Report of the STSM

Host institution: Insight Centre for Data Analytics at UCD, Dublin
Student: Sergio Oramas Martín
Keystone COST Action IC1302

Purpose of the STSM

During the proposed Short Term Scientific Mission, our work will be focused on creating a methodology for semantic annotation and expansion of keyword queries, and the use of the extended query for searching through the structured data source provided by the tagging ontology of Freesound. The main idea is to apply Entity Linking and Word Sense Disambiguation to the keyword query, and expand it using WordNet and DBpedia. Currently available semantic technologies will be tested and customized algorithms will be implemented. On the other hand there is a huge amount of unstructured data in Freesound that might be exploited to create a more semantically meaningful data structure. To this end, tags, text descriptions and user reviews will be exploited and structured into a domain ontology. Once the data is structured and the system is able to search into the structured data source the user experience should improve. Finally, we will carry out an evaluation with real users. This evaluation will serve as a proof of concept for the utility of the proposed methodology for semantic expansion of keyword queries.

Description of the work carried out

- The domain ontology of Freesound has been expanded by applying entity linking to tags and text descriptions and disambiguating the detected entities to WordNet using Babelfy. Then, an initial ontology manually defined has been populated with tags and keywords using the WordNet hyperonymy relations. This work was an extension of an initial work previously done on a small portion of the Freesound data.

- User reviews in Freesond has been analysed using a framework for opinion mining developed in the host institution. In every review, opinion features are extracted, together with sentiment words and sentiment score. The idea was to expand the structured knowledge base with sentiment information.
- An entity recognition algorithm has been trained for the specific domain to be used for query expansion. We developed a classifier trained with linguistic and contextual features using Conditional Random Fields over a dataset of annotated musical entities. This entity recognizer was tested together with other state of the art NER systems, showing a clear improvement over the specific data. This research was written down in a paper entitled "Is it a song? Benchmarking Named Entity Recognition in the Music Domain" that will be submitted to LREC 2016.
- The entity recognizer was applied to the Freesound query search and the detected entities were disambiguated against the structured data source. Then the query was converted into a graph where the nodes are the detected entities in the query and it was expanded using the h-hop neighbourhood subgraph of every entity in the data source. Then similarity between the query graph and the subgraph of every sound was computed applying Maximal Common Subgraph measure. This methodology for graph similarity was published during the STSM by the authors in [1]. The sounds are then ranked by similarity score and presented to the user following this order. The evaluation environment has been developed but the experiments haven't been carried on yet. We plan to do the experiments in September and write a paper about the complete methodology and results for query expansion in the music and sound domain.
- In addition to the work on Freesound, we applied the opinion mining framework to a dataset of music reviews from Amazon CDs and Vinyls. The dataset contains 3,778,294 reviews of 486,361 products. Features, opinion words and sentiment scores were extracted from every review. Then, topic modelling was applied to the extracted features in order to identify the different dimensions that define the music. We tried with two

different approaches for topic modelling, LSI and NMF. The ultimate end of this research is to be able to give explanations to recommendations, giving a sentiment score to every music dimension in every product. The same methodology can be applied to search results. This may help the user in understanding the ranking in the result list, and can be used combined with the user profile in order to provide more personalized search results.

Future collaboration and foreseen publications

As depicted before, there are three research lines to be finished and three publications will be send to different conferences or journals. First, the domain specific entity recognition tool will be send to LREC 2016. Second, the application of the entity recognition and disambiguation to the query search and the results of the search experiments will be written down into a paper and will be sent to a conference to be determined. Third, the work on opinion mining and topic modelling over Amazon reviews for recommendations will be continued in collaboration with the host institution and will be also published.

Bibliography

[1] Oramas S., Sordo M., Espinosa-Anke L., Serra X. (2015). *A Semantic-based approach for Artist Similarity*. 16th International Society for Music Information Retrieval Conference ISMIR 2015. In Press.