Exploration of Cultural Heritage Information via Textual Search Queries

Liliana Ardissono
Maurizio Lucenteforte
University of Torino
Torino, 10149 Italy
{liliana.ardissono,
maurizio.lucenteforte}@unito.it

Noemi Mauro
Adriano Savoca
University of Torino
Torino, 10149 Italy
{noemi.mauro,
adriano.savoca}@edu.unito.it

Angioletta Voghera
Luigi La Riccia
Politecnico di Torino
Inter-University Department of Regional and Urban Studies and Planning
Torino, 10125 Italy
{angioletta.voghera,
luigi.lariccia}@polito.it

Abstract
Searching information in a Geographical Information System (GIS) usually imposes that users explore pre-compiled category catalogs and select the types of information they are looking for. Unfortunately, that approach is challenging because it forces people to adhere to a conceptualization of the information space that might be different from their own. In order to address this issue, we propose to support textual search as the basic interaction model, exploiting linguistic information, together with category exploration, for query interpretation and expansion. This paper describes our model and its adoption in the OnToMap Participatory GIS.

Author Keywords
GIS, text-based information search, community maps.

ACM Classification Keywords
H.3.3. Information Search and Retrieval. H.3.5. Online Information Services. H.4.3. Organizational Impacts.

MobileHCI '16 Adjunct, September 06-09, 2016, Florence, Italy
ACM 978-1-4503-4413-5/16/09.
http://dx.doi.org/10.1145/2957265.2962648

Introduction
The protection and promotion of Cultural Heritage (CH) has become a core aspect of European and international cultural policies in the last decades: this attention regards not only tangible but also intangible elements (UNESCO 2003). Various factors challenge the development of ICT tools supporting the knowledge and fruition of CH. On the one

hand, its intrinsically dynamic nature demands for an information sharing model facilitating the continuous publication and revision of knowledge items. On the other hand, the wide variety of available resources raises the need to help people explore the territory in a holistic way, considering not only artworks and places to visit, but also the services supporting their fruition, exposing users to a possibly relevant information overload. This situation suggests a convergence of mobile guides, information sharing services and search engines within a unified framework extending GIS with intelligent information search and crowdsourcing.

Our work aims at enhancing the interaction with the user during the exploration of a geographical information space. Our goals are to let her/him specify information needs in a natural way and to make the system capable of flexibly interpreting search queries. For this purpose, we use a semantic representation of geographic information, following the Geospatial Semantic Web approach (Janowicz, 2012). Specifically, we propose a Natural Language interaction model to allow people to express textual queries for multi-faceted information search. We applied our model to extend the information search functions offered by the OnToMap Participatory GIS (Voghera et al., 2015; Voghera et al., 2016). OnToMap supports information sharing for Participatory Decision-Making processes. It employs Natural Language Processing (NLP) techniques to extract geographic focus and relevant concepts from search queries; then it matches the extracted concepts to those of the domain ontology underlying it in order to identify the information to visualize in the map. Different from previous work on query expansion, which extends search queries with linguistic knowledge and can incur in word sense disambiguation problems, we annotate the system ontology with linguistic knowledge targeted to the represented domain. This prevents the introduction of wrong meanings in the description of the domain, which can be checked at system set-up time, and makes it possible to directly match search queries to domain concepts, without searching for alternative meaning of words.

In the following, we provide a background about GIS and geographical information search. Then we describe the OnToMap project and our proposed approach for geographical search management. Later on, we compare our approach to some related work and, at last, we conclude the paper with a summary and an outline of our future work.

Background

The fruition of cultural heritage attracted the research of the HCI community starting from the early '80ies. Since that time, several projects were devoted to the digitalization of artworks (e.g., Europeana Collections – www.europeana.it, and Virtual Reality applications), and to the development of applications supporting information search. Moreover, mobile guides emerged for the exploration of physical sites such as exhibitions, historical places, museums, etc.; see (Ardissono et al., 2012). At the same time, GIS evolved from isolated applications to Participatory GIS - PGIS (Dunn, 2007): these are web-based applications that support the sharing of geographical information, leveraging a visual paradigm for the presentation of data in dynamic maps. The revolution brought by this trend is traditionally associated with the term VGI (Volunteered Geographic Information, see (Goodchild, 2007)), by comparing the human being to an "intelligent, mobile sensor" able to acquire geospatial information in both a spatial and temporal dimension. Another successful concept widely used in GIS literature is “geo-crowdsourcing” (Goetz and Zipf 2013) which outlines work, involving the collection of geospatial information, performed by a network of people. Even though VGI and geo-crowdsourcing have slightly different meanings, they are usually treated as synonyms or combined (Sui et al. 2003).

Driven by the recent trends in GIS, including VGI and crowdsourcing/geo-crowdsourcing, and fostered by continuous technological advances, the collection and dissemination of geospatial information by ordinary people has become commonplace. Applications involving user-generated geospatial content show diversified patterns in terms of incentive, type and level of participation, purpose of the activity, data/metadata provided and data quality. In this sense, geographical maps have proven to support an intuitive and expressive representation of the territory leveraging the geographical position to help users orientate themselves during the exploration of the information space. *Community maps* are a visual representation of geographical information concerning the values, resources, territorial identity and describe different points of view, each one representing a specific lens on the overall situation. However, considering the contribution of human and automated information sources to content generation, a very large amount of data can be available. Thus, it is important to understand the user’s information needs in order to filter out irrelevant information.

PGIS, and their convergence with community maps, are powerful tools for the knowledge and the fruition of Cultural Heritage. First, they base the presentation of information on a visual approach that, thanks to data geo-localization, enhances user orientation and comprehension. Second, they support information diffusion using the internet as an economic, pervasive communication channel, by supporting information access from both desktop and mobile settings. However, during an information-seeking task, PGIS base the interaction with the user on the selection of pre-compiled information categories. We claim that this approach limits the user’s capability of expressing her/his interests for various reasons. For instance, it forces her/him to adhere to the domain conceptualization and terminology adopted in the system, which might not be immediately understandable, especially as far as technical data is concerned. Moreover, it exposes the user to the browsing of a possibly complex list or network of concepts. Furthermore, it fails to support the creation of very specific queries for restricting the set of relevant items. For instance, as queries can only be composed of a set of relevant concepts, more selective search criteria, which refer to the attributes of items (e.g., “*Baroque* buildings), cannot be specified. Indeed, this interaction model can also challenge mobile users with the visualization of large amounts of information in the screens of possibly small devices.

We aim at overcoming these limitations by enhancing information filtering in PGIS via textual search query specification and query expansion. In that way, the user is free to express information needs using her/his own terminology and providing details for the selection of the data (s)he is looking for.

OnToMap

The “Mappe di Comunità 3.0” project (ontomap.dyndns.org:10000/, 2014-2015), funded by Fondazione CRT, Torino, was carried out by the Università di Torino and Politecnico di Torino, in collaboration with CSI Piemonte, and gave birth to other spin-off projects, which frame the work described in this paper. “Mappe di Comunità 3.0” pursued the development of a knowledge-sharing model, and an online platform, for the integration of official cartographies developed by the Public Administration, and spontaneous ones, in a unified framework supporting information search and crowdsourcing of territorial Open Data. The application domain was that of Participatory Decision-Making processes, aimed at favoring the inclusion of citizens in the design of public policies and, orthogonally, a reflection on territorial identity. The goal was to convey the knowledge of the perception of places, creating an indicator of the territorial identities inhabiting the city. Within this context, the project focused on the development of a service for the creation and management of Community Maps that enable communities to maintain their own distributed information sharing spaces for representing viewpoints on the local territory.

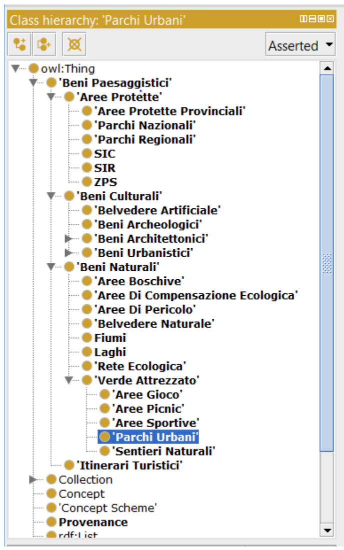

Figure 1: Portion of the OnToMap Ontology, visualized using the Protégé editor (stanford.protege.edu).

The main result of the project is the web application OnToMap - Mappe di Comunità 3.0, which can be used to consult spatial data, create custom maps, report critical issues or proposals to the local administration. OnToMap employs an ontological layer that makes it possible to:

1. integrate heterogeneous data, originated from different sources, and manage them as linked data;
2. describe semantic relations among information items to express spatial relations, different levels of abstraction in the description of entities and thematic relations.

The OnToMap ontology is used as a model for enabling the user to browse the information space and, for this purpose, a set of thematic views make it possible to visualize portions of it. For instance, Figure 1 shows a portion of the OnToMap ontology and Figure 4 presents a thematic view focused on children (Bambini) issues. By navigating the ontology graph, users can select the spatial information they are interested in and they can search for semantically related data. Given their selections, the system displays relevant data on an interactive map that provides an overview of spatial information and supports the search for information about specific elements of the territory and their relations with other elements; see the central and the right portions of Figure 5.

Browsing the domain ontology has pros and cons. By visiting the graph, it is possible to identify the constitutive dimensions of the territory, which take into account the environmental, urban, cultural and social components. However, the conceptualization might be complex and rather different from the one held by the user. Even though (s)he can select the thematic view of interest for searching information, this does not address the former issue, which we propose to handle via textual query interpretation.

Figure 2: Linguistic information associated to concept “Parchi Urbani” in the OnToMap ontology (visualized using Protégé editor). Each linguistic item is annotated with the reference language (here, Italian – it) and the ontology supports the specification of information in multiple languages.

Domain Knowledge Representation

A flexible interpretation of textual queries presupposes that the system is able to map them to its own domain conceptualization. This kind of semantic mapping is particularly difficult because, as noticed by Belkin (1980), in a search task the user is asked to specify something that (s)he does not know. Indeed, it is very likely that her/his terminology differs from the one of the system and the two have to be reconciled in order to identify the user’s information needs. In document search, some query expansion approaches use Natural Language Processing techniques to enrich the list of keywords of the search query with synonyms, in order to increase the matches with the document corpus. However, that approach is affected by word sense disambiguation problems, especially in case of short queries; see (Grootjen, 2006). In order to address this issue, we propose to enrich the domain ontology with linguistic knowledge to improve matches in a controlled way, selecting the appropriate meanings for each concept. Specifically, we propose to exploit Natural Language for enriching the domain ontology by making the meaning of concepts explicit.

We consider multiple ways to refer to the same concept, through synonyms, as well as linguistic definitions. The purpose of linguistic definitions is twofold: first, they enhance the clarity of the domain conceptualization. Second, they are a source of relevant keywords to refer to the same concept in Natural Language expressions. For instance, as shown in Figure 2, a possible definition of “Parco Urbano” (Urban Park) is “Zone verdi ben coltivate e ordinate, talvolta accessibili con limitazioni di orario” (cultivated green areas, sometimes accessible with time restrictions). This definition makes the concept relevant not only for expressions referring to urban parks, but also to green areas, and cultivated ones.

For simplifying the retrieval of concepts, regardless of the particular form used in the queries, we propose to store, for each concept *C* of the ontology, the following features:

- The lemma of a word *w* associated to *C* and to the synonyms of *w*. For instance, considering “Parchi Urbani”, the involved lemmas are “Parco”, “Urbano”; the synonyms are “Parco pubblico” (Public park) and others.
- The lemmas of all the keywords belonging to the definition(s) of *C*. In the previous example, “Zona” (Zone), “Verde” (Green), and so forth.

Both types of information are language dependent; currently we are working with the Italian language but the representation format enables multi-linguality. For the specification of linguistic definitions and synonyms, as well as for Word Sense Disambiguation, we employ the BabelFy multilingual Entity Linking and Word Sense Disambiguation service (http://babelfy.org), in combination with the Morph-it![1] lemmatization service. Specifically, we have used these services to annotate the ontology in a semi-automated way (starting with an automated annotation of meanings, and then filtering out the inappropriate ones). Moreover, the system invokes the services to extract the lemmas of the words occurring in search queries.

[1] http://sslmitdev-online.sslmit.unibo.it/linguistics/morph-it.php

Interpretation of Textual Search Queries

While interacting with OnToMap, the user can specify textual queries that also include the geographical reference, or (s)he can combine queries with the selection of an area in the map. Therefore, by default, the system exploits the geographical restrictions included in the query, if any; otherwise, it constrains the search using as a bounding box the visible portion of the map.

In order to identify a set of typical geographical query patterns we analyzed a public search engine log containing a set of users' queries (AOL, 2006). Because of the variety of types of queries in the trace, we selected those containing geographical information (e.g., "philadelphia art museum" or "cheerleading in georgia") useful for our work. We manually filtered the trace extracting the queries containing geographically inductive terms such as place names, spatial names or toponym qualifiers. We considered about 20000 records of the log and we found about 100 geographical queries. We noticed that the geographical component is often located at the beginning or at the end of the query and we found some relevant patterns. In particular, the 65% of the queries presents the pattern <where, what>, where the "what" component represents the concept which the user is searching for and the "where" component represents the geographical information; the 23% of the queries we analyzed, present the pattern <what, where> and the 12% follows the pattern <what, preposition, where>. Starting from this information, we developed a search model that first identifies the geographical references occurring in the query for defining the bounding box for the selection of items within an area of the map, and then retrieves the concepts from queries.

Identification of the geographical references: There are different patterns for the specification of geographical locations and the system tries to identify them in cascade:

1) First, it analyzes the query to identify any toponym qualifiers (e.g., square, route), which denote the presence of an address. If there is one, then the system submits the complete address to Google Geocoder (https://developers.google.com/maps/documentation/javascript/geocoding) to retrieve the referenced geographical area.

2) If there are no toponym qualifiers, the system checks whether the user has specified a geographical entity (e.g., a region, or a town as in "schools in Torino") by matching the words of the query with the names in the Geonames geographical database (www.geonames.org). Then, it submits the geographical names it found to the Geocoder to retrieve the related area.

Analysis of the search query, after having removed the geographical reference. The system identifies relevant ontology concepts by interpreting the query as follows:

1) First, it splits the query in individual words and simplifies it via stop-word removal. Then it generates a new query ("normalized query"), composed of the lemmas of each retained word.

2) Then it attempts to match the words of the normalized query on the ontology concepts. There are different degrees of matching:

- *Direct match between lemmas of the normalized query and ontology concepts.* This happens when the user has used a vocabulary consistent with the ontology. The concepts are identified by considering single lemmas of the search query as well as adjacent tuples of lemmas, in order to look for the most specific concepts corresponding to the query. For instance, if the query includes the lemmas of “public” and “service”, and the ontology includes both “services” and a sub-concept “public services”, the latter concept is identified as a match.
- *Match between the lemmas of the normalized query and those of the synonyms of concepts in the ontology*. This occurs when the user has used a similar vocabulary with respect to the ontology, but not exactly the same words.
- *Match between the lemmas of the normalized query and the keywords of concepts in the ontology.* In this case, there is no strict match between query and ontology concepts; however, the system tries to identify a set of relevant concepts by matching the query with the keywords extracted from their linguistic definitions.

Figure 3: Sample suggestions for query expansion in a case in which OnToMap does not find any solutions to the search query. Here, the user asked for “animali” (animals) and the only possibly related concepts found by the system are “Rete Ecologica” (Ecological Network) and “Aree Protette” (Protected Areas), which are loosely related to animals through linguistic knowledge.

OnToMap retrieves the instances of the relevant concepts visualizes them in the map. Moreover, it uses them to suggest query expansions, or related queries to retrieve more results. Specifically, if there is a direct match between the query and a set of concepts, we assume that the system successfully interpreted the query. Thus, it presents the instances on the map and a set of related concepts as suggestions that may interest the user (see Figure 5). The user can then modify the query if (s)he receives irrelevant results. Otherwise, the system exploits query expansion to suggest alternative search queries that correspond to types of information modeled in the ontology. In this case, it presents the suggestions (if any) in a menu sorted by degree of matching: first, it lists concepts identified through the analysis of synonyms, then those found via keywords of descriptions; e.g., see Figure 3. The user can then select the relevant concepts to visualize the results in the map, or (s)he can formulate a different query.

Related Work

Various techniques exist to improve information search. For instance, Google’s search engine manages the Knowledge Graph to relate facts, concepts and entities depending on their co-occurrence in search queries (https://developers.google.com/knowledge-graph/). On a related perspective, CoSeNa (Candan, et al., 2009) exploits keyword co-occurrence and ontological knowledge to support the exploration of text collections using a keywords-by-concepts graph, which “supports navigations using domain-specific concepts as well as keywords that are characterizing the text corpus”. These works are complementary to our own and suggest paths for future work.

Several GIS employ ontologies for conceptualizing the domain (Fonseca, 2000) and helping users in the information search process. For instance, SIAPAD (Molina and Bayarri, 2011) combines semantic knowledge representation with task-based information to map the keywords occurring in search queries with the ontology concepts related to the corresponding activities. With respect to that work, we adopt a more general approach, referring to generic linguistic knowledge, because the same type of geographical information might be relevant for tasks belonging to different domains. Other GIS, such as TripAdvisor (www.tripadvisor.it), ask for a separate specification of geographical entities and searched information. They use the keywords included in the search query to match geo-data names, item reviews, etc., providing mixed results that include heterogeneous geo-data (e.g., if the query includes term “sport”, results can include a café whose name includes the term “sport”, a list of sport facilities, etc.). In comparison, OnToMap interprets the user’s terminology referring to specific ontology concepts and as such can retrieve coherent results, e.g., all the sport facilities of an area.

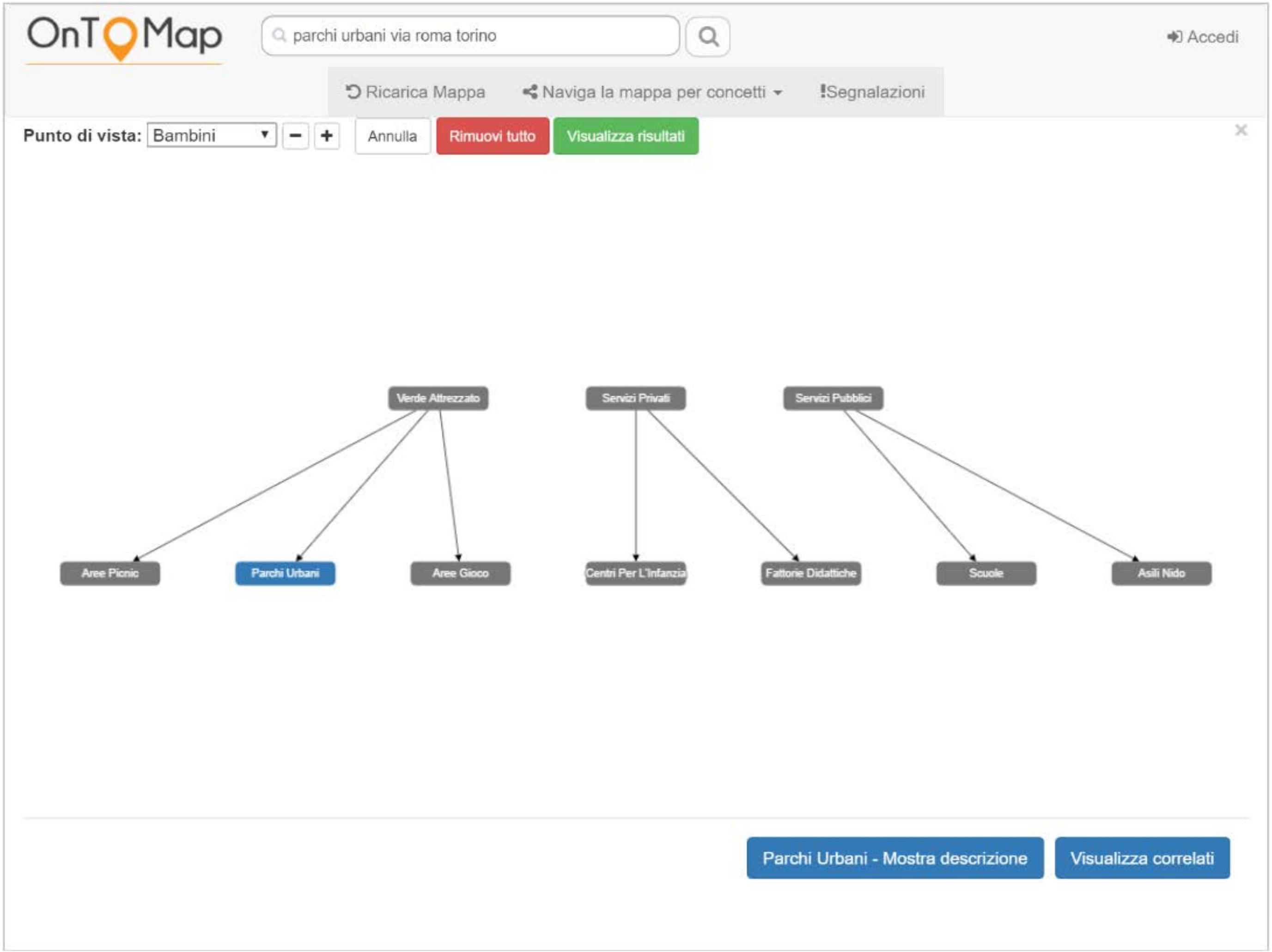

Figure 4: Thematic view of the OnToMap ontology graph focused on children issues (tablet view). The concepts selected by the user for retrieving results are blue (here, "Parchi Urbani", urban parks).

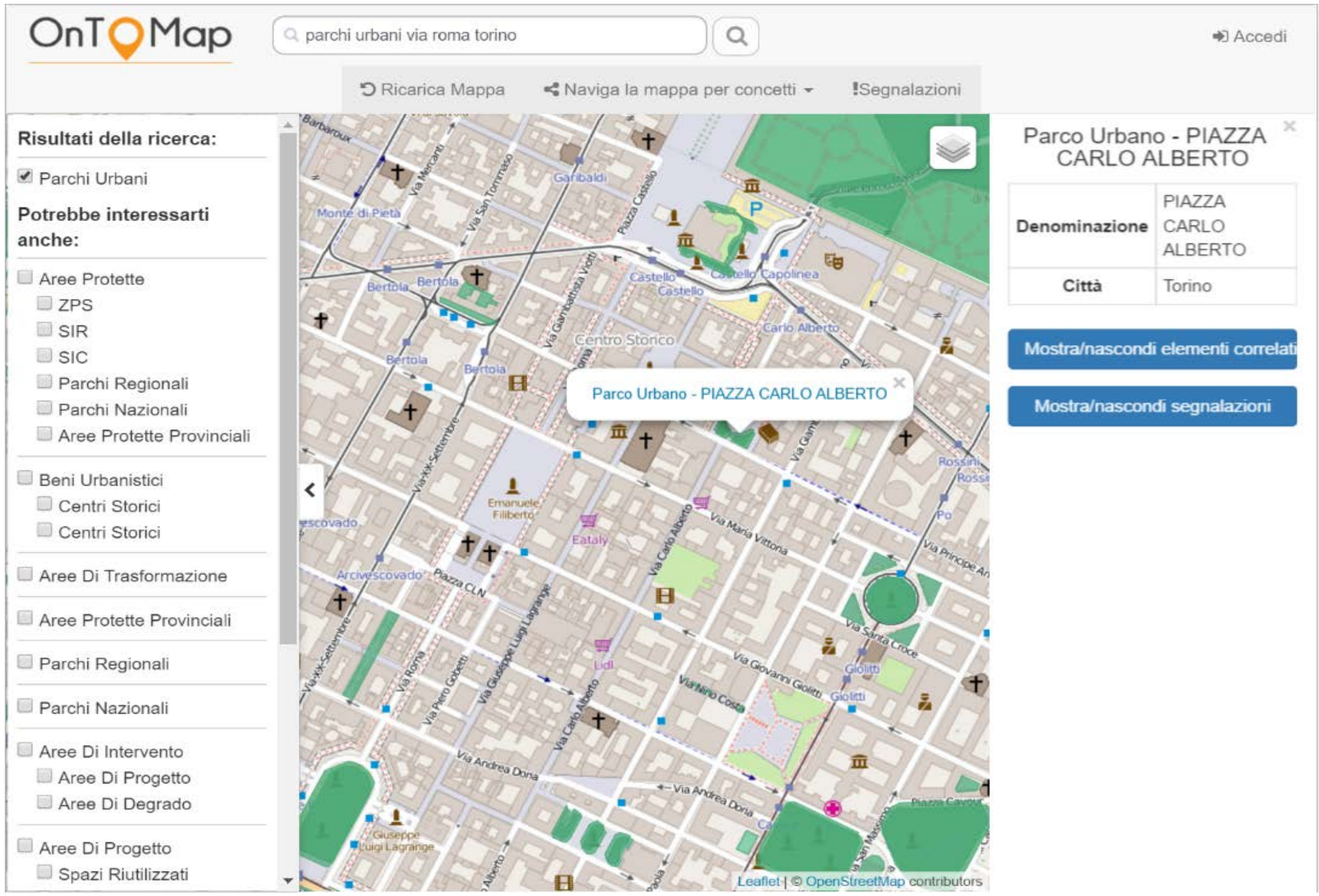

Figure 5: Search results and suggestions menu for query "parchi urbani via roma torino" (looking for urban parks located in via Roma, Torino). Visualization for tablets.

Conclusions

This paper presented a model for the interpretation of textual search queries within a PGIS that supports a holistic search and sharing of information about the resources of a territory and its cultural heritage.

The proposed approach exploits NLP to match search queries to the conceptualization of the domain in order to enable users to express their information needs in a natural way. We applied this model in the OnToMap system and we plan to test it with users by proposing activities in Torino city: we will compare user's performance using OnToMap with respect to that using a well-known GIS, such as OpenStreetMap. Moreover, we will collect feedback on user experience with the two systems.

Acknowledgments

This work is partially funded by project MIMOSA (MultIModal Ontolgy-driven query system for the heterogeneous data of a SmArtcity, "Progetto di Ateneo Torino_call2014_L2_157", 2015-17).

References

1. L. Ardissono, T. Kuflik, D. Petrelli. 2012. Personalization in Cultural Heritage: The Road Travelled and the One Ahead. *USER MODELING AND USER-ADAPTED INTERACTION - The Journal of Personalization Research,* 22, 1-2 (2012), 73-99.
2. K. S. Candan, M. Cataldi, L. Di Caro, M. L. Sapino, and C. Schifanella. 2009. CoSeNa: a Context-based Search and Navigation System. In *MEDES '09 Proc. of the Int. Conference on Management of Emergent Digital EcoSystems.* Article n. 33.
3. C. E. Dunn. 2007. Participatory GIS a people's GIS? *Progress in Human Geography*, 31.
4. F. T. Fonseca, M. J. Egenhofer, C. A. Davis Jr., and K. A.V. Borges. 2000. Ontologies and knowledge sharing in Urban GIS. *Computers, Environment and Urban Systems,* 24(3) (2000), 251-272.
5. M. Goetz, A. Zipf. 2013. The evolution of geo-crowdsourcing: bringing volunteered geographic information to the third dimension. In: Sui D., Elwood S., Goodchild M. F. (Eds.), *Crowdsourcing Geographic Knowledge*, 139-159.
6. M. F. Goodchild. 2007. Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2 (2007), 24-32.
7. F.A. Grootjem. Th. P. van der Weide. 2006. Conceptual query expansion. *Data & Knowledge Engineering,* 56, 174-193.
8. K. Janowicz, S. Scheider, T. Pehle, and G. Hart. 2012. Geospatial Semantics and Linked Spatiotemporal Data – Past, Present, and Future. *Semantic Web - On linked spatiotemporal data and geo-ontologies*, 3(4). 321-332.
9. M. Molina, S. Bayarri. 2011. A multinational SDI-based system to facilitate disaster risk management in the Andean community. *Computers & Geosciences*, 37, 1501-1510.
10. D. Sui, S. Elwood, M. Frank Goodchild (Eds.). 2013. *Crowdsourcing Geographic Knowledge*. Springer.
11. UNESCO. 2003. Convention for the Safeguarding of the Intangible Cultural Heritage, Paris, October 17 2003.
12. A. Voghera, R. Crivello, L. Ardissono, M. Lucenteforte, and A. Savoca. 2015. Il progetto OnToMap: costruire conoscenza territoriale attraverso l'uso di un'applicazione ICT. In *Proc. of the 19th Conference ASITA.*

13. A. Voghera, R. Crivello, L. Ardissono, M. Lucenteforte, A. Savoca, and L. La Riccia. 2016. Production of spatial representations through collaborative mapping. An experiment. To appear in *Proc. of INPUT 2016*.