A novel summarisation approach from embedded Markup on the Web

---

STSM Scientific Report - COST Action IC1302 KEYSTONE

STSM Applicant

Davide Taibi, Consiglio Nazionale delle Ricerche, Istituto per le Tecnologie Didattiche, Palermo (ITALY)
davide.taibi@itd.cnr.it

STSM Host

Dr. Stefan Dietze, L3S Research Center, Hannover (DE)
dietze@l3s.de

Period

14/11/2016-18/11/2016

Purpose of the STSM

The purpose of the STSM is to investigate the possibility to develop a novel approach for retrieving and analyzing entity summaries from embedded markup on the Web. Embedded markup languages enable the annotation of unstructured Web pages with structured facts through Microdata, RDFa and Microformats. Such annotations are used by major search engines to facilitate the interpretation of Web content, but at the same time, represent an unprecedented source of knowledge. In particular, the Schema.org initiative, driven by Google, Yahoo!, Yonder, and Bing, has led to an increasing adoption of embedded markup by providing a common vocabulary for describing a wide variety of entities.

However, RDF statements extracted from markup are fundamentally different to traditional RDF graphs: entity descriptions are flat, facts are highly redundant and granular, and co-references are very frequent yet explicit links are missing. Therefore, carrying out typical entity-centric tasks such as retrieval and summarization cannot be tackled sufficiently with state of the art methods. To this aim, it is necessary to investigate new approaches specifically designed to be effective with this type of data source.

Particular emphasis will be posed to the application of the proposed approach to Web resources containing markup related to learning resources metadata initiative (LRMI) properties [1]. In April 2013 the metadata schema developed by the Learning Resource Metadata Initiative (LRMI) to describe educational resources has been added to the Schema.org vocabulary and is currently under development by the LRMI task group of the Dublin Core Metadata Initiative (DCMI). Finally, a preliminary investigation into the coverage and complementarity of retrieved facts, compared to existing knowledge graphs, will be also undertaken.

Description of the work carried out and main results obtained during the STSM

Given that markup constitute a knowledge resource fundamentally different to traditional Linked Data and knowledge graphs, our work contribute specifically to the problem of search and retrieval in structured markup, i.e. RDF data extracted from embedded annotations.

In particular, during the period of the STSM the following activities have been carried out:

- Research studies on recent approaches on summarization approaches based on embedded Markup on the Web. The state of the art in summarisation approaches from embedded Markup have been analysed [2][3].
- Participation to meetings with the host researcher to select the approaches to be tested, with particular references to learning resources.
- Experimentation of different approaches based on clustering techniques such as: K-means, X-means, LDA (Latent Dirichlet Allocation).
- Definition of a preliminary structure for a research paper aimed at describing the experimentation results.

Future collaboration with the host institution

The activities carried out during the STSM have been developed in the framework of previous collaborations between the two research centres. Further collaborations in the topic of summarization approach from markup on the Web especially focused on learning resources have been planned.

The preliminary results obtained during the STSM have proven the feasibility of the proposed approach, therefore future research aimed at improving the quality of the results will be undertaken. Moreover, an expected result of this STMS visiting period is the submission for the publication of the work in one of the major Semantic Web and Web mining conferences of high relevance to the KEYSTONE targets. In particular, the 9th International ACM Web Science Conference (WebSci 2017), that will be held from June 26 to June 28, 2017 in Troy, NUY (USA) has been identified as a possible target.

Other comments

During the STSM visits the researcher Davide Taibi has been invited to participate to other meetings with L3S researches to discuss further collaboration opportunities on topic related to KEYSTONE.

References

[1] Davide Taibi, Stefan Dietze: Towards Embedded Markup of Learning Resources on the Web: An Initial Quantitative Analysis of LRMI Terms Usage. WWW (Companion Volume) 2016: 513-517

[2] Ran Yu, Ujwal Gadiraju, Xiaofei Zhu, Besnik Fetahu, Stefan Dietze: Towards Entity Summarisation on Structured Web Markup. ESWC (Satellite Events) 2016: 69-73.

[3] Ran Yu, Besnik Fetahu, Ujwal Gadiraju, Stefan Dietze: A Survey on Challenges in Web Markup Data for Entity Retrieval. International Semantic Web Conference (Posters & Demos) 2016.