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 inWalk, an interactive web-based system for linked data exploration featured by the notion of inCloud 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 inWalk, an interactive system for the exploration of linked data based on the notion of inCloud. An inCloud 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 inWalk functionalities that are demonstrated.

2. THE INWALK SYSTEM
inWalk is suited to support both skilled and non-skilled users on two critical aspects related to linked data exploration. Two main contributions of inWalk are the target of this demo. On one side, the goal of inWalk 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, inWalk 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 inWalk system is characterized by the following key features.

Abstraction-by-aggregation Given a set of linked data \( S \), the inWalk system provides a high-level, conceptual view of such data through the notion of inCloud [2]. An inCloud 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 inCloud is the result of an abstraction process based on linked data aggregation. A node \( n_i = (cl_i, d_i) \) of an inCloud 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 inCloud. 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.

Exploration-by-walks The inWalk system enforces linked data exploration by relying on two possible actions, namely the inside walk and the thematic walk, defined over a considered inCloud. 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 inWalk interface listing the complete essential \( d_i \) as well as all the linked data contained in \( cl_i \). The user can select an

(c) 2014, Copyright is with the authors. Published in Proceeding of the 17th International Conference on Extending Database Technology (EDBT 2014) on OpenProceedings.org. Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.
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 inCloud. By selecting a cluster, the corresponding neighborhood is shown on the borderline of the cluster circle. The neighborhood of a cluster $c_i$ is the set of clusters directly connected to $c_i$, through a proximity link in the inCloud. The borderline of the cluster $c_i$, 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 inCloud. 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 inWalk to the linked data view in the repository of provenance.

Filtering-by-patterns The inWalk system supports filtering operations over the inCloud structure to enable the user in focusing the exploration of a portion of inCloud that satisfy a certain selection criterion. The inCloud 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 “a 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 inWalk 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 inWalk and it is in charge of constructing inClouds. 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 inCloud for subsequent walk exploration through the application interface component. The extraction manager acquires from a repository $R$ the linked data set $S$ upon which the inCloud is built. The classification manager first executes a data matching step over the set $S$ 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 inCloud (see [5] for more technical details). Finally, the abstraction manager finalizes the construction of the inCloud graph structure out of similarity clusters.

The web application interface is the front-end component of inWalk and it has the role of supporting the web users in the exploration of inClouds (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 inClouds 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-
ity degree. The query assistant is responsible of supporting the user in selecting the portion of the inCloud 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 inWalk repository and the application memory.

3. THE INWALK SEARCH FUNCTIONALITIES
The web interface of inWalk supports users in exploring and searching a given inCloud.

At startup, the inWalk interface asks the user to select the inCloud to walk from a list of available inClouds stored in the inWalk repository. Since the inCloud 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 inCloud 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 inCloud 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 inCloud as selection threshold. The complete inCloud can be visualized by selecting the Full view button of the interface (Figure 2(E)). The default portion of inCloud (sub-cloud) shown at the inWalk 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 inWalk will be described by addressing specific user tasks, corresponding to increasing levels of complexity of the inWalk functionalities. The inWalk 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 inCloud, 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). 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 inWalk (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

\[ ^2 \text{The numeric index is used to distinguish the clusters resulting with the same (label, type) descriptor.} \]
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 inClouds 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 inWalk 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 inWalk demo presented in the paper is available online (http://islab.di.unimi.it/inwalk). The inWalk system is based on inCloud, a flexible data structure that is also suited to work with data other than linked data. For instance, we experienced the use of inWalk to explore a dataset of social data extracted from Twitter. This further dataset is available for exploration in the online inWalk system.

6. REFERENCES