inWalk: Interactive and Thematic Walks inside the Web of Data

Silvana Castano
Università degli Studi di Milano
DI - Via Comelico, 39
20135 Milano, Italy
silvana.castano@unimi.it

Alfio Ferrara
Università degli Studi di Milano
DI - Via Comelico, 39
20135 Milano, Italy
alfio.ferrara@unimi.it

Stefano Montanelli
Università degli Studi di Milano
DI - Via Comelico, 39
20135 Milano, Italy
stefano.montanelli@unimi.it

ABSTRACT

The goal of this paper is to demonstrate *in*Walk, an interactive web-based system for linked data exploration featured by the notion of *in*Cloud and thematic walk. The demonstration focuses on the key functionalities of the system for smart data aggregation and navigation.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

General Terms

Thematic web-data exploration, similarity-based data aggregation

1. INTRODUCTION

The availability of large datasets of linked data makes it possible to access information, and knowledge on the semantic web through URIs (Universal Resource Identifier), RDF (Resource Description Framework), and new query languages like SPARQL [1]. However, users searching data about a specific topic of interest need usually to face a multi-step and loosely-intuitive browsing activity to build a comprehensive picture of the data of interest [4, 7]. In this paper, we present *in*Walk, an interactive system for the exploration of linked data based on the notion of *in*Cloud. An *in*Cloud is a high-level thematic graph where nodes represent *clusters* of similar linked data and edges represent relations of *proximity* between nodes. A system demonstration is available online at `http://islab.di.unimi.it/inwalk` and it is articulated as a sequence of user tasks corresponding to increasing levels of skills in using data-driven applications and increasing levels of complexity of the *in*Walk functionalities that are demonstrated.

2. THE INWALK SYSTEM

*in*Walk is suited to support both skilled and non-skilled users on two critical aspects related to linked data exploration. Two main contributions of *in*Walk are the target of this demo. On one side, the goal of *in*Walk is to overcome the rigid web interfaces of linked data repositories by providing thematic, high-level data views built through similarity-based aggregation techniques. On the other side, *in*Walk aims at supporting users that are interested in querying the linked data, but are unfamiliar with RDF-based query languages (e.g., SPARQL, MQL) by providing intuitive keyword-based and SQL-like query tools.

2.1 System features

The *in*Walk system is characterized by the following key features.

Abstraction-by-aggregation Given a set of linked data $\mathcal{S}$, the *in*Walk system provides a high-level, conceptual view of such data through the notion of *in*Cloud [2]. An *in*Cloud is a graph $iC = (N, E)$, where a node $n_i \in N$ represents a *cluster* of similar linked data whose meaning is expressed through a synthetic and representative description called *cluster essential*, and an edge $e(n_i, n_j) \in E$ represents a *proximity link* between the clusters n_i and n_j. An *in*Cloud is the result of an abstraction process based on linked data aggregation. A node $n_i = (cl_i, d_i)$ of an *in*Cloud represents a *theme* and it is defined as a similarity cluster cl_i, with an associated essential description d_i. A cluster cl_i is a set of linked data that are found to be similar according to a considered list of properties. An essential d_i is the set of the top-k most representative *labels/types* featuring the linked data belonging to cl_i. Each node $n_i \in N$ is also characterized by a *prominence* value p_i, which expresses the importance of n_i within the overall *in*Cloud. A *proximity link* $e(n_i, n_j)$ is set between n_i and n_j to denote a similarity-based relationship between the involved nodes and a *degree of proximity* x_{ij} is associated with e to express the strength of such a relationship[1].

Exploration-by-walks The *in*Walk system enforces linked data exploration by relying on two possible actions, namely the *inside walk* and the *thematic walk*, defined over a considered *in*Cloud. The inside walk corresponds to the idea of a "walk-in-the-deep" of a certain cluster cl_i to explore its contents. A panel is shown on the right-hand side of the *in*Walk interface listing the complete essential d_i as well as all the linked data contained in cl_i. The user can select an

[1]Technical details about *in*Cloud construction, cluster-essential definition, and proximity-link specification are out of the scope of this paper; technical details can be found in [2, 3].

10.5441/002/edbt.2014.60

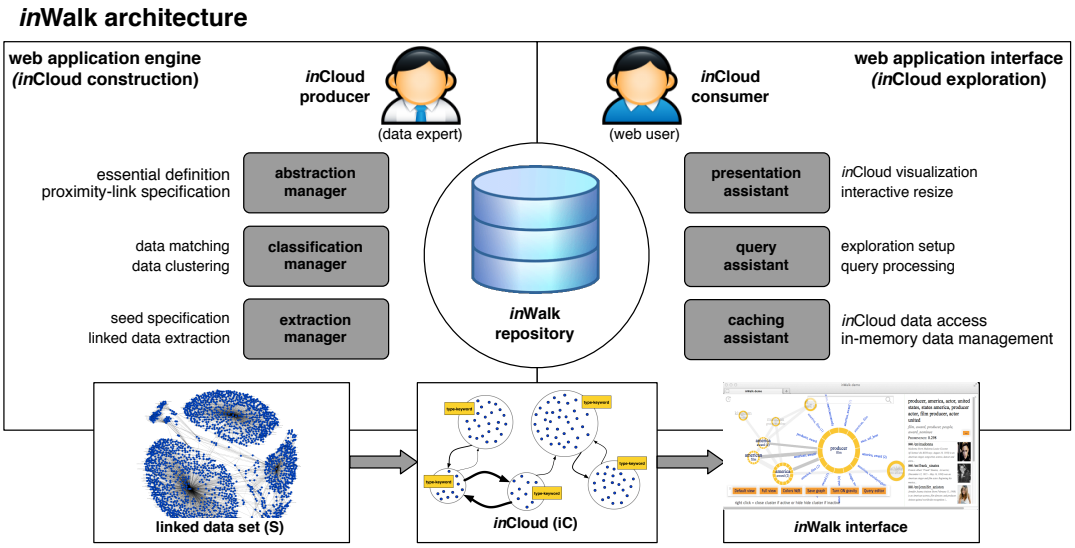

Figure 1: Architecture of the *in*Walk system

item of interest within the panel to visualize the corresponding linked data view in the repository of provenance (e.g., the Freebase web interface). The thematic walk corresponds to the idea of a "walk-in-the-surface" of the *in*Cloud. By selecting a cluster, the corresponding neighborhood is shown on the borderline of the cluster circle. The neighborhood of a cluster cl_i is the set of clusters directly connected to cl_i through a proximity link in the *in*Cloud. The borderline of the cluster cl_i is segmented into slices and each slice is a pointer to a neighbor cluster. Each slice is associated with the name of the neighbor cluster. By clicking on a slice, the user moves from the current cluster to the neighbor cluster which becomes active in turn. By moving from one cluster to another on the basis of the proximity links (i.e., by slice selection), the user describes what we call a "thematic path" within the *in*Cloud. We stress that cluster essentials and names are automatically extracted from original linked data to bring out the theme/topic characterizing the cluster contents, by also maintaining the original terminology of the repository of provenance. This way, a thematic path chains the sequence of themes/topics explored by the user while enabling a seamless shift from the thematic view of *in*Walk to the linked data view in the repository of provenance.

Filtering-by-patterns The *in*Walk system supports filtering operations over the *in*Cloud structure to enable the user in focusing the exploration of a portion of *in*Cloud that satisfy a certain selection criterion. The *in*Cloud Query Language (IQL) has been developed to enforce user-friendly, assisted modalities of query formulation, namely *keyword-based* and *query-based*. Keyword-based editing allows the user to specify a query "á la search-engine" where a list of target keywords is entered by the user. Only the clusters that contain all of them, either as an essential label or a type, are shown in the result. Auto-completion and controlled-vocabulary mechanisms are enforced to support the user editing. A history of past keyword searches is also provided. Query-based editing allows the user to specify an IQL query according to a sort of SQL-like syntax. Four different query patterns are provided in IQL to support i) cluster selection, ii) cluster join, iii) cluster path retrieval, and iv) intersection, union, and difference of sub-clouds.

2.2 System architecture

The *in*Walk architecture is featured by two main components, namely the *web application engine* and the *web application interface* (see Figure 1).

The **web application engine** is the back-end component of *in*Walk and it is in charge of constructing *in*Clouds. It transforms a raw set of linked data extracted from a given repository (e.g., Freebase - `http://www.freebase.com`, DBpedia - `http://dbpedia.org`) into a corresponding *in*Cloud for subsequent walk exploration through the application interface component. The *extraction manager* acquires from a repository $\mathcal{R}$ the linked data set $\mathcal{S}$ upon which the *in*Cloud is built. The *classification manager* first executes a data matching step over the set $\mathcal{S}$ to detect pairs of similar linked data through the HMatch 2.0 suite. An agglomerative hierarchical clustering step is then performed to generate the set of similarity clusters that constitutes the *in*Cloud (see [5] for more technical details). Finally, the *abstraction manager* finalizes the construction of the *in*Cloud graph structure out of similarity clusters.

The **web application interface** is the front-end component of *in*Walk and it has the role of supporting the web users in the exploration of *in*Clouds (see Figure 2). The interface is a HTML5+Javascript GUI and it can be visualized through any HTML5-compliant browser. The *presentation assistant* is in charge of managing the visual layout of *in*Clouds according to the number of cluster nodes and proximity links to represent on the screen. A cluster is visualized as a circle whose *name* is the pair ⟨label, type⟩ corresponding to the first label and first type of the cluster essential, and whose *size* is proportional to the cluster prominence. A proximity link is represented as an edge between two concepts whose thickness on the screen is proportional to the proxim-

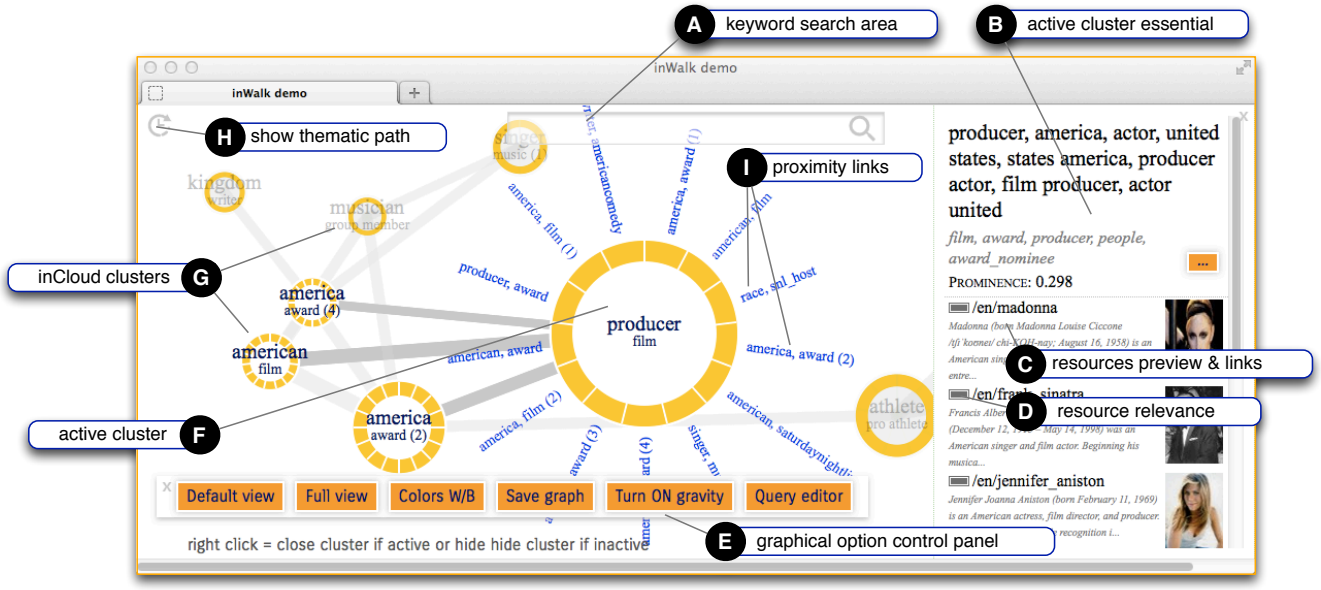

Figure 2: The *in*Walk graphical user interface (GUI)

ity degree. The *query assistant* is responsible of supporting the user in selecting the portion of the *in*Cloud to visualize as the starting point for exploration. This assistant is in charge of supporting query processing (both keyword-based and query-based modalities). Finally, the *caching assistant* has the role of managing data loading/unloading from/to the *in*Walk repository and the application memory.

3. THE INWALK SEARCH FUNCTIONALITIES

The web interface of *in*Walk supports users in exploring and searching a given *in*Cloud.

At startup, the *in*Walk interface asks the user to select the *in*Cloud to walk from a list of available *in*Clouds stored in the *in*Walk repository. Since the *in*Cloud to show can be large in size, a threshold mechanism is enforced to choose the top-k most prominent clusters to show in the default view. The demonstration is based on an *in*Cloud generated from a real dataset *AC* extracted from Freebase. The AC dataset is a collection of about 2900 linked data resources and 400 classes with more than 4000 predicates describing both sports and movies celebrities. The *in*Cloud generated from AC is composed by 44 clusters, 368 proximity links with an average prominence of 4.28. In the demonstration, the default view is generated by using the average prominence in the whole *in*Cloud as selection threshold. The complete *in*Cloud can be visualized by selecting the Full view button of the interface (Figure 2(E)). The default portion of *in*Cloud (sub-cloud) shown at the *in*Walk startup can be visualized by selecting the Default view button and contains 13 clusters of which 7 about movie and music celebrities and 6 about sports athletes. The main functionalities of *in*Walk will be described by addressing specific user tasks, corresponding to increasing levels of complexity of the *in*Walk functionalities. The *in*Walk demo is organized in two main user tasks.

User task 1 (U1 - thematic walk). In this task, users create thematic paths corresponding to their search interests on the default view of the *in*Cloud, without typing text or queries. For example, we start by selecting the cluster labeled producer/film. This action activates the cluster and highlights its proximity links. At the same time, a panel on the right-hand side is shown to visualize the details of the active cluster, namely full cluster essential, resource preview, and resource relevance in the cluster, (see Figure 2(B-D)). The walk starts by selecting the neighbor cluster labeled america/award(4)[2]. This action adds a step in the walk as well as the cluster america/award(4) to the current thematic path (see Figure 3(A)). For the next walk step, we choose to click on the slice labeled kingdom, award of america/award(4). This action highlights the cluster kingdom/award, which contains British actors. This cluster is now activated and we go one step ahead in the walk by choosing the neighbor cluster labeled people/award. We decide to stop here the walk, which originated the thematic path producer/film → america/award(4) → kingdom/award → people/award (see in Figure 3(A)).

User task 2 (U2 - inside walk). In this task, users search for a cluster of interest by exploiting the keyword search functionalities of *in*Walk (Figure 2(A)), where they can freely type any keyword for search with the help of an auto-completion mechanism, of a list of recent searches, and of the list of the most relevant keywords in the dictionary (Figure 3(B-D)). In particular, if we are interested in searching for sports, we can type the string "spor" and we choose "sports" from the keywords suggested by the auto-completion. The result is a sub-cloud of 16 clusters, containing data about athletes. We then activate the cluster italy/pro athlete. Cluster information and the linked data

[2]The numeric index is used to distinguish the clusters resulting with the same ⟨label, type⟩ descriptor.

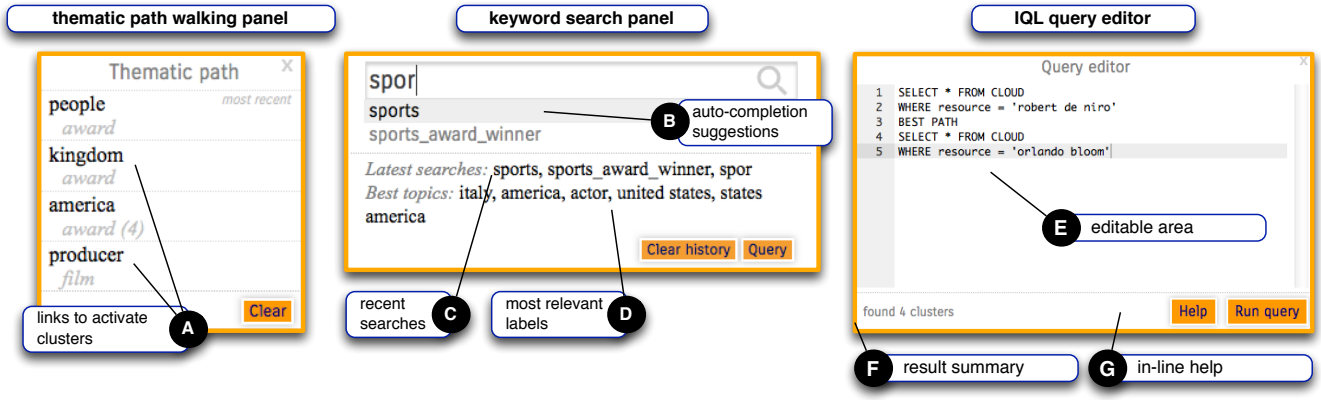

Figure 3: Main panels of the *in*Walk search interface

contained in this cluster are shown in the right-side panel of the interface (see Figure 2(B,D)). This cluster can be chosen now for a walk-in-the-deep: the essential tells us that the cluster is about Italian athletes and contains 24 people and we can access a more detailed description of one of the athletes by selecting the athlete's name and visualizing her corresponding Freebase web page.

More complex user tasks based on the IQL editor will be illustrated in the demo (Figure 3(E-G)). The online prototype provides a *how-to guide* to support the interested web users in exploiting the syntax of the IQL language and in formulating complex queries on the underlying *in*Cloud.

4. RELATED WORK

In the last years, a lot of effort has been focused on automatically matching, classifying and representing the contents of a knowledge repository. Examples of interesting solutions in the literature are Google Knowledge Graph, Parallax, gFacet, and Microsoft Pivot. Mainly, the idea of these tools is to work on the presentation aspects of the Web of Data and to provide functionalities for smart browsing of single data in the form of visual interfaces based on graphs, mashups, and histograms. However, these solutions are limited in providing mechanisms for shifting the granularity from a single data item to an aggregated set of objects with a single representative. For this reason, their possible application to real scenarios, especially in large-scale environments, still requires further investigations and development improvements. As possible alternatives, the use of newly-discovered knowledge associations and visual navigation paths is proposed in [6, 8, 9] to provide aggregation-based tools for exploration of DBpedia and Freebase repositories.

With respect to state-of-the-art, we stress that the use of *in*Clouds as data organization structures capable of representing similarity clusters equipped with cluster essentials and proximity links is an enabling solution towards "object-driven" analysis and exploration of linked data. Moreover, the capability of *in*Walk to combine visual, interactive, and thematic exploration functionalities with intuitive and flexible query functionalities through IQL is a new feature in the field of web applications for linked data exploration.

5. CONCLUDING REMARKS

The *in*Walk demo presented in the paper is available online (`http://islab.di.unimi.it/inwalk`). The *in*Walk system is based on *in*Cloud, a flexible data structure that is also suited to work with data other than linked data. For instance, we experienced the use of *in*Walk to explore a dataset of social data extracted from Twitter. This further dataset is available for exploration in the online *in*Walk system.

6. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *Int. Journal on Semantic Web and Information Systems*, 5(3), 2009.

[2] S. Castano, A. Ferrara, and S. Montanelli. Clouding Services for Linked Data Exploration. In *Proc. of the Int. Conference on Advanced Information Systems Engineering (CAiSE 2012)*, Gdansk, Poland, 2012.

[3] S. Castano, A. Ferrara, and S. Montanelli. *Search Computing - Broadening Web Search*, chapter Thematic Clustering and Exploration of Linked Data. Springer, 2013.

[4] S. Davies, J. Hatfield, C. Donaher, and J. Zeitz. User Interface Design Considerations for Linked Data Authoring Environments. In *Proc. of the Int. WWW-LDOW Workshop*, Raleigh, NC, USA, 2010.

[5] A. Ferrara, L. Genta, and S. Montanelli. Linked Data Classification: a Feature-based Approach. In *Proc. of the Int. EDBT-LWDM Workshop*, Genoa, Italy, 2013.

[6] C. Hirsch, J. Hosking, and J. Grundy. Interactive Visualization Tools for Exploring the Semantic Graph of Large Knowledge Spaces. In *Proc. of the IUI-VISSW Workshop*, Florida, USA, 2009.

[7] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the Pedantic Web. In *Proc. of the Int. WWW-LDOW Workshop*, Raleigh, NC, USA, 2010.

[8] N. Marie, F. Gandon, M. Ribière, and F. Rodio. Discovery Hub: on-the-fly Linked Data Exploratory Search. In *Proc. of the 9th Int. Conference on Semantic Systems (ISEM 2013)*, Graz, Austria, 2013.

[9] R. Mirizzi et al. Semantic Wonder Cloud: Exploratory Search in DBpedia. In *Proc. of the Int. ICWE-SWIM Workshop*, Vienna, Austria, 2010.