A Semantic Clouding Approach for Cross-Webs Interoperability

Silvana Castano, Alfio Ferrara, Stefano Montanelli, Gaia Varese

Università degli Studi di Milano
Dipartimento di Informatica e Comunicazione
Via Comelico, 39 – 20135 Milano, Italy
{castano, ferrara, montanelli, varese}@dico.unimi.it

Abstract. The classical vision of the Web as a merely publishing environment for information-consuming users is being replaced by a plural vision where multiple webs, like Web 2.0, Social Web, and Semantic Web, co-exist and interoperate to make information sharing more effective and socially pervasive. In this paper, we propose a *semantic clouding approach* for the construction of cross-web, disciplined, and intuitive information organization structures called *i*-clouds. An overview of the proposed semantic clouding approach is presented in the paper, as well as an example of *i*-cloud over real web resources about a movie dataset.

1 Introduction

Over the recent years, the classical vision of the Web as a merely publishing environment for information-consuming users is being replaced by a plural vision where multiple webs, like Web 2.0, Social Web, and Semantic Web, co-exist and interoperate to make information sharing more effective and socially pervasive. The experience of active research projects like OKKAM and Linked Data places the accent on the growing need to recognize identity and similarity relations between data descriptions provided by different web sources in different domains. The variety of webs data, spanning from textual tags to RDF(S) structural descriptions up to formal OWL instances, makes the above mentioned need of identity/similarity recognition even more crucial and challenging. In such a complex scenario, a new generation of information search techniques is required to cope with the following needs: i) the capability to span across multiple Webs, to properly consider the wide variety of available web resources and pieces of knowledge by properly assessing their information contribution nature; ii) the capability to anticipate the user needs by providing a focused but comprehensive set of web resources prominent for his/her target; iii) the capability to semantically organize all retrieved prominent resources into an intuitive and coherent structure [1,2].

In this paper, we propose a *semantic clouding approach* for the construction of cross-web, disciplined, and intuitive information organization structures called *i*-

clouds. An *i*-cloud is built to organize all the web resources about a certain *target entity of interest* into a graph on the basis of their level of *prominence* and reciprocal *closeness*. An overview of the proposed semantic clouding approach is presented in the paper, as well as an example of *i*-cloud over real web resources about movies. A more technical discussion about construction and formal properties of *i*-clouds is provided in [3].

2 Semantic clouding of web resources

An *i*-cloud is built around a certain target entity, which is a keyword-based representation of a topic of interest, namely a real-world object/person, an event, a situation, or any similar subject that can be of interest for the user. The notions of *closeness*, and *prominence* are define for an *i*-cloud to capture how similar web resources are each other and with respect to the target entity of the *i*-cloud, and the relative importance of a resource within the *i*-cloud, respectively. The following properties characterize *i*-clouds:

- *Cross-webness*. An *i*-cloud collects web resources coming from multiple webs to provide a comprehensive picture of all the available information, both objective and subjective, about the specified target entity for which the *i*-cloud is built.
- *Discipliness*. The web resources in an *i*-cloud are not only those directly related to target entity (i.e., those trivially matching the target entity) but also those that are in some way related to the target and are close to it.
- *Intuitiveness*. The *i*-cloud organization borrows the graphical representation commonly used for folksonomies and tag-clouds. This supports the user in browsing the *i*-cloud more effectively according to closeness and prominence of the web resources therein contained.

For *i*-cloud construction, we propose a semantic clouding approach articulated in three main phases (see Fig. 1): *acquisition of web resources*, *classification of web resources*, and *clouding of web resources*.

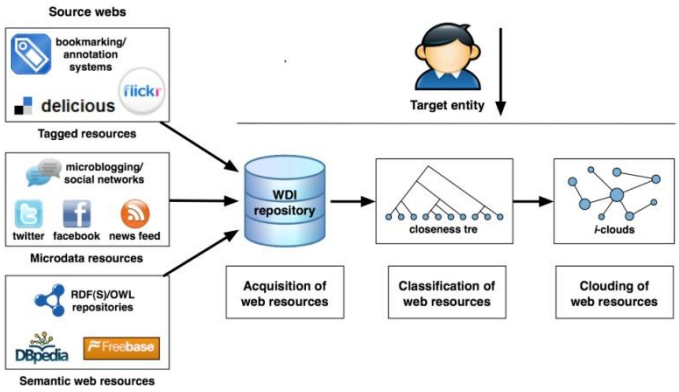

Fig. 1. The semantic clouding approach

Acquisition of web resources. For semantic clouding, all the different web resources are acquired from their respective source webs according to a reference data model called *WDI model*. The WDI model is capable of dealing with a variety of web resources. In particular, a WDI representation is provided for *tagged resources* that are resources from social annotation systems like Delicious, *microdata resources* that are resources from microblogging systems like Twitter, and *semantic web resources* that are resources from RDF/OWL knowledge repositories like Freebase. Each web resource *wr* is stored in a support repository called WDI repository in the form of a *web data item wdi(wr)* where terminological, structural and logical information about *wr* are properly represented.

Classification of web resources. The acquired web resources are grouped together according to their level of closeness. To this end, tailored matching techniques have been developed in the framework of the HMatch 2.0 system [4]. First, term and structural matching techniques are adopted to calculate the level of closeness $CC(wdi_i, wdi_j)$ between any pair of web data items wdi_i and wdi_j in the WDI repository. Then, a hierarchical clustering procedure is adopted to determine a *closeness tree* where all the wdis are properly grouped according to their closeness coefficients previously calculated.

Clouding of web resources. Given a target entity *e* specified by the user, an *i*-cloud is built for *e* by firstly extracting from the WDI repository a *ground set* of web data items that syntactically match *e*. Given a closeness threshold, the wdis in the ground set are used to select a number of candidate clusters in the closeness tree, namely all the clusters containing the wdis of the ground set and the wdis whose closeness is greater than or equal to the threshold. The candidate clusters originate the graph structure of the resulting *i*-cloud through a graph construction procedure. A labelling function is finally applied to assign to nodes and edges of the *i*-cloud their corresponding closeness and prominence values, respectively.

3 An example of *i*-cloud for cross-web interoperability

As an example, we consider the *i*-cloud of Fig. 2 where a number of web resources related to the target entity "Star Wars" are shown. We can observe that resources in the *i*-cloud are not only those directly related to this popular movie, such as the titles of the six movies of the Star Wars saga, but also resources that are close to the movie saga even if not directly matching the target, such as some of the most important characters in the movies. The dimension of each node in the *i*-cloud is proportional to the prominence of the corresponding web resource for "Star Wars" and the edges connecting the nodes are labelled with their closeness degree. We observe that different kinds of web resources populate the "Star Wars" *i*-cloud. In particular, this *i*-cloud is built over resources acquired from the Delicious annotation systems (e.g., wdi(delicious1)), the Twitter microblogging system (e.g., wdi(twitter1)), and the Freebase Linked Data repository (e.g., wdi(iimb1)).

In this example of *i*-cloud, the prominence of the various web resources is calculated through a *popularity-based* mechanism. This means that the prominence of a resource *wr* depends on the "centrality" of *wr* with respect to the *i*-cloud and it corresponds to the degree of connection of *wr* with the other nodes in the graph of the *i*-cloud [5]. Other techniques can be used for prominence computation based on the provenance of the web resources in the *i*-cloud (e.g., [6]).

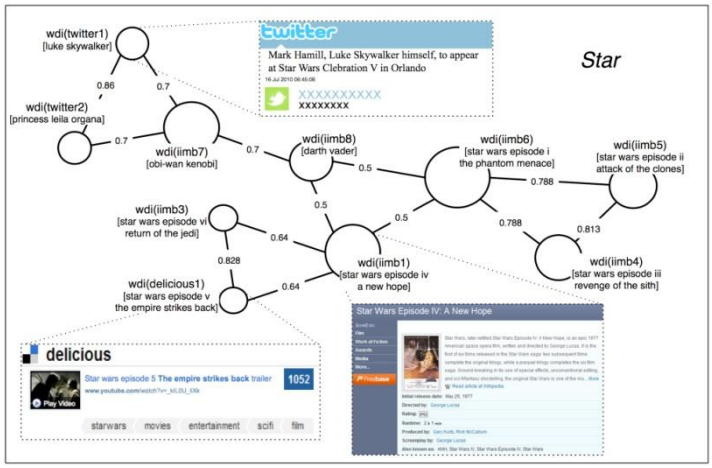

Fig. 2. Example of *i*-cloud for the entity "star wars".

4 Concluding remarks

A more detailed description of the approach and related support techniques can be found in [3]. A prototype for *i*-cloud construction has been developed on top of the HMatch 2.0 environment (http://islab.dico.unimi.it/hmatch) and it has been evaluated on the OAEI-IIMB 2010 dataset (http://www.instancematching.org/oaei/imei2010.html). The

positive results we obtained during evaluation encourages to continue working on *i-cloud* research issues. In particular, a focused search application is being developed in the domain of tourism and entertainment related to the city of Milan.

5 References

1. Kuo, B., Hentrich, T., Good, B., Wilkinson, M.: Tag Clouds for Summarizing Web Search Results. In: *Proc. of the 16th Int. Conference on World Wide Web (WWW 2007)*. Banff, Alberta, Canada, 2007.
2. Koutrika, G., Zadeh, Z., Garcia-Molina, H.: Data Clouds: Summarizing Keyword Search Results over Structured Data. In: *Proc. of the 12th Int. Conference on Extending Database Technology (EDBT 2009)*. Saint Petersburg, Russia, 2009.
3. Castano, S., Ferrara, A., Montanelli, S.: Semantic Data Clouding across Multiple Webs. Submitted to *Information Systems*. Elsevier, 2010.
4. Castano, S., Ferrara, A., Montanelli, S.: Dealing with Matching Variability of Semantic Web Data Using Contexts. In: *Proc. of the 22nd Int. Conference on Advanced Information Systems Engineering (CAiSE'10)*. Hammamet, Tunisia, 2010.
5. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, 2007.
6. Gil, Y., Artz, D.: Towards Content Trust of Web Resources. *Web Semantics: Science, Services and Agents on the World Wide Web* 5(4), 2007.