Leveraging Linked Data in Social Event Detection

Timo Hintsa
VTT
P.O.Box 1000
FI-02044, VTT, Finland
+358 40 837 2723
timo.hintsa@vtt.fi

Sari Vainikainen
VTT
P.O.Box 1000
FI-02044, VTT, Finland
+358 50 525 5794
sari.vainikainen@vtt.fi

Magnus Melin
VTT
P.O. Box 1000
FI-02044, VTT, Finland
+358 40 589 6384
magnus.melin@vtt.fi

ABSTRACT
In this paper, we present our approach and results for the MediaEval 2011 Social Event detection task. VTT participated in Challenge 2 where a given dataset of Flickr photos were matched to events in certain places. We used Linked Data to enhance the dataset by adding event information and other related data and then searching the enhanced dataset. Additional information relating to venues and places were used for creating a subset of photos for each place; Barcelona and Amsterdam. The event profiles including semantically enhanced metadata were used in media retrieval. The approach of combining additional data from the Internet and limiting the queries to limited subsets improved the relevance of photos relating to the events.

Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]:, H.3.1 Content Analysis and Indexing; H 3.3 Information Search and Retrieval; H5.3 On-line Information Services

General Terms
Experimentation

Keywords
events, Linked Data, metadata enhancement, media retrieval

1. MOTIVATION AND RELATED WORK
The challenges, dataset and evaluation methods of the Social Event detection task are described in [3]. VTT participated in challenge 2 where the task was to find all events that took place in May 2009 at defined venues, Parc del Forum in Barcelona and Paradiso in Amstedam, and to find all photos associated with the events.

In our earlier research [4] we have worked with personalized recommendations where events were recommended to the end user based on the user's interests. The approach was to test similar methods for "recommending" relevant media items to the event. In our earlier work with user profiles we have used Linked Data[1] and publicly available semantic databases such as Freebase[2], DBpedia[3] and GeoNames[4] for enhancing the user profile with additional semantic information [2,4]. In this challenge we used Linked Data for enhancing the event descriptions and for multi-language support. The information was used for creating the "profile" for the event and matching it with the metadata of photos.

2. DESCRIPTION OF THE APPROACH
The main point of the approach was to connect the given photos to events that were found using the Linked Data sources on the Internet. Linked Data was used to get additional information relating to events, artists, venues and places.

2.1 Enhancing Dataset with Linked Data
First we used publicly available event services such as Last.fm[5] and Upcoming[6] to find information about the relevant events. The event descriptions including title, description, artist, time and venue information was stored in a database.

By using Freebase we looked up the unique identifiers for the artists and bands. Based on these URIs, additional information such as genre and band members were collected and stored in a database. This additional information was used for updating the "profile" of the event.

We used Freebase and GeoNames for getting additional information relating to places. This included getting coordinates for the venues and cities, as well as different language versions for the cities and countries. We used Freebase for getting information about the tourist attractions in Barcelona and Amsterdam, and GeoNames for getting places near venues by utilizing coordinates. An assumption was that these were things that users commonly use for describing the photos.

We created a limited dataset for each place based on the photo location information. The tourist attractions, nearby places and coordinates that were too far from the venues were used to exclude irrelevant photos from the limited dataset. The goal was to be able to create more relevant matches between the events and photos.

2.2 Run Configurations
2.2.1 First Run
In the first run, searching for photos that matched the relevant events was made against the datasets in which the photos were limited based on the places.

The run consisted of a set of queries that include matching the artist name and the time of the event, the venue name and the time of the event, and the event name and the time of the event with the metadata (title, tags, description and time taken) of the photos in the dataset.

[1] http://linkeddata.org

[2] http://www.freebase.com

[3] http://dbpedia.org

[4] http://www.geonames.org

[5] http://www.last.fm

[6] http://upcoming.yahoo.com

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

The goal of this run was to get a set of highly relevant matches between events and photos.

2.2.2 Second Run

In the second run we used the results of the first run, but we created additional searches for the total dataset of photos for finding more relevant photos.

Event names without time restriction were queried against the metadata of photos. In the case of Parc del Forum, event names were quite unique such as Primavera Sound 2009 and the queries found relevant images. In the case of Paradiso the name of events were often same as the artist that were performing in the event. If time restriction was not used together with the event name, quite a lot of irrelevant photos were attached to the events. We used this query only in the case of Parc del Forum.

The event profiles and their tag clouds were enhanced with the results of the first run, namely the tags from the photos that were found relevant to the event. In this phase, the event profiles consisted of the event name, venue, city, artists, genre, band member information, and the photo tags from the previous run.

Apache Solr[7] and Lucene[8] were used in free-text indexing and searching the textual photo metadata, namely tags and photo descriptions. The photo index was searched with the information in the event profile. The Lucene score limit for accepted result was set relatively high (i.e. 0.5) so that the irrelevant photos would be left out. To further increase the relevance the searches were run on the limited datasets of the Barcelona photos and the Amsterdam photos as described in chapter 2.1.

3. RESULTS AND DISCUSSION

The results of our submitted runs can be seen in the table 1. The evaluation measures are described in [3].

Table 1. The results of the submitted runs

Run	Precision	Recall	F-score	NMI
1	72,18	48,41	57,96	0,5839
2	73,79	64,21	68,67	0,6782

As expected, the recall of the first run was low due to the use of the limited set of photos, however the photos were quite relevant. Our additions to the second run improved the results and more relevant photos were found.

Our approach of limiting searches to the subset of photos, which was created based on additional information gathered from Linked Data, increased the relevance of photos.

One challenge in the development was the unreliability of the photo metadata. We could see that the photo timestamps that are created by different cameras were not always reliable. This made it difficult to match different images to events using the time information. The same problem was noted with the GPS coordinates where even the inherent error in location precision in city environments is tens of meters [1]. This is particularly shown in the Paradiso case where distances as low as 100 meters from the center of the building yield false positives.

When analysing the irrelevant photos in the results of the second run we found that more logic should be developed for checking the reliability of the results. To enhance the quality of the second run, the event profile created from the users' tags should have been cleaned up from irrelevant tags regarding the image content, e.g. camera makers and models. Further analysis of tag relevance based on occurrence and co-occurrence could have been made to further define the tag relevancies to images and the event.

We planned to make the semantic analysis [2] of users' tags, but did not do it due to the time needed to analyse all the images and seemingly high variance on the quality of the tags themselves. However, the analysis would have helped to better determine the place-related tags and remove false positives in the result sets.

A search for other photos from the same user within same timeframe as the ones found in the first run was not conducted. This search would have helped to find photosets where only one or few of the photos are tagged, but the rest of the photos are from the same event.

Solr parameters, like the score parameter, can be adjusted further and more logic can be added to find irrelevant photos especially when the score parameter value is lowered. Other Lucene functionality like MoreLikeThis would also be worth exploring.

4. ACKNOWLEDGMENTS

The work presented in this paper was partially funded by the OpenSEM project funded by EIT ICT Labs. We would like to thank Onni Ojutkangas, Asko Ollila, Johannes Peltola, Antti Nummiaho and Mika Timonen for code snippets, thoughts and ideas while planning and realizing this task.

5. REFERENCES

[1] Modsching M., Kramer R. and ten Hagen K. 2006. Field trial on GPS Accuracy in a medium size city: The influence of builtup. 3rd Workshop on Positioning, Navigation and Communication 2006, WPNC'06 Hannover, Germany March 16 2006. Proceedings.

[2] Nummiaho A., Vainikainen S., Melin M. 2010. Utilizing Linked Open Data Sources for Automatic Generation of Semantic Metadata. Metadata and Semantic Research 4th International Conference, MTSR 2010, Alcalá de Henares, Spain, October 20-22, 2010. Proceedings. Metadata and Semantic Research,Communications in Computer and Information Science, 2010, Volume 108, 78-83, DOI: 10.1007/978-3-642-16552-8_8.

[3] Papadopoulos S., Troncy R., Mezaris V., Huet B. and Kompatsiaris I. Social Event Detection at MediaEval 2011: Challenges, Dataset and Evaluation. In MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy.

[4] Vainikainen S., Laakko T., Giesecke R., Vesikivi P. 2011. Context awareness – portable profiles, HTML5 and advertiser's metadata. Next Media deliverable D3.0.1.2.

[7] http://lucene.apache.org/solr

[8] http://lucene.apache.org