Urban Information Integration in MI-Search: Results and Future Research Activities*

Stefano Montanelli, Lorenzo Genta, and Silvana Castano

Università degli Studi di Milano
DI - Via Comelico, 39 - 20135 Milano

{stefano.montanelli,lorenzo.genta,silvana.castano}@unimi.it

Abstract. In this paper, we present the main achievements of the MI-Search project for "multi-web" information integration around topics relevant for urban users like for example city events and points of interest. In particular, we discuss the results of our experimental evaluation over a considered case study about the city of Milan as well as ongoing/future research activities in the framework of MI-Search.

1 Introduction

The recent success of mobile urban applications like point-of-interest exploration apps and thematic event-publishing walls has produced a new attention on information integration issues in pervasive and highly-dynamic scenarios. Existing tools in this field are mainly focused on exploiting conventional geo-local information extracted from pre-organized integrated maps [5,8,9]. User-generated contents taken from Social Web platforms (e.g., microblogging posts, RSS news) and/or semantic web data taken from Linked Data repositories (e.g., Freebase, DBpedia) are mostly ignored by such a kind of mobile applications. The MI-Search project aims at providing the capability to dynamically mix up and integrate "multi-web" information around topics relevant for urban users like for example city events and points of interest [6].

In this paper, we present the main achievements of the MI-Search project. First, we overview the MI-Search solutions for *urban-oriented*, *event-centric* surfing of web contents (Section 2). Then, we discuss the results of our experimental evaluation over a considered case study about the city of Milan (Section 3). In particular, the experimentation is aimed at assessing the effectiveness of the MI-Search techniques in retrieving pertinent and integrated information about a given event of interest. A discussion about the MI-Search effectiveness from the scalability point of view is also provided. Finally, we outline ongoing/future research activities in the framework of MI-Search (Section 4) and we provide our concluding remarks (Section 5).

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

2 Overview of MI-Search

MI-Search is featured by an approach for urban information integration based on the notions of *smart city view* and *similarity cluster* (see Figure 1).

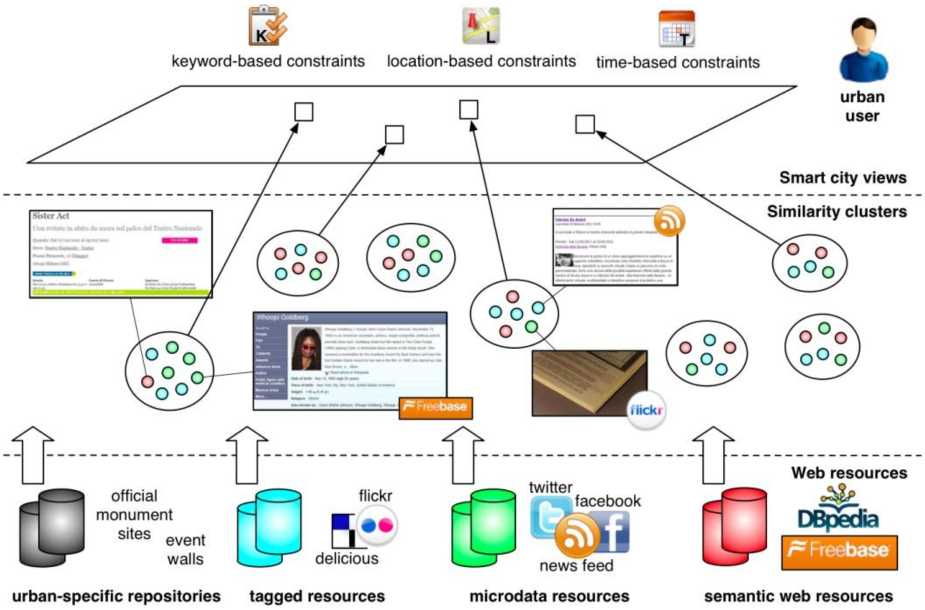

Fig. 1. The MI-Search approach for urban information integration

By **smart city view**, we mean a set of similarity clusters about the events of a considered urban space that satisfies the *selection criterion* specified by an interested urban user, like a citizen, a tourist, or a business agent. The idea of smart city views is defined in MI-Search to enforce on-the-fly filtering operations over the available data about the events of a considered urban space (e.g., a city, a metropolitan area, a region). The selection criterion consists in one or more user requirements featuring keyword-, location-, and/or time-based constraints to use for choosing the pertinent similarity clusters. The creation of a smart city view can be manually triggered by an interested user that specifies a set of keywords. In a more realistic scenario, smart city views are automatically generated by MI-Search to show potentially interesting urban events based on the current user position or a planned appointment in the personal user agenda. For instance, consider an user that plans to attend the musical Sister Act at Teatro Nazionale, Milan, Italy on September, 20th 2012 8-11PM. When an entry is inserted in the user personal agenda, MI-Search generates a smart city view featured by pertinent similarity clusters that contain events/web resources concerned with Sister Act, nearby Teatro Nazionale, in a time period compatible with September, 20th 2012 8-11PM. The user can explore the cluster contents to discover additional events (e.g., a special menu offer for musical attendants by a restaurant next to Teatro Nazionale) or useful information of interest (e.g., curiosities about the Sister Act musical).

A **similarity cluster** is built by collecting web resources that can have a different nature (e.g., official sites, event walls, user posts/comments), but are similar in content (e.g., they refer to the same event, such as an art exhibition). Each similarity cluster is characterized by a set of keyword-, location-, and time-based descriptors that are exploited by MI-Search for matching with the user requirements to generate smart city views.

The approach for urban information integration of MI-Search has the goal to define a set of similarity clusters through the execution of three main phases, that are *web content acquisition*, *web content matching*, and *web content classification*. In the following, we provide a summary overview of the MI-Search approach for urban information integration. Further technical details can be found in [4,6].

Web content acquisition. MI-Search is based on a support repository called MI-Search-DB capable of storing through a uniform representation all the different kinds of web contents considered for advertisement of urban events. In MI-Search-DB, web contents are distinguished in *events* that represent official initiatives like art exhibitions or concerts and other *resources* that are acquired from web and then classified in Tagged Resource, Microdata Resource, and Semantic Web Resource. An event is acquired from electronic publishing walls and they are 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). Events are also associated with information about contact-points (e.g., Phone, Facebook page, Twitter channel) and geo-coordinates where the event takes place, respectively. 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 of the web-of-data (e.g., Freebase, DBpedia). These resources are characterized by a structured description composed of a set of assertions denoting their specification in the web document of origin.

Each web content, either event or resource, is associated with a set of tags denoting the keywords that more prominently characterize the event/resource.

Web 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 common 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:

$$\sigma(wc_i, wc_j) = \frac{2 \cdot |\{< tag_x, tag_y >: tag_x \sim tag_y\}|}{|Tag^{wc_i}| + |Tag^{wc_j}|}$$

where $tag_x \sim tag_y$ denotes that $tag_x \in Tag^{wc_i}$ and $tag_y \in Tag^{wc_j}$ are matching tags according to a string matching metric that considers the syntax 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 [3].

Web content classification. Similarity clusters are built by relying on a clique percolation method (CPM) [7]. 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 $\sigma(wc_i,wc_j) \geq th_m \in (0,1]$, where th_m 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.

3 Experimental results

An experimental evaluation has been performed to assess the effectiveness of the MI-Search approach for urban information integration. To this end, two datasets called MI-DS-focused and MI-DS-large have been defined for experimentation. These datasets are built by exploiting well-known publishing walls related to events about the city of Milan, Italy. In particular, MI-DS-focused is fully based on the http://www.milanodabere.it/ publishing wall, while MI-DS-large stores events extracted from heterogeneous sources (e.g., http://www.milanodabere.it/, http://eventi-milano.it/, http://www.eventiesagre.it/). The two datasets mainly differ in the number of stored events (i.e., 134 events in MI-DS-focused vs. 253 events in MI-DS-large) and in the number of tags associated with events (i.e., 1115 tags in MI-DS-focused vs. 247 tags in MI-DS-large). The choice of having two datasets storing a strongly different number of events is motivated by the idea to provide a basic measurement of scalability performance of MI-Search (see below). Moreover, we note that tags are associated with events in two modalities: i) they are extracted from the considered publishing walls by exploiting predefined wall categories (e.g., art-exhibition, entertainment, theater), and ii) they are manually inserted by the wall staff. The difference of the two datasets in the number of tags depends on the fact that most of the existing publishing walls about Milan provide a poor event categorization, and thus events are associated to a small number of tags (usually only one in MI-DS-large). The publishing wall http://www.milanodabere.it/ (that has been exploited to generate MI-DS-focused) provides a more accurate event categorization that is also coupled with a manual event annotation by the wall staff. As a consequence, in MI-DS-focused, each event is associated with 8 tags on average ($1115/134 \approx 8$) that are appropriate to ensure interesting matching results.

Capability of MI-Search to retrieve pertinent events. This experiment aims at evaluating the quality of the similarity clusters on top of which smart city views are built. In particular, given a dataset of web resources, we apply matching and classification techniques and we analyze the resulting similarity clusters to assess whether the events therein contained have been correctly clustered. In other words, we are interested in measuring the capability of the matching techniques of MI-Search in detecting similarities among events and web contents. Besides the basic similarities that can be detected through the conventional search functionalities of publishing walls, we want to evaluate the capability of MI-Search to discover non-trivial mappings among the dataset elements. The MI-DS-focused dataset has been employed in this experiment due to the high number of tags-per-event that characterizes this dataset. In this experiment, we used the event classification of http://www.milanodabere.it/ as baseline where two events are set to be similar if they are placed in the same category by the wall staff. We executed matching over the events of MI-DS-focused using HMatch 2.0 and we analyzed the overlap between the mappings detected by HMatch 2.0 and the baseline. We observed that with low values of matching threshold (e.g., $0.2 \leq th_m \leq 0.5$) the MI-Search techniques are capable of detecting most of the mappings in the baseline (recall ≈ 80%). With higher values of similarity threshold (e.g., $0.6 \leq th_m \leq 1.0$), recall decreases but interesting values of precision are obtained (precision ≈ 90%). Furthermore, we note that a number of mappings that are not contained in the baseline are found by HMatch 2.0. In some cases, these additional mappings are false positives and they cause a precision decrease. This side-effect mostly depends on the quality of the tags associated with the events in the dataset. Tags are manually provided by the wall staff and usually they lack of accuracy in the sense that they are too much generic for actually describing the event features. However, in some other cases, these additional mappings represent non-trivial similarity mappings between pairs of events that are somehow related. The pair of events in Figure 2 is an example of non-trivial mapping between two events differently categorized in the baseline but actually similar since they refer to events about the same historical character (i.e., Gian Giacomo Poldi Pezzoli)[1].

Appetizer at Poldi Pezzoli, Milan (EventID:267)
Il Museo Poldi Pezzoli propone un aperitivo, orchestrato dal Ristorante Don Lisander, in tandem con l'attuale mostra dedicata a Gian Giacomo Poldi Pezzoli, noto collezionista risorgimentale, oltre che uno dei protagonisti delle Cinque Giornate di Milano…
tag: appetizer, centre, historical, Manzoni, Milan, muse, Pezzoli, Poldi, street…

(a)

Exhibition Gian Giacomo Poldi Pezzoli, Milan (eventID:195)
Nobiluomo colto e raffinato, Gian Giacomo Poldi Pezzoli fu uno dei protagonisti delle Cinque Giornate di Milano. Una mostra allestita nelle stesse sale in cui il collezionista visse e lavorò, ne ricorda passioni e impegno civico…
tag: centre, exhibition, Giacomo, Gian, historical, house, Manzoni, Milan, muse, painting, Pezzoli, Poldi, Risorgimento, street…

(b)

Fig. 2. Example of a non-trivial mapping detected by HMatch 2.0 in the MI-DS-focused dataset

[1] In this example, the web contents are left in the original Italian language while tags and other metadata have been translated into English for reader convenience.

The example of Figure 2 represents the positive impact of using similarity matching techniques in the construction of event clusters. This result further highlights that the choice of appropriate tags for describing web contents is a key aspect for enforcing an effective event classification (see Section 4 for further details about this topic).

Scalability performance of MI-Search. This experiment aims at evaluating the scalability of MI-Search when the number of elements to consider for clustering increases. In this respect, the scalability performances of MI-Search mostly depend on the efficiency of the CPM techniques employed for generating similarity clusters. The MI-DS-large dataset has been exploited in this experiment by executing CPM with a progressively-increasing number of considered events stored in MI-DS-large. Furthermore, we also executed CPM by varying the matching threshold th_m used for detecting similar web contents. This way, we can observe the scalability performances of MI-Search on change of the degree of interconnection for the graph G used by CPM (see Table 1).

Table 1. Scalability results obtained with the MI-DS-large dataset

#nodes	th_m=0.4	th_m=0.6	th_m=0.9	th_m=0.97
30	~250ms	~50ms	~25ms	~25ms
40	~1s	~130ms	~30ms	~25ms
50	~6s	~200ms	~30ms	~25ms
100	~800s	~11s	~50ms	~30ms
250	∞	~800s	~200ms	~40ms

In the results, we note that CPM scales very well with high values of matching threshold $th_m \geq 0.9$. Performances become critical when low/intermediate values of th_m are employed and more than 100 nodes belong to the graph G. In general, an intermediate value of matching threshold (e.g., $0.55 \leq th_m \leq 0.65$) is suggested to ensure a good trade-off between precision and recall (see the experiment above). For this reason, we observe that CPM can be suitably employed when small urban spaces are considered (with less than 100 nodes to consider at a time). For larger urban spaces, clustering solutions more efficient than CPM need to be enforced (see Section 4 for further details about this topic).

4 Ongoing and future research activities

The following research activities about MI-Search are ongoing and/or planned in the next future.

Periodic refresh of similarity clusters. The similarity clusters of MI-Search need to be periodically updated to refresh the information about existing events and to include new events in the system. On this topic, two different research activities are planned. On one side, we are working on a strategy for the incremental, on-the-fly update of the similarity clusters. The basic idea is to refresh the acquisition of each stored web content and to compare past and present descriptions to detect possible changes. If the tag descriptions are changed, we consider to re-place the web content in a different similarity cluster by evaluating the degree of similarity between the new associated tags and the cluster descriptors. This strategy is adequate for contents characterized by a low obsolescence such as semantic web resources. On the other side, we plan to work on a strategy for the batch, from-scratch reconstruction of the entire set of similarity clusters. The basic idea is to start a new session of acquisition, matching, and classification when the current similarity clusters are becoming obsolete. This strategy is adequate for contents characterized by a high obsolescence such as event descriptions on electronic walls and tagged/microdata resources. We also note that these two strategies can be used in combination to reduce the overall computational effort. For instance, the incremental, on-the-fly refresh of existing clusters can be employed for rapid update of the current classification, while the batch, from-scratch reconstruction of similarity clusters can be executed when a sufficient number of new events are found to be included in the MI-Search system.

Scalability of content classification techniques. The CPM techniques are inadequate for cluster aggregation when a high number of web contents needs to be managed. To overcome this limitation, we aim to equip MI-Search with a suite of different aggregation methods to be dynamically activated according to the number of web contents stored in the MI-Search-DB. On one side, we plan to investigate the use of hierarchical clustering techniques [2]. This choice has a twofold motivation. First, the complexity of the algorithm is quadratic in the worst case under the assumption that an agglomerative strategy is adopted. Second, agglomerative hierarchical clustering enforces a bottom-up approach which allows to stop the cluster computation once that a desired level of aggregation is obtained. As a result, we argue that the adoption of hierarchical clustering techniques enables to improve the content classification techniques of MI-Search in terms of efficiency and flexibility at the same time. On the other side, we plan to study supervised clustering techniques with predefined seeds [1]. This kind of clustering techniques are based on a predefined set of cluster representatives around which all the other elements are then aggregated. In MI-Search, cluster representatives are urban events taken from publishing walls, while other web contents like tagged, microdata, and semantic web resources are the elements to be aggregated with events according to similarity coefficients. We argue that, the use of supervised clustering techniques allows to further improve scalability performances of web content classification due to the fact that the identification of cluster representatives is immediate and similarity coefficients can be exploited for choosing the most appropriate cluster to place the other contents.

Extraction of effective tag descriptions for web contents. For the web contents of the MI-Search-DB, the set of associated tags are extracted from the websites of origin. Usually, manually inserted tags and/or basic event categorizations of the electronic walls are exploited to derive the tag descriptions to use for populating MI-Search-DB. From the experimentation, we observe that this strategy provides a low number of associated tags, which are insufficient to effectively support the matching operations. For this reason, we plan to integrate the current tag-extraction techniques with tag-mining techniques derived from the literature on information retrieval. We set a threshold that expresses the minimum number of tags to be associated with each web content stored in the MI-Search-DB. Thus, tag-mining techniques are invoked to complement the results of tag-extraction techniques when the required number of extracted tags is not reached. State-of-the-art techniques in the field of text analysis will be employed to this end, such as stop-word dropping, term tokenization/normalization, and word stemming/lemmatization.

5 Concluding remarks

In this paper, we presented the main features and results of the MI-Search project for the construction of smart city views based on web-content integration techniques. Ongoing and future research activities in the framework of the MI-Search project have been also outlined. The design of a complete MI-Search prototype for the Android platform is currently under development. Further details are available at http://islab.di.unimi.it/misearch/.

References

1. Basu, S., Banerjee, A., Mooney, R.: Semi-supervised Clustering by Seeding. In: Proc. of the 19th Int. Conference on Machine Learning (ICML 2002). Sydney, Australia (2002)
2. Castano, S., De Antonellis, V., De Capitani Di Vimercati, S.: Global Viewing of Heterogeneous Data Sources. IEEE Transactions on Knowledge and Data Engineering 13(2) (2001)
3. Castano, S., Ferrara, A., Montanelli, S.: Matching Ontologies in Open Networked Systems: Techniques and Applications. Journal on Data Semantics V (2006)
4. 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)
5. Kanhere, S.S.: Participatory Sensing: Crowdsourcing Data from Mobile Smartphones in Urban Spaces. In: Proc. of the 12th IEEE Int. Conference on Mobile Data Management (MDM 2011). Luleå, Sweden (2011)
6. Montanelli, S., Castano, S.: MI-Search: a Smart Approach for Urban Information Clouding. In: Proc. of the 4th Interop-Vlab.it Workshop. Rome, Italy (2011)
7. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature 435 (2005)
8. Papadopoulos, S., Zigkolis, C., Kompatsiaris, Y., Vakali, A.: Cluster-based Landmark and Event Detection on Tagged Photo Collections. IEEE Multimedia Magazine 18(1) (2011)
9. Schmeiß, D., Scherp, A., Staab, S.: Integrated Mobile Visualization and Interaction of Events and POIs. In: Proc. of the 18th Int. ACM Conference on Multimedia. Firenze, Italy (2010)