Université
franco-allemande
Deutsch-Französische
Hochschule

Semantic Snippets via Query-Biased Ranking of Linked Data Entities

Mazen ALSAREM

Abstract

In our knowledge-driven society, the acquisition and the transfer of knowledge play a principal role. Web search engines are somehow tools for knowledge acquisition and transfer from the web to the user. The search engine results page (SERP) consists mainly of a list of links and snippets (excerpts from the results). The snippets are used to express, as efficiently as possible, the way a web page may be relevant to the query.

As an extension of the existing web, the semantic web or “web 3.0” is designed to convert the presently available web of unstructured documents into a web of data consumable by both human and machines. The resulting web of data and the current web of documents coexist and interconnect via multiple mechanisms, such as the embedded structured data, or the automatic annotation.

In this thesis, we introduce a new interactive artifact for the SERP: the “Semantic Snippet”. Semantic Snippets rely on the coexistence of the two webs to facilitate the transfer of knowledge to the user thanks to a semantic contextualization of the user’s information need. It makes apparent the relationships between the information need and the most relevant entities present in the web page.

The generation of semantic snippets is mainly based on the automatic annotation of the LOD[1]’s entities in web pages. The annotated entities have different level of impor- tance, usefulness and relevance. Even with state of the art solutions for the automatic annotations of LOD entities within web pages, there is still a lot of noise in the form of erroneous or off-topic annotations. Therefore, we propose a query-biased algorithm (LDRANK) for the ranking of these entities. LDRANK adopts a strategy based on the linear consensual combination of several sources of prior knowledge (any form of con- textual knowledge, like the textual descriptions for the nodes of the graph) to modify a PageRank-like algorithm.

For generating semantic snippets, we use LDRANK to find the more relevant entities in the web page. Then, we use a supervised learning algorithm to link each selected entity to excerpts from the web page that highlight the relationship between the entity and the original

information need.

In order to evaluate our semantic snippets, we integrate them in ENsEN (Enhanced Search Engine), a software system that enhances the SERP with semantic snippets.

Finally, we use crowdsourcing to evaluate the usefulness and the efficiency of ENsEN.

Keywords: Semantic Snippets, Entity Ranking, Web of Data.