2], where the authors describe an algorithm based on graph matching. This algorithm constructs a graph similarity that consists on calculating all possible graph mappings between pairs of resources in the different RDF graphs. Each mapping has a measure associated which is computed applying a string comparison algorithm to the literals attached to the resources. New links are created when the similarities are higher than a given threshold to avoid dissimilar mappings in the result.

The difference regarding to our method relies on how the contexts are exploited. In our case, we do not need to explore the graphs for performing the disambiguation. We do this by using information retrieval techniques. The process implemented is as follows.

First, we have indexed the content available on DBpedia using SIREn[4]. For doing so, we have created an individual document for each DBpedia entity, associating the URI with its respective context. This allows us to keep the structure of the information, so that lately it can be retrieved using similar expressions to SPARQL queries combined with full text search. Note that the index will contain the target data sets we want to interlink our data to, allowing the interlinking of a source data set to multiple target data sets at the same time.

Figure 4: Interlinking threshold scale

[4]http://siren.sindice.com/

The second step consists on getting the context of each entity in the source data set. This context will be used to construct a query that will retrieve all the documents in the index related to the entity. The number of queries to perform coincides with the amount of different entities in the source data set. The grade of similarity is measured using the internal score approach implemented in Lucene[5]. Afterwards, Lucene scores are normalized to make them independent on the number of documents indexed. Scores higher than a certain threshold are selected to create the new links. We have defined a threshold scale to split the scores in three different ranges as depicted in figure 4. The first range is composed by documents with high scores that are suitable for creating *owl:sameAs* links. The second range is composed by those documents whose score is not high enough for using a predicate of equivalence, but are somehow related to the identifier. We model those links with *rdfs:seeAlso*. In the third range are the documents not related to the identifier.

4. CONCLUSIONS

In this work we have introduced playence Media, a suite for the management of media assets relying on semantic technologies. We have described our efforts on multimedia annotation and retrieval focusing on the role that interlinking approaches play in our system. Through the development of this work we can conclude the following issues regarding to the annotation, search and management of media information:

1. Automatic processing of multimedia is limited by its nature. The information represented by media content is richer and at the same time more complex than the one represented by text-based formats. This requires a special processing to generate accurate annotations.

2. Users consume the knowledge on the Web in a very inefficient way. They require machine support to navigate network of interlinked knowledge.

3. Integration of heterogeneous data and different formats. In media asset management, not only text to text integration is required but also the integration of text with media content. In addition, for the description of multimedia resources, a plethora of metadata formats are in use, causing interoperability issues.

4. The annotation of content must be done in such a way that generated metadata is consumable by any user without the need of an expert acting as mediator. Linked Data leads to a convergence of vocabulary and some agreement of the users leading to better annotations and search.

5. Searching content. Given high quality metadata, integrated networks of information, within and beyond organizational boundaries, the full potential and power of semantic search can be exploited. However, semantic search must be extended towards needs and requirements of Linked Data.

Relying on these issues, we can establish the following steps for extending our work.

- The NER approach we have introduced has shown its limitation in order to disambiguate entities with a similar spelling in cases where there is a lack of information in the URI or the associated labels. However, due to its simplicity, this technique is characterized by a high scalability that pushes us to use it in our systems for those cases where it behaves accurately. We plan further efforts on improving the linguistic resources and text engineering techniques underlying the NER functionality.

- The described context-based interlinking approach relies on how good the internal queries for consulting the index are modeled as they are responsible for retrieving the set of documents used to define the linkage associated to a given entity. Right now, these queries are implemented in a generic fashion, however we are aware that a better accuracy could be reached considering specific properties of the data to be interlinked. A trade-off needs to be found in order to keep the user transparency while still improving the results.

- Along this document we have shown how useful media annotations are for our systems, neglecting any detail about how third parties could also benefit from them. On this aspect, we plan to develop a publication strategy compatible with the Open Annotation Collaboration (OAC)[6] and accomplishing with any kind of intellectual rights of the data involved in the annotations. Our aim is to make the annotation set available as Linked Open Data.

5. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst*, 5(3):1–22, 2009.

[2] Y. Raimond, C. Sutton, and M. Sandler. Automatic interlinking of music datasets on the semantic web, 2008.

[3] S. Woelger, K. Siorpaes, T. Buerger, E. Simperl, S. Thaler, and C. Hofer. A survey on data interlinking methods. Technical report, Semantic Technology Institute, March 2011.

[5] http://lucene.apache.org/java/3_0_1/api/core/org/apache/lucene/search/Similarity.html

[6] http://www.openannotation.org/