Semantic Data Clouding over the Webs

- Ph.D. Thesis[*] Abstract -

Gaia Varese

Università degli Studi di Milano
Via Comelico, 39 - 20135 Milano, Italy
gaia.varese@dico.unimi.it

Abstract. Very often, for business or personal needs, users require to retrieve, in a very fast way, all the available relevant information about a focused *target entity*, in order to take decisions, organize business work, plan future actions. To answer this kind of "entity"-driven user needs, a huge multiplicity of web resources is actually available, coming from the Social Web and related user-centered services (e.g., news publishing, social networks, microblogging systems), from the Semantic Web and related ontologies and knowledge repositories, and from the conventional Web of Documents. The Ph.D. thesis is devoted to define the notion of *i*-cloud and a *semantic clouding approach* for the construction of *i*-clouds that works over the Social Web, the Semantic Web, and the Web of Documents. *i*-clouds are built for a target entity of interest to organize all relevant web resources, modeled as *web data items*, into a graph, on the basis of their level of *prominence* and reciprocal *closeness*.

Keywords: Semantic clouding, *i*-clouds, Social Web, Semantic Web

1 The research question of the thesis

The user expectations on the quality of results of web information searches are becoming more and more high. Very often, for business or personal needs, users require to retrieve, in a very fast way, all the available relevant information about a focused *target entity*, in order to take decisions, organize business work, plan future actions. A target entity 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. To answer this kind of "entity"-driven user needs, a huge multiplicity of web resources is actually available, coming from the Social Web and related user-centered services (e.g., news publishing, social networks, microblogging systems), from the Semantic Web and related ontologies and knowledge repositories, and from the conventional Web of Documents. Each kind of web resource is differently structured according to a variety of formats, ranging from short, unstructured, and ready-to-consume news/posts, to well-structured, formal ontology, and each one can provide unique information for a given target entity. For example, only web resources coming from the Social Web are able to provide subjective information reflecting users opinions or preferences about

[*] Ph.D. thesis supervisor: Prof. Silvana Castano - Università degli Studi di Milano.

the target entity, which complement in a useful way the more objective information provided by web resources coming from the other webs. To satisfy user expectations, a new generation of web information search techniques has to cope with different requirements: 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 relevant for the target entity; iii) the capability to semantically organize all retrieved web resources into an intuitive and coherent structure for the given target entity.

With respect to this scenario, the Ph.D. thesis is devoted to define the notion of *i*-cloud and a *semantic clouding approach* for the construction of *i*-clouds that works over the Social Web, the Semantic Web, and the Web of Documents. *i*-clouds are built for a target entity of interest to organize all relevant web resources, modeled as *web data items*, into a graph, on the basis of their level of *prominence* and reciprocal *closeness*. Prominence captures the importance of a web resource within the *i*-cloud, by distinguishing, also in a visual way "a la tag-cloud", how much relevant web resources are with respect to the target entity. The level of closeness between web resources is evaluated using matching and clustering techniques, with the goal of determining how similar web resources are to each other and with respect to the target entity.

The research methodology followed for the Ph.D. activity is based on the following main phases: i) *literature review* with the aim of providing a critical comparison of the state of the art solutions for semantic data clouding, ii) *conceptual design* where requirements and foundational aspects related to the Ph.D. issues are formally addressed, iii) *prototype implementation* where a prototype tool is developed according to the defined architecture, and iv) *evaluation* of the proposed techniques on a number of real test cases.

2 Related work

Relevant research work with respect to the Ph.D. thesis regards Linked Data, instance matching, and data clouds.

Linked Data. A new generation of web applications for the integration of both data and services is being emerging in the context of the Linked Data project [2]. Linked Data is mainly focused on the idea of improving interoperability and aggregation among large data collections already available on the web, such as for example DBLP [1], DBPedia [2], CiteSeer [3], IMDB [4], and Freebase [5], which are available as retrievable RDF datasets or SPARQL query endpoints. Linked Data is a step beyond the simple availability of data and syntactic compatibility, in that it promotes some important principles in making web data available and sharable to the Semantic Web community. Such principles are

[1] http://www.informatik.uni-trier.de/~ley/db

[2] http://dbpedia.org

[3] http://citeseerx.ist.psu.edu

[4] http://www.imdb.com

[5] http://www.freebase.com

the following: i) all the web resources have to be referenced by a URI; ii) URIs have to be resolvable on the web to RDF descriptions; iii) RDF triples have to be consumed by a new generation of Semantic Web browsers and crawlers [15]. However, Linked Data does not take into account the web resources originated from user-generated contents like comments, posts and personal feeds, that are characterized by poor structure and rapid obsolescence. Moreover, Linked Data builds a flat graph structure of interconnected URIs, without distinguishing the prominence and closeness of web resources.

Instance matching. The same real-world object can be described multiple times in different knowledge repositories, possibly using different perspectives and by emphasizing different properties of interest. The capability of finding similar object descriptions assumes particular relevance in the field of Semantic Web, to promote effective web resource sharing on the global scale and to correctly interoperate/reuse individual knowledge chunks coming from disparate information repositories, disregarding their specific URIs. Such task is called *instance matching*, and consists in finding instances (i.e., object descriptions), coming from different sources, which describe the same real-world object in a different and heterogeneous way. Some contributions in this direction have been focused on defining techniques and approaches for the generation and management of identifiers at object-level, like, for example, the OKKAM project [3]. Other approaches have been proposed for the unification of different URIs associated to the same object [13]. Moreover, a problem related to instance matching is the one of finding object descriptions referring to similar objects. To this end, suitable matching techniques are required. Such techniques are mainly provided by the research work in the field of record linkage, which has been widely studied in the databases community [8]. More recently, some new techniques have been proposed to specifically match ontology instances [9] and to identify similar web resources [11]. However, none of the proposed approaches is able to compare different kinds of object.

Data clouds. In the recent years, the traditional World Wide Web based on "user-consuming" applications and informative web pages has changed into a more complex vision composed of a plurality of webs, where semantic-intensive applications as well as interactive "user-generated" platforms like microblogging, and news feeds are becoming more and more popular. In this scenario, the research efforts towards the development of solutions for organizing this huge amount of web resources according to semantic clouding or similar approaches is still at an initial stage [10]. Some interesting work has been done in the field of news aggregation, with the aim of providing techniques for their semantic organization and classification. Examples of proposed systems are NewsInEssence [14] and Relevant News [1], which automatically group news related to the same topic by exploiting hierarchical clustering algorithms and tag/keyword-based search functionalities. For what concerns microdata sources, like Twitter or Facebook, tools for semantic aggregation are still missing. In the same direction, structured and collaborative search engines are being emerging as a promising solution for presenting the query results in a sort of structured form. Examples in this field are Wolfram Alpha [6] and Google Wonder Wheel [7]. In particular, Wolfram Alpha

[6] http://www.wolframalpha.com

[7] http://www.googlewonderwheel.com

is a computational knowledge engine based on data extraction from popular knowledge repositories, like Wikipedia. The goal of this engine is to provide answers to the user requests by returning a comprehensive picture of the available data retrieved about the given request. The same idea is enforced by Google Wonder Wheel, which provides also a graphical, cloud-oriented view of the query results based on terminological similarities among different web resources. However, all these proposed solutions still lack the integration between Social and Semantic Web resources, and provide a poor support of semantic matching techniques for identifying similar web resources.

2.1 Contributions of the thesis

With respect to the state of the art, the contributions of the Ph.D. thesis are mainly the following.

- Definition of a *cross-web* approach considering the different kinds of available web resources (e.g., tagged resources, microdata resources, Semantic Web resources), and considering both objective and subjective information. As far as we know, our semantic clouding approach represents a first attempt to bridge the gap between Semantic Web resources (typically managed in Linked Data) and other kinds of web resource, such as, for example, tagged and microdata resources.
- Definition of *i-cloud* as a new data structure for organizing relevant web resources for a given target entity on the basis of their prominence and closeness.
- Definition of *matching techniques* for comparing different kinds of web resources.

In particular, in Table 1, the differences between Linked Data and *i*-clouds are summarized.

Linked Data	***i*-cloud**
Resulting structure: graph	Resulting structure: graph
Aim: connect different RDF descriptions of the same object	Aim: organize the relevant web resources for a target entity
Off-line process	On-line process
One general graph (connecting different repositories)	One graph for each target entity
Directed graph	Undirected graph
Unweighted graph	Weighted graph
The nodes can be URIs or literals	The nodes are web data items (wdis)
The edges can be labeled with properties or with *owl:sameAs*	The edges are labeled with the value of closeness between wdis
Connected data are described using RDF	Connected data are described using the WDI model
No distinction between the nodes	Each node has a different prominence
Only descriptions referred to the same object are connected	Similar wdis are connected by different closeness values
Data which are not described using RDF cannot be included	Each kind of web resource can be included

Table 1. Comparison between Linked Data and *i*-cloud

3 The proposed semantic clouding approach

In Figure 1, we show the semantic clouding approach developed for *i*-cloud construction. The approach is articulated in three phases: i) modeling of web resources, ii) classification of web resources, and iii) clouding of web resources.

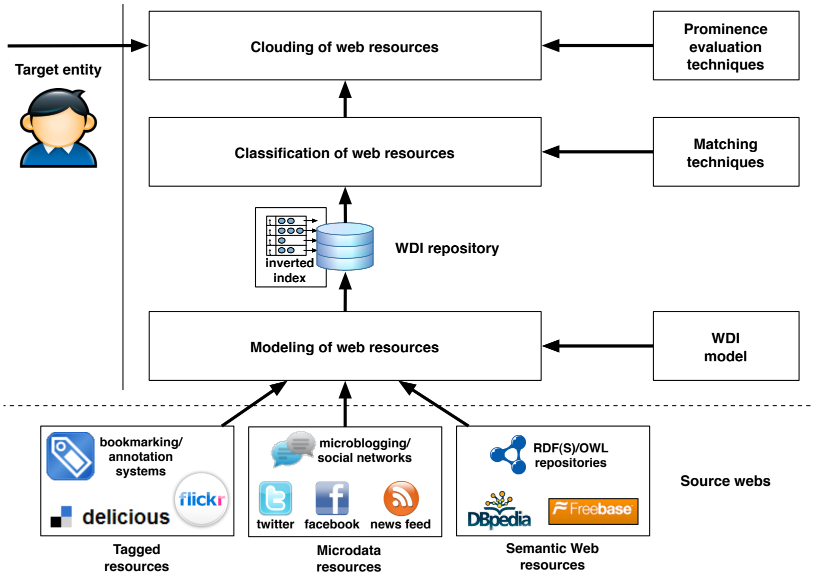


Fig. 1. The semantic clouding approach

Modeling of web resources: WDI model. Building *i*-clouds by mixing up both objective and subjective information about a certain target entity requires the capability to deal with a variety of web resources coming from different webs. For semantic clouding, all the different web resources are acquired from their respective source web and they are stored in a support repository, called *WDI repository*, according to a reference data model, called *WDI model* [6], based on the notion of *web data item* to represent the metadata featuring the various kinds of web resource.

Classification of web resources. Our semantic clouding approach is based on the capability of grouping the web data items on the basis of their closeness. The closeness between two web data items wdi_i and wdi_j captures the level of similarity/semantic relation holding between them and it is represented by a closeness coefficient $cc(wdi_i, wdi_j) \in [0, 1]$, calculated by comparing wdi_i and wdi_j. Such closeness coefficient $cc(wdi_i, wdi_j)$ is calculated for each possible pair of web data items stored in the WDI repository using appropriate matching techniques, and the corresponding values are then used by a hierarchical clustering procedure in order to produce a *closeness tree* where each leaf corresponds to a web data item, and inner nodes denote the closeness coefficient values. To choose the matching techniques to use, we take into account the nature and the different complexity that can characterize the different web resources, and consequently, the corresponding web data items. In [4], we address the problem of matching Semantic Web resources; in [6], we analyze the problem of classifying and comparing microdata; in [7], we provide specific methods and techniques for organizing and matching tags extracted from the Social Web. Moreover, in [5, 12], we present a system for integrating Social and Semantic knowledge in a P2P environment.

Clouding of web resources. The clouding phase is based on the results of the classification activity and aims at constructing the appropriate *i*-cloud organization for a given target entity by prominence and closeness levels. An *i*-cloud is formally defined as an undirected weighted graph $\mathcal{IC}_e = (N, E)$ associated with a target entity e. A node $n_i \in N$ represents a web data item wdi_i relevant for e, while an edge $(n_i, n_j) \in E$ between two nodes n_i and n_j represents the level of closeness between wdi_i and wdi_j. $\mathcal{IC}_e$ is equipped with a labeling function $\rho : N \rightarrow [0, 1]$, that associates each node $n_i \in N$ with a value $p(n_i) \in [0, 1]$, and a labeling function $\sigma : E \rightarrow [0, 1]$, that associates each edge $(n_i, n_j) \in E$ with a value $c(n_i, n_j) \in [0, 1]$. A value $p(n_i)$ denotes the level of *prominence* of the web data item wdi_i in $\mathcal{IC}_e$. A high value of $p(n_i)$ denotes that the web resource corresponding to wdi_i is very relevant for e. Different techniques are possible for the evaluation of the prominence in an *i*-cloud and these techniques can be used alone or in combination. We devise three main categories of techniques for prominence evaluation, namely provenance-base, target-based, and popularity-based techniques. A value $c(n_i, n_j)$ denotes the level of *closeness* between the web data items wdi_i and wdi_j in $\mathcal{IC}_e$. In particular, $c(n_i, n_j)$ is equal to the closeness coefficient $cc(wdi_i, wdi_j)$ calculated in the previous phase.

An example of *i*-cloud is shown in Figure 2, collecting web resources related to the target entity “Star Wars”. We can observe that web 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 labeled with their closeness degree.

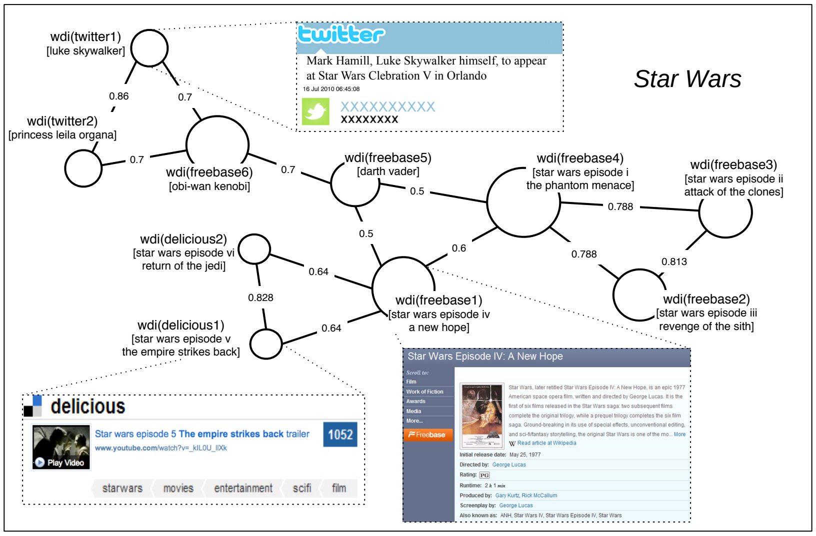
4 Ongoing and future work

We have presented the thesis work we are undergoing for semantic data clouding. Ongoing and future work will be devoted to formally define the properties of *i*-clouds and the operations that can be applied between different *i*-clouds (e.g., selection, projection, join). Furthermore, some preliminary evaluation of our semantic clouding approach has been performed using data extracted from Delicious [8], Twitter [9], and Freebase [10]. *i*-clouds are evaluated on the basis of their level of accuracy and by analyzing the dependency between their size (i.e., the number of web data items) and their cohesion (i.e., the average level of closeness between web data items). The accuracy of an *i*-cloud is defined as its capability to collect web resources which are really relevant with respect to the given target entity, and it depends on the matching techniques that are used for

[8] `http://www.delicious.com`

[9] `http://search.twitter.com`

[10] `http://www.freebase.com`



Mark Hamill, Luke Skywalker himself, to appear at Star Wars Clebration V in Orlando
16 Jul 2010 06:45:08

Star wars episode 5 The empire strikes back trailer
www.youtube.com/watch?v=_kIL0U_llXk

Star Wars Episode IV: A New Hope

Star Wars, later retitled Star Wars Episode IV: A New Hope, is an epic 1977 American space opera film, written and directed by George Lucas. It is the first of six films released in the Star Wars saga: two subsequent films complete the original trilogy, while a prequel trilogy completes the six film saga. Ground-breaking in its use of special effects, unconventional editing, and sci-fi/fantasy storytelling, the original Star Wars is one of the mo... More
W Read article at Wikipedia

Initial release date: May 25, 1977
Directed by: George Lucas
Rating: PG
Runtime: 2 h 1 min
Produced by: Gary Kurtz, Rick McCallum
Screenplay by: George Lucas
Also known as: ANH, Star Wars IV, Star Wars Episode IV, Star Wars

Fig. 2. Example of *i*-cloud for the target entity "Star Wars"

clustering web data items. In order to evaluate the quality of our matching techniques, we exploited the IIMB 2010 dataset [11] and related tools, that are used also for the international instance matching evaluation contest of the Ontology Alignment Evaluation Initiative (OAEI) [12]. The obtained results show that the accuracy of our matching tool HMatch 2.0 is significantly higher than the one of a simple string matching algorithm. The effective applicability of the semantic clouding approach in real application contexts and how it is affected by the number of web data items stored in the WDI repository is also under study.

References

1. Bergamaschi, S., Guerra, F., Orsini, M., Sartori, C., Vincini, M.: RELEVANT News: a Semantic News Feed Aggregator. In: Proc. of the 4th Workshop on Semantic Web Applications and Perspectives (SWAP 2007). Bari, Italy (2007)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5(3) (2009)
3. Bouquet, P., Stoermer, H., Mancioppi, M., Giacomuzzi, D.: OkkaM: Towards a Solution to the "Identity Crisis" on the Semantic Web. In: Proc. of the 3rd Italian Semantic Web Workshop. Pisa, Italy (2006)
4. Castano, S., Ferrara, A., Montanelli, S., Varese, G.: Matching Semantic Web Resources. In: Proc. of the 8th Int. Workshop on Web Semantics (WebS 2009) co-located with the 20th

[11] http://www.instancematching.org/oaei/imei2010.html
[12] http://oaei.ontologymatching.org/2010

Int. Conference on Database and Expert Systems Applications (DEXA 2009). Linz, Austria (2009)

5. Castano, S., Ferrara, A., Montanelli, S., Varese, G.: Semantic Coordination of P2P Collective Intelligence. In: Proc. of the Int. ACM Conference on Management of Emergent Digital EcoSystems (MEDES 2009). Lyon, France (2009)
6. Castano, S., Ferrara, A., Montanelli, S., Varese, G.: Similarity-based Classification of Microdata. In: D'Atri, A., Ferrara, M., George, J., Spagnoletti, P. (eds.) Information Technology and Innovation Trends in Organizations. Springer-Verlag (2010)
7. Castano, S., Varese, G.: Building Collective Intelligence through Folksonomy Coordination. In: Bessis, N., Xhafa, F. (eds.) Next Generation Data Technologies for Collective Computational Intelligence. Springer-Verlag (2011), to appear
8. Hernández, M., Stolfo, S.: The Merge/Purge Problem for Large Databases. In: Proc. of the ACM SIGMOD Int. Conference on Management of Data. San Jose, California, USA (1995)
9. Isaac, A., Van der Meij, L., Schlobach, S., Wang, S.: An Empirical Study of Instance-Based Ontology Matching. In: Proc. of the 6th Int. Semantic Web Conference (ISWC 2007). Busan, Korea (2007)
10. Koutrika, G., Bercovitz, B., Ikeda, R., Kaliszan, F., Liou, H., Zadeh, Z., Garcia-Molina, H.: Social Systems: Can We Do More Than Just Poke Friends? In: Proc. of the 4th Biennial Conference on Innovative Data Systems Research (CIDR 2009). Asilomar, CA, USA (2009)
11. Langegger, A., Wöß, W., Blöchl, M.: A Semantic Web Middleware for Virtual Data Integration on the Web. In: Proc. of the 5th European Semantic Web Conference (ESWC 2008). Tenerife, Spain (2008)
12. Montanelli, S., Castano, S., Ferrara, A., Varese, G.: Managing Collective Intelligence in Semantic Communities of Interest. International Journal of Organizational and Collective Intelligence (IJOCI), Special Issue on Collectively Intelligent Information and Knowledge Management 1(4) (2010)
13. Nikolov, A., Uren, V., Motta, E., Roeck, A.D.: Handling Instance Coreferencing in the KnoFuss Architecture. In: Proc. of the 1st ESWC Int. Workshop on Identity and Reference on the Semantic Web (IRSW 2008). Tenerife, Spain (2008)
14. Radev, D., Otterbacher, J., Winkel, A., Blair-Goldensohn, S.: NewsInEssence: Summarizing Online News Topics. Communications of the ACM 48(10) (2005)
15. Tummarello, G., Delbru, R., Oren, E.: Sindice.com: Weaving the Open Linked Data. In: Proc. of the 6th Int. Semantic Web Conference (ISWC 2007). Busan, South Korea (2007)