MI-Search: a Smart Approach for Urban Information Clouding[1]

Stefano Montanelli and Silvana Castano

Università degli Studi di Milano
DICo - Via Comelico, 39 - 20135 Milano
{stefano.montanelli,silvana.castano}@unimi.it

Abstract. In this paper, we present an approach for *urban-centered* and *calendar-oriented* surfing of web contents according to the personal preferences of a user profile and interests.
Real examples related to the city of Milan are discussed in the paper to illustrate the technical peculiarities of the proposed approach.

Keywords: Urban Information Clouding, Web Content Classification

1 Introduction

The recent innovations in the field of Web 2.0 and Semantic Web have radically changed the way that web contents are surfed and explored. On one side, the growing availability of user-generated contents, like microblogging posts and RSS news, typical of Web 2.0 and Social Web platforms, has posed the question of how to effectively handle and index this huge amount of short and rapidly-obsolescent data [5]. On the other side, the success of the Linked Data paradigm is enforcing the upcoming data-oriented vision of the Semantic Web, in spite of the conventional resource-oriented model [1]. The result is that the existing techniques for web content classification, search, and presentation are actually inadequate to satisfy the user needs in such a pervasive and highly-dynamic scenario.

In this paper, we present an approach for *urban-centered* and *calendar-oriented* surfing of web contents according to the personal preferences of a user profile and interests. Such an approach has been developed in the framework of the MI-Search project co-funded by Regione Lombardia and Fastweb S.p.A..

A distinguishing feature of MI-Search is the capability to go beyond the actual interoperability problems concerned with the capability to exploit traditional web sites and spontaneous user comments/posts in an integrated way. This allows to enable a *cloud-based* web exploration, where all the available information about a topic/event of interest are delivered to the user in a comprehensive, intuitive picture. By *urban-centered*, we mean the MI-Search is tailored to work on the specific scenario of a

[1] This work is funded by Regione Lombardia and Fastweb S.p.A. in the framework of the *Dote in Ricerca* project.

selected city, such as the city of Milan in our case, with the goal to focus the web contents to consider on a selected target. By *calendar-oriented*, we mean that the events and meetings noted in the personal agenda can be exploited to automatically select the web contents that can be suggested as potentially interesting for the final user. Real examples related to the city of Milan are discussed in the paper to illustrate the technical peculiarities of MI-Search.

2 Overview of MI-Search

The MI-Search approach is shown in Figure 1 and it is characterized by the following distinguishing features.

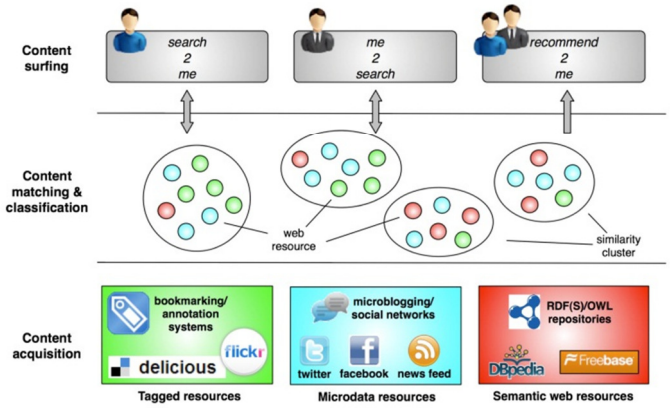

Fig. 1. The MI-Search approach

Capability to consider different kinds of web contents in a seamless way. MI-Search is conceived to deal with contents extracted from different kinds of web resources. In particular, in MI-Search, we distinguish three different kinds of web resources, that are *tagged resources*, *microdata resources*, and *semantic web resources* (see the bottom part of Figure 1). Tagged resources are traditional web resources (i.e., web pages) and they are characterized by a raw structure with few metadata. Microdata resources are posts/comments coming from news feeds and microblogging systems (e.g., Facebook, Twitter posts). A microdata resource is characterized by a short textual content and a set of metadata/properties, like title, author, and creation date, that are commonly employed to describe publishing items. Semantic web resources are instances/individuals coming from RDF(S) knowledge repositories and OWL ontologies and they are characterized by a structured description composed of a set of assertions denoting its specification in the web document of origin. MI-Search successfully supports content interoperability by providing a support repository where all the considered contents are stored according to a reference data model developed in the project.

Capability to perform a similarity-based aggregation of web contents. The web contents stored in the support repository are submitted to a matching and classification process where contents referring to the same argument are first detected and then aggregated in *similarity clusters* (see the middle part of Figure 1). A similarity cluster is defined to collect web resources that can have a different nature, but are similar in content. In other words, a similarity cluster represents a specific argument and it contains all the web resources, either tagged, microdata, and semantic web resource, that refer to that argument.

Capability to tailor the contents to deliver according to the user profile and interests. The similarity clusters are exploited by the final users during their content surfing activities. By relying on the user profile/interests, the similarity clusters that are interesting for the user are selected. This way, MI-Search succeeds in tailoring the most appropriate information and/or the suggestions to deliver to the user according to the specific scenario that is currently enforced (see the top part of Figure 1).

The MI-Search project distinguishes two different kinds of user categories: the *personal users* and the *business users*. Personal users are users interested in receiving information and they exploit the MI-Search technology for obtaining contents and suggestions about their personal interests and events in the agenda. Business users are users interested in public events and other possible situations that are suitable for promoting their business activities. In this respect, the following three main scenarios have been envisaged in MI-Search:

- *Search-2-me scenario*. This is the typical scenario of personal users and it is triggered when new personal events are planned by the user in the agenda. By exploiting the user agenda, the MI-Search technology discovers the user interests and it can provide a complete set of information about a planned event. In particular, MI-Search retrieves spontaneous information and user-generated contents related to the considered event, like comments from other similar users and special user offers joint with the participation to the event. As an example, we consider an art-exhibition event about the singer Fabrizio de André located in Milan at Rotonda della Besana. The user plans to visit this exhibition and a personal event is inserted in the agenda for a certain date. Through specialized websites (e.g., http://www.fabriziodeandrelamostra.com), MI-Search automatically provides to the user all the available information about the exhibition and about the singer. Moreover, other information are extracted by the MI-Search technology from social networks (e.g., Facebook[2], Twitter[3]) to provide comments of other users that previously visited the exhibition.
- *Me-2-search scenario*. This is the typical scenario of business users and it is triggered by the user when she/he start browsing the available suggestions that the system provides as potentially interesting opportunities for promoting the

[2] http://www.facebook.com/.
[3] http://www.twitter.com/.

user business. The user can browse a suggestion list of public events that can be interesting from the business point of view and she/he can decide to insert in the system a new business offer joint with a suggested event. For example, we consider a business user that has a sushi restaurant located in Milan, Viale Montenero (near to Rotonda della Besana). When the user starts browsing the possible suggestions, the art exhibition about Fabrizio de André at Rotonda della Besana is retrieved (due to a geo-locality proximity). The business user can decide to insert in the system a special menu price for the exhibition visitors. Such an offer will be linked to the art exhibition event and it will be visualized by personal users that plan to visit the exhibition.

- *Recommend-2-me scenario*. This is a basic scenario of personal users and it is permanently active without requiring any triggering event. The recommend-2-me scenario is based on the user interests expressed in the personal profile to suggest events and/or (promotional) initiatives that can be potentially interesting. In this scenario, the user periodically receives a report with a list of upcoming events, either public and business events, that match her/his preferences for possible selection (and subsequent insertion in the personal agenda). As an example, we consider a personal user who specified an interest for sushi restaurants in her/his profile. Receiving the periodic report of interesting upcoming events, the user becomes aware of the special menu price of the sushi restaurant in Viale Montenero joint with the art exhibition about Fabrizio de André. The user can decide to visit the exhibition with the goal to subsequently take advantage of the special sushi offer. A personal event is inserted in the user agenda to plan the visit and to receive further information about the event (see the search-2-me scenario).

3 The MI-Search techniques

In the following, we discuss some technical details of MI-Search with special reference to those aspects of the project that are concerned with interoperability issues. In particular, *web content acquisition* and *web resource matching and classification* of MI-Search will be presented.

3.1 Web content acquisition

MI-Search is based on a support repository called MI-Search-DB capable of storing all the different kinds of web contents considered in the project through a uniform representation. The representation of specific features for event localization, such as spatial/temporal coordinates, is also enforced in MI-Search-DB. The repository is implemented as a PostgreSQL relational database, whose ER schema is shown in Figure 2. In the schema, we note that any kind of considered web content is represented through the entity Web Content. Web contents are distinguished in events (entity Event) and resources (entity Resource).

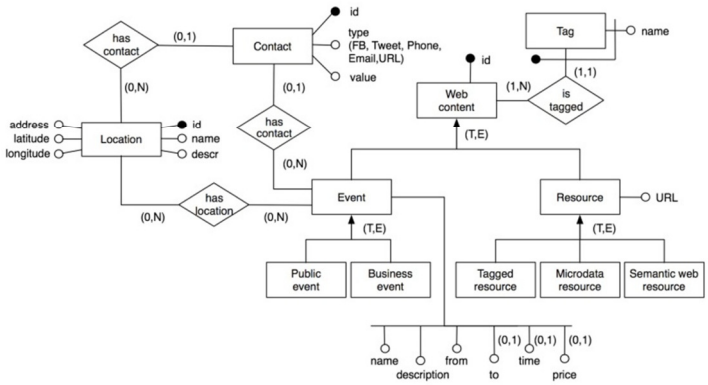

Fig. 2. The schema of the MI-Search-DB repository for web content acquisition

Event. Events are classified in public events (entity Public Event), that represent official initiatives like art exhibitions or concerts, and business events (entity Business Event) that represent commercial initiatives inserted by business users. An event is characterized by attributes that describe its temporal frame (i.e., from-date, to-date, time, and frequency) and other features, like description and price (where needed). The entity Event is associated with the entities Contact and Location to represent the different contact-points for the event (e.g., Phone, Facebook page, Twitter channel) and the geo-coordinates where the event takes place, respectively.

Resource. Resources are web contents acquired from outside the system and they distinguished in Tagged Resource, Microdata Resource, and Semantic Web Resource as discussed in Section 2.

Tag. Each web content, either event or resource, is associated with a set of tags (entity Tag) denoting the keywords that mostly characterize the event/resource. For an event, the set of tags can be automatically extracted from one or more reference website. This usually happens with public events. Otherwise, tags can be manually inserted by the user that inserts the event. This usually happens with business events. For a resource, the set of tags is automatically extracted from the resource content itself. In a tagged resource, tags are extracted from bookmarking and social annotation systems (e.g., Delicious, Flickr). In a microdata resources, tags are extracted from the textual resource content and from other available metadata/properties, like the title. In a semantic web resource, tags are extracted from literals, property names, and property values contained in the RDF/OWL assertions of the resource specification. We note that, before insertion in the entity Tag, a tag is submitted to a normalization procedure for word-lemma extraction and for compound-term tokenization [4,7].

Example. In Figure 3, we consider two examples of acquired web contents.

Fig. 3. Examples of web resource acquisition

Figure 3(a) shows a RSS post published on a well-known electronic wall about events planned in the city of Milan (http://blog.milano-italia.it/). This is an example of public event related to the art-exhibition about Fabrizio de Andrè located at at Rotonda della Besana. Besides the featuring attributes expected in MI-Search-DB for a public event, contact and location information are also provided. Figure 3(b) shows a comment posted on the Facebook social network published by a user that visited the art-exhibition about Fabrizio de André. This is an example of microdata resource featured by its URL on the web as expected in MI-Search-DB. Moreover, either the public event and the microdata resource, are associated with a set of tags automatically extracted from the two posts as a sort of synthetic characterization of each web contents.

3.2 Web content matching and classification

The goal of matching and classification in MI-Search is to detect and build the similarity clusters to use for content delivery to the final users.

Content matching. This step has the goal to evaluate the degree of similarity between each pair of web contents stored in the MI-Search-DB. Given two web contents wc_i and wc_j, the *similarity coefficient* $\sigma(wc_i,wc_j) \in [0,1]$ denotes the level of similarity of wc_i and wc_j based on their commons tags. We define $Tag_{wc} = \{tag_1, \ldots, tag_m\}$ as the set of tags associated with the web content wc in MI-Search-DB. The similarity coefficient $\sigma(wc_i,wc_j)$ is calculated as follows:

$$\sigma(wc_i, wc_j) = \frac{2 * |tag_x \sim tag_y|}{|Tag_{wc_i}| + |Tag_{wc_j}|}$$

where $tag_x \sim tag_y$ denotes that $tag_x \in Tag_{wci}$ and $tag_y \in Tag_{wcj}$ are matching tags according to a string matching metric that considers the structure of tag_x and tag_y. For σ calculation, we employ our matching system HMatch 2.0, where state-of-the-art metrics for string matching (e.g., I-Sub, Q-Gram, Edit-Distance, and Jaro-Winkler) are implemented [2].

Content classification. Similarity clusters are built by relying on a clique percolation method (CPM) [6]. This method receives in input a graph G where nodes are the web contents stored in the MI-Search-DB repository and edges are established between any pair (wc_i, wc_j) of similar contents for which $\sigma(wc_i,wc_j) \geq th$ ($th \in (0,1]$ is a matching threshold denoting the minimum level of similarity required to consider two web contents as matching contents). The CPM returns a set of similarity clusters where each cluster collects a region of nodes in G that are more densely connected to each other than to the nodes outside the region. The CPM is based on the notion of *k-clique* which corresponds to a complete (fully-connected) sub-graph of k nodes within the graph G_s^+. Two k-cliques are defined as *adjacent k-cliques* if they share $k - 1$ nodes. The CPM determines clusters from k-cliques. In particular, a cluster, or more precisely, a k-clique-cluster, is defined as the union of all k-cliques that can be reached from each other through a series of adjacent k-cliques. More technical details about the CPM and the construction of similarity clusters can be found in [3].

Example. We consider the example shown in Figure 3. We call wc_1 the public event of Figure 3(a) and wc_2 the microdata resource of Figure 3(b). The similarity coefficient of wc_1 and wc_2 is $\sigma(wc_1,wc_2) = 0.35$ due to the matching tags in Tag_{wc1} and Tag_{wc2}. With a matching threshold $th = 0.3$, the web contents wc_1 and wc_2 are considered as matching contents and an edge (wc_1, wc_2) is set in the graph G that is passed to the CPM method for calculation of the similarity clusters. An example of similarity cluster is shown in Figure 4. Besides the web contents wc_1 and wc_2, the cluster of Figure 4 contains a Flickr image taken from the exhibition (tagged resource), another Facebook user comment (microdata resource), the Freebase page about Fabrizio de André (semantic web resource), and the contact information of the Aoyama restaurant, a sushi restaurant that published in MI-Search a discounted dinner offer associated with the art-exhibition at Rotonda della Besana (i.e., business event). Such a cluster will be exploited by the delivery services of MI-Search when a request about the Fabrizio de André art-exhibition is submitted by a user.

4 Concluding remarks

In this paper, we presented the main features of the MI-Search project for urban-centered and calendar-oriented surfing of web contents. Technical issues about web content acquisition, matching, and classification as well as real examples applied to the city of Milan are also discussed.

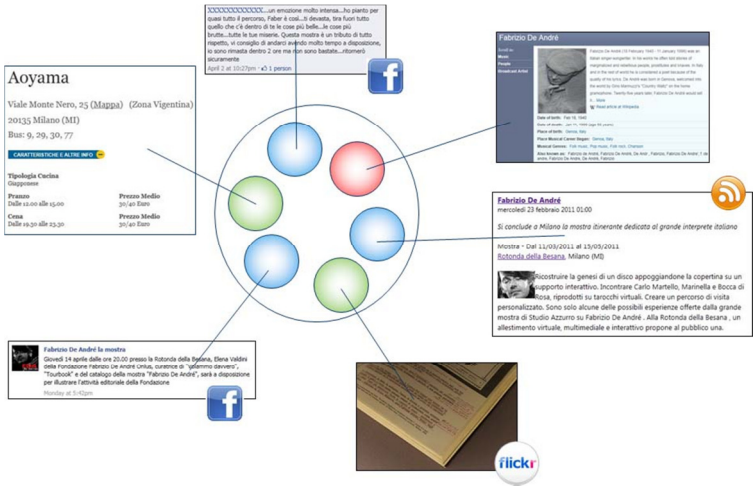

Fig. 4. An example of similarity cluster

Ongoing research work is devoted to study the problem of periodically refreshing the contents of the MI-Search-DB repository and to complete the acquisition of a dataset about the city of Milan to be employed for experimentation. Moreover, matching techniques combining both string-based techniques and position-based techniques are currently under development as well as techniques for content delivery based on similarity cluster exploitation. Finally, next-future activities will be focused on the development of a mobile prototype based on the presented ideas.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. Int. Journal on Semantic Web and Information Systems 5(3) (2009)
2. Castano, S., Ferrara, A., Montanelli, S.: Matching Ontologies in Open Networked Systems: Techniques and Applications. Journal on Data Semantics V (2006)
3. Castano, S., Ferrara, A., Montanelli, S.: Thematic Exploration of Linked Data. In: Proc. of the 1st VLDB Int. Workshop on Searching and Integrating New Web Data Sources (VLDS 2011). Seattle, USA (2011)
4. Castano, S., Varese, G.: Next Generation Data Technologies for Collective Computational Intelligence, chap. Building Collective Intelligence through Folksonomy Coordination. Springer (2011)
5. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press (2010)
6. Palla, G., Derènyi, I., Farkas, I., Vicsek, T.: Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature 435 (2005)
7. Sorrentino, S., et al.: Schema Normalization for Improving Schema Matching. In: Proc. of the 28th Int. ER Conference. Gramado, Brazil (2009)