Supporting the exploration of cultural heritage information via search behavior analysis

This is the author's manuscript

Original Citation:

Availability:
This version is available http://hdl.handle.net/2318/1711677 since 2023-02-10T15:00:03Z

Publisher:
Association for Computing Machinery, Inc

Published version:
DOI:10.1145/3314183.3323862

Terms of use:
Open Access
Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)
Supporting the Exploration of Cultural Heritage Information via Search Behavior Analysis

Noemi Mauro
University of Torino
Torino, Italy
noemi.mauro@unito.it

ABSTRACT

Thematic maps, traditionally developed to present specific themes within defined geographical areas, are an interesting information presentation model for Cultural Heritage exploration because of the abstract view on the territory they provide. However, in order to cope with possibly heterogeneous user interests, they should be adapted to the individual user by including the relevant types of information, given her/his specific interests.

In a previous paper, we proposed an approach to the integration of thematic maps in the OnToMap Participatory GIS (Geographic Information System), in order to support query expansion during an exploratory search task. The proposed maps were built on the basis of a survey in which we asked people to rate the relevance of a set of concepts to five main themes around which we developed the maps. In this paper we go one step forward and we propose a more general approach to information search support in order to automatically create thematic maps, based on the analysis of frequently co-occurring search interests in a search engine query log. This type of analysis supports the identification of clusters of concepts that people frequently search within the same sessions and helps the identification of co-occurring topics that can be proposed to users when exploring an information space. In this way, when the user browses a catalog of Cultural Heritage information, (s)he can both visualize the thematic maps relevant to the search context, and be guided in the navigation within types of information, looking for possibly complementary types of data to satisfy her/his needs.

CCS CONCEPTS

- Information systems → Geographic information systems;
- Ontologies; Search interfaces; Personalization; Information retrieval. Query suggestion;

KEYWORDS

Information search; Thematic maps; Session-based Concept Suggestion; Personalization.

ACM Reference format:
https://doi.org/10.1145/3314183.3323862

1 INTRODUCTION

The exploration of the Cultural Heritage (CH) patrimony of a territory is a source of inspiration for the integration of information filtering and Adaptive Hypermedia techniques with Geographical Information Systems for several reasons. Firstly, it challenges the user with a large amount of possibly heterogeneous information (from monuments to museums to areas of conservation), which requires a rational organization in order to help its inspection. Secondly, for many CH items the digital representation is just a mirror of their physical presence and the spatial perspective on information exploration should be taken into account, e.g., when planning a visit to the physical sites. Therefore, the presentation of CH information should be geolocalized and possibly integrated in geographic maps helping the user visualize the spatial distribution of items.

Starting from this consideration, we propose a semantic approach to the representation of information, which makes it possible to describe information items in a rich way, to relate them on the basis of semantic associations and, at the same time, to provide spatial navigation means in order to help the user orientate her/himself in a region of interest. The idea is that of providing a data integration approach that makes it possible to generate custom views of the territory, focused on the user’s information needs, but easily extensible to support the inspection of topically related contents. For this purpose, the geographical dimension of a territory should be integrated with a semantic exploration of content, as described in Semantic GeoSpatial Semantic Web [19], and with a proactive information retrieval support.

Thematic maps, traditionally developed to present specific themes within geographical areas, are an interesting information presentation model because of the abstract view of the territory they provide. However, it is difficult to determine the types of data that should be presented a priori: on the one hand, the maps should adapt to users’ search behavior in order to reflect changes in the interests of the population. On the other hand, they might not suit the individual user’s needs; e.g., (s)he might be interested in a subset of a map, or in a mixture of the content provided by two of them. Therefore, the management of static information layers might not be the best solution to be adopted.
In order to support the exploration of territorial data under different points of view, in [22] we developed a model that combines the interpretation of the user’s queries with a dynamic selection of thematic maps reflecting the information needs expressed by the user while (s)he searches for data on a geographical map. The model is integrated in the OnToMap Participatory GIS (Geographic Information System) [2, 4, 5, 32]: the thematic maps represented a set of broad points of views on the information space, e.g., a tourist view on the territory.

In the present paper we go one step forward by introducing a general model for identifying frequently co-occurring search interests, thus supporting a flexible type of query expansion. The user is free to inspect the thematic maps suggested by the system because they are relevant to the ongoing search task, or to expand her/his search queries with specific concepts suggested “on the fly” by the system. The clusters of frequently co-occurring interests we developed are derived from an analysis of general search behavior, available in the logs of search engines. Therefore, they cover a broader perspective on information than the Cultural Heritage one and can help users find useful information in a more general way than only relying on a static set of themes. The cluster of concepts can be seen as dynamic thematic maps derived from common search behaviours.

The identification of suitable concept clusters for query expansion is based on: (i) an ontological domain knowledge representation for describing the meaning of concepts and their semantic relations, as well as for handling geographical data as Linked Data [33]; (ii) a semantic query expansion model for identifying the concepts referenced in the user’s queries, and thus the types of information relevant to her/his search; (iii) the analysis of the co-occurrence of concepts in search sessions.

The remainder of this paper is organized as follows: Section 2 positions our work in the related research. Section 3 presents the main features of the OnToMap application and the introduction to thematic maps. Section 4 describes the approach we adopted to identify the co-occurrences of concepts in search sessions and some query expansion strategies to make concepts suggestions. Section 5 concludes the paper and outlines our future work.

2 RELATED WORK
The spatial search perspective is a key aspect of our work, with respect to other repositories of artworks, such as online collections (e.g., see [18]) and large-scale projects supporting the digitalization and diffusion of artistic and historical content across Europe; e.g., Europeana Collections [11]. The main difference is that, instead of providing the user with a catalog of artworks organized on the basis of specific features (e.g., topic, etc.), we aim at integrating digital information about artworks in a spatial representation of the territory to which they belong. The geographical aspect is a key element to understand the identity of a territory, as demonstrated in the experience of the Parish Maps and of community mapping [30].

Thematic maps (e.g., see [17]) are useful to present specific aspects of a territory in an abstract way, by exploiting the geographical dimension to facilitate data visualization and exploration. In other projects, they have been used to describe colors, smells, perception of danger, etc. of a geographical area; e.g., see [1, 10]. In OnToMap, we have exploited the thematic maps model to support knowledge discovery, focusing them on the information relevant to the user’s interests, which can be inferred from the search queries (s)he submits. This has lead us to move from an initial model in which a static set of themes was identified (e.g., family, tourism, culture, etc.) to a more flexible approach, based on frequent co-occurrence of interests in general information search behavior.

The identification of interest co-occurrence we propose is related to a few works supporting query expansion, query reformulation, and term suggestion in Information Retrieval. Some researchers propose to analyze session-based user behavior in order to detect co-occurrence relations useful to improve search queries, taking the search context into account. For instance, in [7] Cao et al. suggest queries on the basis of the context provided by the user’s recent search history, by clustering queries on the basis of the search results visited by users. Moreover, Huang et al. [16] and Chen et al. [8] detect term co-occurrence in search sessions to group sets of relevant words that can be mutually suggested. Our work is different because we adopt a linguistic interpretation approach (based on lemmatization and Word Sense Disambiguation) to find the concepts referenced in the queries; see [21]. Therefore, we extract information about concept co-occurrence, which is more general than term co-occurrence.

Knowledge graphs describe item features and relations among entities, supporting the analysis of item relatedness, as well as similarity for information filtering and top-N recommendation. In several works these graphs are extracted from document pools and/or from the Linked Data Cloud. For instance, in [28], Di Noia et al. create a relatedness graph, by analyzing external data sources such as DBPedia, in order to support the evaluation of semantic similarity between items. Analogously, item features have been extracted from the Linked Data Cloud to improve recommendation performance in [25–27, 31]. Some works attempt to extend the relations among information items by integrating data deriving from the observation of different types of user behavior. E.g., Google search engine manages the Knowledge Graph [14] to relate facts, concepts and entities depending on their co-occurrence in queries. Moreover, entity2rec learns user-item relatedness from knowledge graphs by analyzing external data sources such as DBPedia, in order to support the evaluation of semantic similarity between items. In comparison, our work exploits knowledge graphs based on concept networks in order to model the relatedness among types of information, instead of dealing with individual items.

Some projects proposed the adoption of ontologies for the integration of geographical data as Linked Data, in order to support information retrieval, but focused on data modeling, leaving the management of personalized views on information spaces apart; e.g., OSM Semantic Network [6] and OSMonto [9], and [12, 13].

3 ONTOMAP
The OnToMap Participatory GIS [4, 5, 32] supports the management of interactive community maps for information sharing and participatory decision-making. The application enables the user both to
consult spatial data and to create custom maps, which reflect individual information needs and can be enriched with crowdsourced content (information items and geo-localized comments) to help project design and group collaboration. Figure 1 shows a portion of the user interface of the application. In particular, the higher portion of the figure shows the results of a search query with the visualization of the details of a geographic item.

3.1 Knowledge Representation
As described in [32], the system exploits a semantic representation of domain knowledge based on a geographical ontology that organizes information in concepts related to one another by generalization or semantic relations [12]. Concepts characterize the attributes of data and thus support multi-faceted data retrieval and presentation [13], as well as the integration of heterogeneous Linked Data [33]. Moreover, each ontology concept is enriched with linguistic and encyclopaedic knowledge that makes it possible to use query expansion techniques to interpret queries by addressing the word sense disambiguation problem; see [3] for details.

3.2 Information Search Support
OnToMap supports multiple ways to specify search queries, including free-text queries and a menu-based navigation of the network of concepts forming the domain ontology. The system semantically interprets the free-text queries by matching the terms they contain with the ontology concepts, and filters the results on the basis of the qualifiers specified in the queries, if any. For all the matches, the system employs synonyms and linguistic descriptions of concepts to abstract from the terminology used by the user; see [3, 21].

OnToMap visualizes search results on the geographical map as pointers or, when available, as geometries [2]. Results can be clicked to retrieve detailed information about them. Figure 1 shows the visualization of the search results of free-text query *Museums in Torino* (see the upper left portion of the user interface). By clicking on an item in the map, the user can view a sticky note and the descriptive table displayed in the right portion of the page.

During the interaction with the user, OnToMap helps her/him discover possible relevant topics within the information space by proposing up to three thematic maps that are particularly relevant to the queries (s)he submitted within the interaction session. The user can view the proposed thematic maps by clicking on button "show related thematic maps". With the approach described in this paper, we aim at proposing to the users clusters of concepts with the same visualization paradigm used for static thematic maps.

3.3 Thematic Maps
As described in [22], OnToMap exploits a set of thematic maps that we defined by spreading an online survey in which we asked people to select the concepts that are most relevant to five themes: culture, tourism, nature, family, and services. In the survey, for each theme, we presented the list of concepts forming the OnToMap ontology and we asked people to select one or more that suited the theme. Given the results of the questionnaires, we computed the relevance of each concept with respect to the five themes, depending on how many people associated the concept to the theme. We assumed that the evidence of the relevance of a concept $c$ to a theme $T$ is determined by people agreement. In the definition of the thematic maps, we considered a concept as relevant to a theme, and thus belonging to the respective map, if it were selected by at least 50% of people who filled in the survey.

Two examples of static thematic maps are:

1. **Tourism:** [Landscape Heritages, Cultural Heritages, Archaeological Heritages, Architectural Heritages, Museums, Natural Heritages, etc.]
We complement the suggestion of relevant information provided by static thematic maps showing to the users a set of clusters of concepts that are dynamically created by analyzing frequent search behaviour. Similar to the case of the thematic maps, the suggestion of relevant concepts for query expansion is based on the interpretation of the queries posted by the user and by the identification of concepts that are strongly related to those mentioned by the users in general information search behavior.

4 SUGGESTION OF CONCEPTS FOR QUERY EXPANSION

We complement the suggestion of relevant information provided by static thematic maps showing to the users a set of clusters of concepts that are dynamically created by analyzing frequent search behaviour. Similar to the case of the thematic maps, the suggestion of relevant concepts for query expansion is based on the interpretation of the queries posted by the user and by the identification of concepts that are strongly related to those mentioned by the users in general information search behavior.

4.1 Identification of Frequently Co-occurring Search Interests

In order to describe the strength of the associations between concepts in search behavior, we analyze their co-occurrence in the search sessions of a query log. By co-occurrence we mean the fact that two or more concepts are referred by the queries belonging to the same session. The identification of concepts in search queries is done by matching the lemmatized knowledge (e.g. synonyms, keywords) associated to the ontology concepts with the lemmatized words of the search queries. The ontology concepts have been enriched using BabelFy multilingual Entity Linking and Word Sense Disambiguation service [24], and Stanford CoreNLP lemmatizer [20]. We deal with the Word Sense Disambiguation problem related to the search queries during the construction of the Concept Co-occurrence Graph (CCG).

In the following, we summarize the analysis of concept co-occurrence, which adopts the approach described in [23].

The Concept Co-occurrence Graph (CCG) represents concept co-occurrence in search sessions: in the CCG, nodes represent the concepts referenced in the analyzed queries and the weight of edges represents the co-occurrence frequency of the connected concepts; i.e., how many times the concepts have been identified within the same search sessions.

The CCG is built as follows: given two concepts $x$ and $y$, the weight of the edge that connects them is defined as:

$$w_{xy} = \sum_{S \in \text{Sessions}} Freq_{S_{xy}}$$

where $Freq_{S_{xy}}$ represents the evidence provided by session $S$ to the co-occurrence frequency of $x$ and $y$. Given $S = \{Q_1, \ldots, Q_n\}$, $Freq_{S_{xy}}$ is computed as the maximum evidence of co-occurrence of $x$ and $y$ in $S$:

$$Freq_{S_{xy}} = \max_{k=1}^n (Freq_{xQ_k}, ev_{xQ_k}, ev_{yQ_k-1})$$

where $Freq_{xQ_k}$ is the co-occurrence evidence of $x$ and $y$ provided by query $Q_k$, and $ev_{xQ_k}$, $ev_{yQ_k-1}$ is the one provided by $Q_1, \ldots, Q_{k-1}$.

The contribution of a query $Q$ to the estimation of co-occurrence evidence is represented as follows:

- If $Q$ contains $k$ terms ($k > 0$), each one identifying a non-ambiguous concept: $T_1 \Rightarrow c_1, \ldots, T_k \Rightarrow c_k$, then, for each concept $c$ of $Q$:
  - The co-occurrence evidence between $c$ and every other concept $d$ of $Q$ is $Freq_{cdQ} = 1$.
  - The co-occurrence evidence between $c$ and every other concept $e$ identified in a non-ambiguous way in the other queries of $S$ is $Freq_{ceQ} = 1$.
  - The co-occurrence evidence between any other concepts $w$ and $z$ identified in $S$ is $Freq_{wzQ} = 0$.

- If $Q$ contains an ambiguous term $t$ that refers to $m$ concepts, the particular concept the user is focusing on cannot be identified. Therefore, the co-occurrence evidence brought by $t$ is computed as above, but the assigned evidence is $1/m$, in order to consider the possible interpretations of $Q$, and divide evidence among ambiguous concepts.

For the development of the co-occurrence graph used in this work we exploited the AOL query log.\(^1\)

The co-occurrence graph of the AOL dataset is strongly connected: almost all of the ontology concepts are linked to each other by an edge whose weight is $> 0$. We thus pruned its weakest arcs in order to select for reasoning the strongest co-occurrence relations among concepts.

4.2 Creation of the Concept Co-occurrence Clusters

Starting from the pruned Concept Co-occurrence Graph (CCG), the clusters are created by applying a community detection algorithm (i.e. COPRA [15]) which identifies the sets of strictly correlated concepts. Our hypothesis is that they correspond to sets of concepts that are frequently searched together.

We selected an algorithm that detects overlapping communities in order to allow to a concept to belong to different clusters at

\(^1\)The log is available at http://www.cim.mcgill.ca/~dudek/206/Logs/AOL-user-et-collection/
the same time. The idea was that, within a session, the user might focus on more than one set of highly correlated concepts, and that, starting from the same query, she might explore different paths of the information space.

4.3 Selection of Thematic Maps

After the identification of clusters we tested different strategies to select the set of clusters to show to the user as thematic maps given the search query submitted by her/him and her/his context.

In [23], three main strategies for cluster selection have been proposed and tested, which result in different concepts suggestion policies. After the offline evaluation of the three strategies done on the AOL query log, we concluded that the strategy called "SLACK" is the most convenient strategy to be integrated in the OnToMap application. It is fairly precise and, at the same time, it suggests more relevant concepts than the other ones. This can be explained by the fact that we aim at providing the user with an overview of the sets of concepts to choose from, given the search context provided by the previous queries. As the user will be enabled to choose relevant concepts in a multiple-choice box, the presence of very few irrelevant concepts, as is the case of SLACK, is not a problem.

A formal definition of the SLACK strategy is described in the following. Given a search session $S = \{Q_1, \ldots, Q_i, \ldots, Q_n\}$, let $S@i$ be the observed portion of $S$, i.e., $\{Q_1, \ldots, Q_i\}$. Moreover, let $C@i = \{c_1, \ldots, c_k\}$ be the set of concepts identified by interpreting the queries of $S@i$.

The SLACK strategy selects a set of clusters $\{CL_{x_1}, \ldots, CL_{x_y}\}$ that contain at least one concept $c$ such that $c \in C@i$; i.e., $c$ has been referenced in the observed portion of the session ($S@i$). For instance, given $C@i = \{c_1, c_5\}$, and three clusters, $CL_1 = \{c_1, c_3, c_7\}$, $CL_2 = \{c_2, c_3, c_5, c_6\}$, $CL_3 = \{c_3, c_9, c_{10}\}$, SLACK selects $CL_1$ and $CL_2$, and it suggests $\{c_2, c_5, c_7, c_8\}$.

After the integration of the model described above in the OnToMap web application, we plan to carry out an online evaluation with real users. The idea is to test if we can achieve similar results to the ones obtained during the offline evaluation where we used the AOL log.

5 CONCLUSIONS

We proposed a concept suggestion model aimed at supporting query expansion in information search. The model is based on the analysis of frequently co-occurring interests in search sessions. It can be applied to suggest types of information relevant to the exploration of large Cultural Heritage catalogs, helping the user orientate her/himself while she browses the concepts available in the information space. The spatial approach to search and data visualization adopted in our work are aimed at presenting information situated in the geographical regions where Cultural Heritage items are located. In this way, information can be presented helping the user orientate her/himself. Our model is based on a semantic representation of geo-data and on a query expansion model based on the recognition of the concepts referenced in the user’s search queries.

ACKNOWLEDGMENTS

This work was supported by the University of Torino through projects “Ricerca Locale” and the Computer Science PhD program.

REFERENCES


