

a semantic search project

z bjelogrić, dw van gulik, a. reggiori, d norheim
www.aseantics.com

Search engines such as Google (google.com), Fast (alltheweb.com) and others offer an extremely fast and broad search of the web. Services like this provide a very good approach during the initial attempts to localize a certain resource on the web, especially if the resource has some unique distinctive characteristics which are known a-priori to the user. However due to the broadness the approach becomes more difficult to use when a narrow or more refined search is needed in a relatively homogeneous dataspace or when the unique characteristics are either unknown or embodied in the relative position, linkage or across third party descriptions of the resource rather than encoded in the resource itself. The search can produce hundreds or thousands of hits, most of which are false positivies, especially when common terms are used.

To improve this situation a different approach, based on Semantic Web technologies is proposed. The approach is summarized in the following steps:

augmented search – the search phrase is analyzed and context and associated activities are identified. The search is augmented by adding data relative to the context or activity. The approach is similar to the one proposed by the “Semantic Search” application of the TAP¹ project. Knowledge base used to determine activities is specialized to Earth Observation-related applications and perhaps shared with the Semantic Web SWEET² project of JPL.

meta-search is applied to target web search engines (Google, etc.) using the augmented serch data above, results are collected, described RDF and consolidated to be used for further processing. Direct search is possible on selected hits/sites according to filtering rules. In the later case, FOAF (Friend-of-a-Friend) relations are created, e.g. to localize images referenced in the hits or people/events mentioned. This opens the door to searches like “give me all images about oil pollution in the Atlantic”.

postprocessing is applied to hits found, even as categorization (e.g. using DMOZ³ directory already written in RDF), site clustering and a more sophisticate topic clustering. The objective of this phase is to organize better the hits, possibly without or with minimal user intervention (e.g. clustering adopted by Vivisimo⁴).

learning process can be further applied to personalize and tune all the steps above for a group of users .

The objectives of the project are to analyse and demonstrate the approach and assess its feasibility.

In particular, all the steps above represent specialistic areas for which prototypes or operational systems were already developed. Our goal is to check also if RDF as standard and tools such as RDFStore⁵ can be used for efficient cooperative deployment of the proposed approach.

¹ <http://tap.stanford.edu>

² <http://sweet.jpl.nasa.gov>

³ <http://dmoz.org/>

⁴ <http://www.vivisimo.com>

⁵ <http://rdfstore.sourceforge.net/> <http://ww.aseantics.com/downloads>