Thematic Exploration of Linked Data

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

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

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

ABSTRACT

Now that a huge amount of data is available in the Linked Data Cloud, providing techniques for its effective exploration is becoming more and more important. In this paper, we propose aggregation and abstraction techniques for thematic exploration of linked data. These techniques transform a basic, flat view of a potentially large set of messy linked data for a given search target, into a high-level, thematic view called *in*Cloud. In an *in*Cloud, thematic exploration is guided by few *essentials* auto-describing their *prominence* for the search target and by their reciprocal *proximity* relations.

Categories and Subject Descriptors

H.5 [**Information Systems**]: Information Interfaces and Presentation; H.3 [**Information Systems**]: Information Storage and Retrieval

General Terms

Linked data aggregation, labeling, and exploration

1. INTRODUCTION

The Linked Data paradigm promoted a new way of exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web, based on URIs (Universal Resource Identifier) and RDF (Resource Description Framework) [1]. Now that a huge amount of data is available in the Linked Data Cloud, providing techniques for effective linked data searching, exploration, and visualization is becoming crucial [7, 9]. In the recent literature, issues related to linked data exploration are getting more and more importance [8, 12]. One of the most challenging questions is to provide effective browsing solutions capable to deal with the inherent flat organization of linked data and to manage the existing huge-sized repositories storing millions of RDF triples.

In this context, we propose abstraction and aggregation techniques to transform a basic, flat view of a potentially large set of messy linked data, into an *in*Cloud, that is, a high-level, thematic view enabling a more effective, theme-driven exploration of the same dataset. Through aggregation techniques, we identify clusters of semantically related linked data in a (even large) collection representing the response to a search target. Through abstraction techniques, we mine suitable *essentials* capturing the theme dealt with a linked data cluster and its relevance for the search target, as well proximity relations reflecting reciprocal degree of closeness between cluster essentials. We motivate the role of *in*Clouds through a real example of linked data collection extracted from the Freebase repository considering Van Gogh as search target. Moreover, we will describe the construction of an *in*Cloud through aggregation and abstraction techniques. Finally, we show how *in*Cloud representation can be used for thematic browsing and exploration of the underlying linked data collection.

2. MOTIVATING EXAMPLE

In a common scenario, the user interested in exploring a linked data repository to satisfy a certain search target usually has to face a long and loosely-intuitive browsing activity. This is due to the inherent flat organization of linked data repositories where the URIs of interest for a given target frequently require the user to follow more than one property link before being explored. In particular, the user exploration is typically characterized by the following steps:

- Submission to the repository of a *search target* (t), namely a keyword (or a list of keywords) that describes the subject of interest for the search. An example of search target is the name of the famous painter Vincent van Gogh.
- Selection of the *seed of interest* (s), namely an URI that represents the "point of origin" for the exploration about the search target. The seed of interest is chosen from the list of URIs returned by the repository as a reply to the search target. In the Freebase linked data repository[1], an example of seed for our target is the URI `/en/vincent_van_gogh`.
- Exploration of the URIs reachable from the seed with the aim to get access to more information about the search target. This requires the user to submit appropriate queries to the repository to extract the seed

[1] `http://www.freebase.com/`.

properties and the URIs directly linked to s through these properties. An example of MQL query for the Freebase repository to extract the artworks directly connected to the seed $s =$ `/en/vincent_van_gogh` through the property `/visual_art/visual_artist/artworks` is [{ "id": "/en/vincent_van_gogh", "type": "/visual_art/visual_artist", "/visual_art/visual_artist/artworks": {}}].

The exploration step can be recursively applied to the visited URIs to progressively discover further URIs at higher distance from the seed according to the user choices and interests.

Due to the huge number of linked data that is usually concerned with a search target, a lot of exploration steps are required to build a (more or less) comprehensive picture of the available information about the target. As an example, in Figure 1, we show the set of linked data extracted from Freebase for the target Vincent van Gogh. In this example, we considered the seed $s =$ `/en/vincent_van_gogh`, we explored the complete set of directly linked URIs and some selected URIs at distance $d = 2$ from s. As it is clear from this simple example, exploring such a flat and huge collection of data is cumbersome. First, because the representation is flat and it is impossible to immediately understand whether some URIs are more important than others. Moreover, possible sets of URIs addressing the same/similar argument about the target are not highlighted nor grouped.

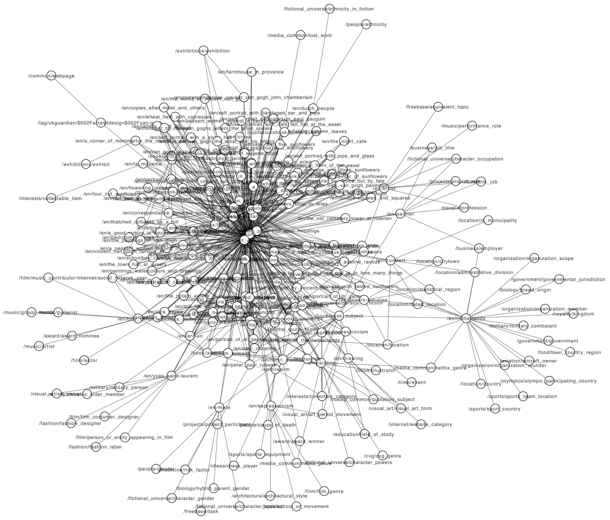

/fictional_universe/ethnicity_in_fiction
/people/ethnicity
/media_common/lost_work
/exhibitions/exhibition
/en/farmhouse_in_provence
/common/webpage
/tag/ukguardian/$002Fartanddesign$002Fvan-g
/en/a_corner_of_montmartre_the_moulin_de
/exhibitions/exhibit
/en/la_mousme
/interests/collectable_item
/en/four_cut_sunflowers
/en/correspondance_generale_tome
/en/thatched_cottages_by_a_hill
/en/a_good_picture_is_equivalent_to_a_good_deed
/en/the_paintings_of_van_gogh
/en/vincent_van_gogh_famous_in
/en/the_town_hall_at_auvers
/en/paintings_watercolours_and_drawings
/film/music_contributor/internet/social_network_user
/music/group_member
/music/guitarist
/award/award_nominee
/music/artist
/film/actor
/en/yves_saint-laurent
/military/military_person
/visual_art/art_owner
/film/film_costumer_designer
/fashion/fashion_designer
/film/person_or_entity_appearing_in_film
/fashion/fashion_label
/people/gender
/medicine/risk_factor
/biology/hybrid_parent_gender
/fictional_universe/character_gender
/freebase/task
/architecture/architectural_style
/film/film_genre
/en/the_works_of_vincent_van_gogh
/en/correspondences_vincent_van_gogh_john_chamberlain
/en/copies_after_millet_and_others
/en/wheat_field_with_cypresses
/en/self_portrait_with_bandaged_ear_and_pipe
/en/self_portrait_dedicated_to_paul_gauguin
/en/self_portrait_with_dark_felt_hat_at_the_easel
/en/falling_autumn_leaves
/en/dutch_people
/en/the_night_cafe
/en/self_portrait_with_pipe_and_glass
/en/vase_with_three_sunflowers
/en/painter_of_sunflowers
/en/artist
/en/double_squares_and_squares
/en/the_old_cemetery_tower_at_nuenen
/en/painter
/freebase/equivalent_topic
/music/performance_role
/business/job_title
/fictional_universe/character_occupation
/people/profession
/location/nl_municipality
/business/employer
/organization/organization_scope
/en/zundert
/location/citytown
/location/administrative_division
/government/governmental_jurisdiction
/biology/breed_origin
/location/statistical_region
/location/dated_location
/en/netherlands
/organization/organization_member
/royalty/kingdom
/military/military_combatant
/government/government
/food/beer_country_region
/aviation/aircraft_owner
/organization/organization_founder
/olympics/olympic_participating_country
/location/country
/sports/sports_team_location
/sports/sport_country
/en/auvers-sur-oise
/book/book_subject
/en/post-impressionism
/location/location
/en/drawing
/en/painting
/book/illustrator
/media_common/netflix_genre
/time/event
/en/the_potato_eaters
/en/arman
/en/peter_paul_rubens
/en/firearm
/en/male
/en/expressionism
/interests/collection_category
/media_common/quotation_subject
/visual_art/visual_art_form
/visual_art/art_period_movement
/internet/website_category
/projects/project_participant
/people/cause_of_death
/education/field_of_study
/award/award_winner
/cvg/cvg_genre
/sports/sports_equipment
/chess/chess_player
/media_common/media_genre
/fictional_universe/character_powers

Figure 1: A graph of linked data extracted from the Freebase repository about the search target Vincent van Gogh

The solution we propose is based on aggregation and abstraction techniques to transform a basic, flat view of linked data like the one in Figure 1, into an *in*Cloud providing a high-level, thematic view of the same data. *in*Clouds are conceived to be coupled with the conventional query interfaces of the existing Linked Data repositories, in that they can be built on top of an extracted dataset to provide a more effective presentation of the result.

An example of *in*Cloud for the seed $s =$ `/en/vincent_van_gogh` is shown in Figure 2. In an *in*Cloud

- a circle-box represents a *a cluster*, namely a group of linked data focused on a specific argument/topic related to the considered seed (e.g., the set of artists that influenced or have been influenced by van Gogh (cluster Cl_3) or the set of van Gogh artworks about sunflowers (Cluster Cl_4);

- a square-box represents an *essential*, namely a concise and convenient summary of the content of a cluster at a glance (e.g., Topic Artwork, Sunflower used to summarize the content of Cluster Cl_4). Clusters in an *in*Cloud are also characterized by a *prominence value* denoting its level of importance for the target in the framework of the overall *in*Cloud. Prominence values determine the size of the cluster circles, thus the most prominent cluster in the *in*Cloud of Figure 2 is Cluster Cl_1;

- an arrow represents a proximity relation between clusters/essentials, namely a closeness relationship between the themes/topics their represent. The arrow thickness denotes the degree of proximity between the two clusters/essentials connected by the arrow.

Aggregation techniques are first employed to enforce a "thematic" clustering of the initial set of linked data, as described in Section 3. *Abstraction techniques* are then applied to synthesize an *in*Cloud over the thematic clusters, as described in Section 4.

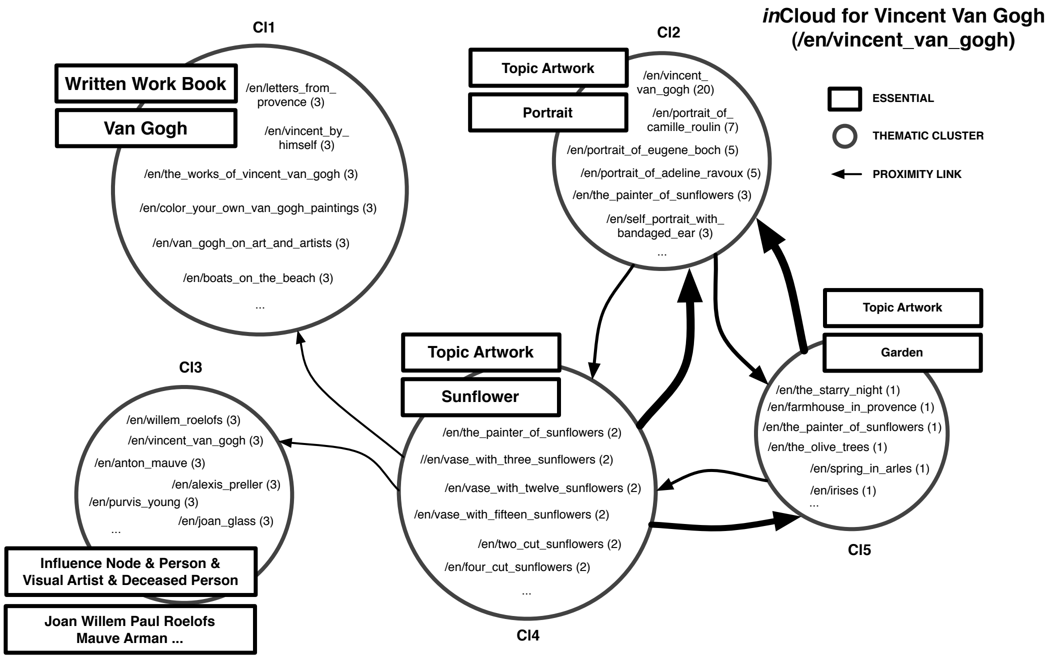
3. LINKED DATA AGGREGATION

The goal of aggregation techniques is to transform an initial set of linked data into a number of thematic clusters. The starting point is a RDF graph $\mathcal{G}_s$ containing the linked data about a certain seed s of interest automatically extracted from a Linked Data repository $\mathcal{R}$. Appropriate extraction queries are defined to this end according to the language (e.g., SPARQL, MQL) supported by the repository $\mathcal{R}$. These queries generally enforce the following extraction/filtering operations:

- *Extraction of properties and corresponding values within a distance $\leq d$ from the seed s.* We consider that an URI in the repository $\mathcal{R}$ is concerned with the seed s if there is a property path of length $\leq d$ between the URI and s. The distance d can be dynamically changed and it has an impact on the number of extracted linked data and thus on the size of the resulting RDF graph. In usual scenarios, a distance $d = 2$ is a good trade-off to obtain a sufficient number of linked data about s and a well-sized RDF graph.

- *Extraction of the URI types.* For each URI within a distance $\leq d$ from the seed s, we extract the list of types (i.e., classes) the URI belongs to. The appropriate property of the repository $\mathcal{R}$ is exploited to this end (e.g., the property type in Freebase).

- *Filtering of non-relevant properties.* Loosely meaningful properties of a repository, like the property *image* of Freebase, can be excluded from the resulting RDF graph since they are poorly useful in providing information about s.



*in*Cloud for Vincent Van Gogh
(/en/vincent_van_gogh)
ESSENTIAL
THEMATIC CLUSTER
PROXIMITY LINK
Cl1
Written Work Book
Van Gogh
/en/letters_from_provence (3)
/en/vincent_by_himself (3)
/en/the_works_of_vincent_van_gogh (3)
/en/color_your_own_van_gogh_paintings (3)
/en/van_gogh_on_art_and_artists (3)
/en/boats_on_the_beach (3)
...
Cl2
Topic Artwork
Portrait
/en/vincent_van_gogh (20)
/en/portrait_of_camille_roulin (7)
/en/portrait_of_eugene_boch (5)
/en/portrait_of_adeline_ravoux (5)
/en/the_painter_of_sunflowers (3)
/en/self_portrait_with_bandaged_ear (3)
...
Cl3
/en/willem_roelofs (3)
/en/vincent_van_gogh (3)
/en/anton_mauve (3)
/en/alexis_preller (3)
/en/purvis_young (3)
/en/joan_glass (3)
...
Influence Node & Person & Visual Artist & Deceased Person
Joan Willem Paul Roelofs Mauve Arman ...
Cl4
Topic Artwork
Sunflower
/en/the_painter_of_sunflowers (2)
//en/vase_with_three_sunflowers (2)
/en/vase_with_twelve_sunflowers (2)
/en/vase_with_fifteen_sunflowers (2)
/en/two_cut_sunflowers (2)
/en/four_cut_sunflowers (2)
...
Cl5
Topic Artwork
Garden
/en/the_starry_night (1)
/en/farmhouse_in_provence (1)
/en/the_painter_of_sunflowers (1)
/en/the_olive_trees (1)
/en/spring_in_arles (1)
/en/irises (1)
...

Figure 2: An example of *in*Cloud extracted from the Freebase repository for the seed /en/vincent_van_gogh

The query result is the graph $\mathcal{G}_s = (N_s, E_s)$ where a node $n \in N_s$, called *linked data entity*, can be an URI, a literal, or a type value that satisfy the query selection, and an edge $e(n_i, n_j) \in E_s$, called *property link*, represents a property relationship of $\mathcal{R}$ between the nodes $n_i, n_j \in N_s$.

Based on the RDF graph $\mathcal{G}_s$, linked data aggregation is articulated in two main steps, namely *similarity evaluation* and *thematic clustering*.

3.1 Similarity evaluation

This step has the goal to analyze the graph $\mathcal{G}_s$ and to generate an *augmented linked data graph* $\mathcal{G}_s^+$ where a *similarity link* is added between each pair of matching linked data entities in N_s. To this end, the level of affinity between the entities of N_s is evaluated as follows. Given two linked data entities n_i, $n_j \in N_s$, the *linked data affinity* $\sigma(n_i, n_j) \in [0, 1]$ denotes the level of similarity of n_i and n_j based on the commonalities of their *terminological equipments*. Each linked data entity $n \in N_s$ is associated with a terminological equipment $\mathsf{Term}_n = \{term_1, \ldots, term_m\}$ where $term_j$, with $1 \leq j \leq m$, is a term appearing in the label of a node adjacent to n in $\mathcal{G}_s$, or a term appearing in the label of n itself. Before inclusion in a terminological equipment, each term is submitted to a normalization procedure for word-lemma extraction and for compound-term tokenization [4, 15].

The affinity σ of two linked data entities n_i, $n_j \in N_s$ is calculated as the Dice coefficient over their terminological equipments as follows:

$$\sigma(n_i, n_j) = \frac{2\cdot \mid term_x \sim term_y \mid}{\mid \mathsf{Term}_{n_i} \mid + \mid \mathsf{Term}_{n_j} \mid}$$

where $term_x \sim term_y$ denotes that $term_x \in \mathsf{Term}_{n_i}$ and $term_y \in \mathsf{Term}_{n_j}$ are matching terms according to a string matching metric that considers the structure of the terms $term_x$ and $term_y$. 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 [2]. A similarity link $e(n_i, n_j)$ is established between the linked data entities n_i and n_j iff $\sigma(n_i, n_j) \geq th$ where $th \in (0, 1]$ is a matching threshold denoting the minimum level of similarity required to consider two linked data entities as matching entities.

3.2 Thematic aggregation

This step has the goal to analyze the graph $\mathcal{G}_s^+$ obtained through similarity evaluation and to identify/mine a set CL of thematic clusters. Given a graph $\mathcal{G}_s^+$, a cluster Cl $= \{(n_1, f_1), \ldots, (n_h, f_h)\}$ is a set of linked data entities $n_1, \ldots, n_h \in N_s$ that are more similar to each other than to the other entities of N_s. Each entity n_j belonging to Cl is associated with a corresponding frequency f_j which denotes the number of occurrences of n_j in Cl.

Clusters are determined by exploiting the graph $\mathcal{G}_s^+$ and by detecting those node regions that are highly interconnected through property/similarity links. The problem of thematic aggregation is analogous to the problem of cluster calculation, also known as *module*, *community*, or *cohesive group*, in graph theory. For this reason, for thematic aggregation, we rely on a clique percolation method (CPM) [13]. The CPM is based on the notion of *k-clique* which corresponds to a complete (fully-connected) sub-graph of k nodes within the graph $\mathcal{G}_s^+$. Two k-cliques are defined as *adjacent k-cliques* if they share $k - 1$ nodes. The CPM determines clusters from k-cliques. In particular, a cluster, or more precisely, a k-clique-cluster, is defined as the union of all k-cliques that can be reached from each other through a series of adjacent k-cliques. As a consequence, a typical k-clique-cluster is composed of several cliques (with size $\leq k$) that tend to share many of their nodes. Since the cliques of a graph can share one or more nodes, we observe that a node can belong to several clusters, and thus clusters can overlap. In our approach, we employ the CPM implemented in the CFinder tool[2]. Although the determination of the full set of cliques of a graph is widely believed to be a non-polynomial problem, CFinder proves to be efficient when applied to graphs like those considered in our approach. Such an algorithm is based on first locating all complete subgraphs of $\mathcal{G}_s^+$ that are not part of larger complete subgraphs, and then on identifying existing k-clique-clusters by carrying out a standard component analysis of the clique-clique overlap matrix [6]. As a result, CFinder produces the full set CL of k-clique-clusters existing in the graph $\mathcal{G}_s^+$ for all the possible values of k. A linked data entity n_i belonging to a cluster $Cl \in CL$ is represented as a pair (n_i, f_i) where the frequency value f_i denotes the number of cliques of Cl which the entity n_j belongs to (see Example of Figure 2). The entities of a cluster are represented with different sizes, proportional to the corresponding frequency values according to a visualization manner "à la tag-cloud"[3].

4. LINKED DATA ABSTRACTION

The goal of linked data abstraction techniques is to build an *in*Cloud, namely a high-level view on top of linked data clusters by synthesizing them through essentials. *in*Cloud clusters are also featured by a level of prominence and by proximity relations that denote the level of overlapping of the different clusters.

4.1 Essential abstraction

An essential Ess_i is a concise and convenient summary of a thematic cluster Cl_i and it is defined as a pair of the form $Ess_i = (C_i, D_i)$ where C_i is the category associated with Cl_i and D_i is a descriptor associated with Cl_i. A category C_i is a set composed by the labels of the most frequent types of the linked data entities in Cl_i, while a descriptor D_i is a set composed by the most frequent terms in the terminological equipments of the entities in Cl_i. If more than one most equally-frequent type and/or term exist, they are all inserted in C_i and D_i, respectively. In the example of Figure 2, the cluster Cl_4 corresponds to a very focused theme expressed by the essential category Topic Artwork (the most frequent type of the entities in the cluster) and by the essential descriptor Sunflower (the most frequent term in the terminological equipments of the entities in Cl_4). In cases where many entities are equally frequent in a cluster, the abstracted essential is less focused and contains more terms. This is the case for example of the cluster Cl_3 of Figure 2, representing persons and visual artists influenced by Van Gogh. In this case, the most frequent terms used as descriptors are the names of the people involved in the cluster, which are all equally frequent in the cluster.

4.2 Prominence evaluation

Clusters (and related essentials) in an *in*Cloud are differently relevant with respect to the original search target. In order to represent this fact, we introduce the notion of prominence of a cluster, namely a value $P_i \in [0, 1]$. The higher P_i is, the higher is also the prominence of Cl_i in the *in*Cloud. In our approach, the level of prominence of a cluster is higher when the cluster is very focused on its theme and its contents are homogeneous. In particular, we formalize two cluster properties that are *variability* and *density.*

Variability v_i is the degree of overlap among the cliques of the cluster Cl_i. For a linked data entity $n_j \in \mathcal{N}_s^+$, we call f_j the frequency of n_j, that is the number of cliques of Cl_i that contain n_j. Variability v_i is measured by a coefficient of variation, which is the ratio between the standard deviation of the linked data entity frequencies in Cl_i and the arithmetic mean of those frequencies, as follows (with $\overline{f}$ denoting the arithmetic mean value of frequencies):

$$v_i = \frac{1}{\overline{f}} \sqrt{\frac{1}{N_i - 1} \sum_{i=1}^{N_i} (f_i - \overline{f})^2}$$

According to this definition, high values of v_i denote a low degree of overlap in the cliques of the cluster Cl_i, while low values of v_i denote a high degree of overlap in the Cl_i cliques.

Density d_i of a cluster Cl_i is the degree of interconnection among the linked data entities of Cl_i. The density coefficient $d_i = 2 \cdot R_i / N_i(N_i - 1)$ is the ratio between the number R_i of links in the cluster Cl_i and the maximum number of possible links. According to this definition, high values of d_i denote a high degree of interconnection among the cluster Cl_i entities, while low values of d_i denote a low degree of interconnection. The prominence P_i of a cluster Cl_i is calculated on the basis of its variability and density as follows:

$$P_i = \frac{2 \cdot (1 - v_i) \cdot d_i}{(1 - v_i) + d_i}$$

According to this approach, most prominent clusters are those which are more focused and homogeneous with respect to their theme. We graphically represent cluster prominence by drawing circles proportional to the prominence values of the corresponding clusters. In our example of Figure 2, clusters Cl_1 and Cl_4 are more prominent (larger circles) because they are more focused and homogeneous. On the opposite, clusters like Cl_3, which collect several entities of different types are considered less prominent (smaller circle). However, other options are possible for the evaluation of prominence in case of specific application needs. A first option is to consider a cluster to be more prominent as it is more close to the seed s of interest. In this case, the prominence P_i of a cluster Cl_i is evaluated by taking into account the average value of similarity between the linked data entities in the cluster Cl_i and s, weighted by the frequency of each entity n_i in Cl_i, as follows:

$$P_i = \frac{\sum_{p=1}^{N_i} \sigma(n_i, s) \cdot f_i}{\sum_{p=1}^{N_i} f_i}$$

where f_i denotes the frequency of the linked data entity n_i in the cluster Cl_i. Another option is to consider the prominence P_i of a cluster Cl_i as proportional to the dimension

[2] Available at http://www.cfinder.org/.

[3] For a more readable visualization of highly-populated clusters, the representation of less-frequent linked data entities can be omitted.

N_i of Cl_i and to the size k_i of the smaller clique in Cl_i, as follows: $P_i = 2 \cdot N_i \cdot k_i / N_i + k_i$.

4.3 Proximity relations

In an *in*Cloud, clusters (and consequently their associated essentials) are connected by reciprocal proximity relations, which represent the degree of overlapping between them. In particular, given two clusters Cl_i and Cl_j, the degree of proximity $X_{ij} =\mid Cl_i \cap Cl_j \mid / \mid Cl_i \mid$ between Cl_i and Cl_j is proportional to the number of linked data entities common to Cl_i and Cl_j over the number of linked data entities in Cl_i. The greater the level of overlapping between Cl_i and Cl_j, the higher the degree of their proximity relation. Proximity relations are graphically represented by arrows with thickness proportional to the proximity degree. In Figure 2, we can see how proximity relations connect those clusters that are more semantically related to each other, such as Cl_2, Cl_4, and Cl_5 which all represent different types of artworks by Vincent van Gogh.

5. USING INCLOUDS FOR THEMATIC EXPLORATION

In this section, we discuss how *in*Clouds can be exploited for thematic exploration of linked data and we provide some considerations about the applicability of the *in*Cloud approach in the large-scale scenario.

5.1 Thematic exploration through inClouds

An *in*Cloud enables different exploration modalities that can be switched on according to the specific user preferences. In particular, the following modalities are defined.

- *Exploration-by-essential.* This is the most intuitive exploration modality and it is based on cluster essentials. A user can consider each essential as a sort of instantaneous picture of the associated cluster and linked data therein contained, thus allowing the user to rapidly choose the most preferred one for starting the exploration.

- *Exploration-by-prominence.* This modality allows the user to organize the exploration according to the prominence values associated with the clusters. The idea is to support the user in moving throughout the clusters according to their relevance with respect to the set of considered linked data. As discussed in Section 4, different criteria can be used to calculate the prominence value. The capability to switch from one criterion to another allows the user to dynamically re-organize the *in*Cloud in light of a different notion of cluster prominence.

- *Exploration-by-proximity.* This modality enables the user to choose a cluster and to browse its constellation, by exploiting the proximity relations. When a user is exploring a certain cluster, the proximity relations provide indication of its fully/partially overlapping neighbors, thus suggesting the possible exploration of clusters that are somehow related in content.

5.2 Linked data exploration in-the-large

The presented *in*Cloud approach can be also exploited for applicability in the large scale scenario. In particular, extension to the multi-repository exploration and to the multi-seed extraction can be performed.

Extension to multi-repository exploration. For a more complete visualization of the available linked data about a certain search target, multiple RDF repositories can be queried to originate a unique, comprehensive *in*Cloud. In the Linked Data Cloud, the property owl:sameAs is used to denote when a linked data entity n_i belonging to a certain RDF repository $\mathcal{R}$ and another entity n_j belonging to a different repository $\mathcal{R}'$ refer to the same real-world object. In a multi-repository scenario, the construction of the graph $\mathcal{G}_s$ can take into account the owl:sameAs relations as a sort of “natural join” operation. The idea is to start the construction of $\mathcal{G}_s$ by querying an initial repository $\mathcal{R}$ and to exploit the owl:sameAs relations to extend the linked data extraction to other RDF repositories. In particular, the URIs connected by a owl:sameAs relation are collapsed in a unique linked data entity of $\mathcal{G}_s$ and the extraction/filtering operations described in Section 3 are applied to the whole set of linked data extracted by the considered RDF repositories.

Extension to multi-seed extraction. In some cases, the user can be interested in exploring the available linked data about more than one seed of interest. In this framework, the *in*Cloud mechanism can be used to build a comprehensive thematic picture that takes into account all the seeds of interest. In a multi-seed scenario, the starting point is a set of seeds $S = \{s_1, \ldots, s_k\}$. The graph $\mathcal{G}_s$ is built by executing the extraction/filtering operations of Section 3 for each element $s_i \in S$. Depending on the seeds of interest, one or more portions of the graph $\mathcal{G}_s$ can be disjoint from the rest of the graph. In particular, when the seeds in S are about completely different arguments, a separate independent cluster is generated through aggregation for each $s_i \in S$. In such a limit case, the usefulness of the *in*Cloud mechanism for exploration is in the capability of providing an effective synthetic essential for each seed $s_i \in S$ and in calculating the relative prominence of each seed with respect to the others.

We stress that linked data exploration in-the-large can require the execution of thematic aggregation techniques over a starting RDF graph $\mathcal{G}_s$ containing a huge number of nodes (e.g., thousands of linked data entities). The clique percolation method we use for cluster calculation best performs when a small-medium number of nodes in the graph $\mathcal{G}_s$ is considered (e.g., hundreds of linked data entities). For example, in our tests, the CPM over a graph $\mathcal{G}_s$ containing 200 nodes takes an execution time of 200ms (considering a matching threshold th=0.9). For linked data exploration in-the-large, when 1.000 (or more) nodes are considered, more efficient clustering algorithms, like hierarchical clustering, can be exploited (see [3] for further details).

6. RELATED WORK

Problems and solutions more strictly related to our work are focused either on improving search and retrieval of information in the Linked Data cloud [14] or on browsing and presentation of linked data contents [5]. Search and retrieval is moving from traditional information lookup to exploratory search, defined as the activity of finding and understanding knowledge about a topic of interest by exploiting aggregation and learning of information in a social context [11]. In this respect, for example, Sig.ma (Semantic Information MAshup) [16] retrieves and integrates linked data, starting from a single URI, by querying the Web of Data and applying machine learning to the data found. In a similar direction, structured and collaborative search engines are being emerging as a promising solution for presenting the query results in a sort of structured form and focusing on the understanding of the user information need. Examples in this field are Wolfram Alpha (`http://www.wolframalpha.com`), Google Wonder Wheel (`http://www.googlewonderwheel.com`), and YAGO2 (`http://www.mpi-inf.mpg.de/yago-naga/yago`). Another category of related work includes approaches aiming at presenting linked data in a more intuitive way. Examples of solutions in this respect are [8, 12] and Freebase Parallax (`http://www.freebase.com/labs/parallax/`), where tools that help users in exploring DBpedia and Freebase are presented, not only via directed links in the RDF dataset, but also via newly discovered knowledge associations and visual navigation. These tools exploit aggregation techniques in order to combine related topics in unified nodes, providing also a textual description of each node. In other approaches, like Marbles (`http://www5.wiwiss.fu-berlin.de/marbles`) and LESS (`http://less.aksw.org`), information about resources of interest is presented exploiting HTML and RSS and by using different colors to distinguish sources.

With respect to the related work, our contribution regards the use of data similarity, proximity, and prominence techniques for *in*Cloud construction, to move from a basic, flat organization of linked data to a high-level, thematic view of them. Moreover, the proposed techniques allow the different themes/topics to directly emerge from the original linked data and their mutual links, by suggesting also an intuitive visualization of data contents in terms of essentials, which synthesize the contents of thematic clusters.

7. CONCLUDING REMARKS

In this paper, we presented *in*Clouds, high-level views of linked data enabling their thematic exploration. Ongoing work is focused on finalizing the development of a web application fully covering the steps of linked data aggregation and abstraction required for *in*Cloud construction. By exploiting an initial prototype implementation, we run some experiments concerning user evaluation of *in*Clouds based on standard user-oriented evaluation methods for interactive web search interfaces and systems [10]. Initial results are promising and *in*Clouds are seen by real users as a valid support to the satisfaction of users information needs [3]. Moreover, ongoing research activity regards the extension of the *in*Cloud approach to consider additional kinds of web data contents, like microdata, microblogging posts, and news. The idea is to propose *in*Clouds as a comprehensive exploration tool considering also actual, up-to-date social web information about the search target for possible fruition in the framework of event-promoting applications.

8. 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):1–22, 2009.

[2] S. Castano, A. Ferrara, and S. Montanelli. Matching Ontologies in Open Networked Systems: Techniques and Applications. *Journal on Data Semantics*, V:25–63, 2006.

[3] S. Castano, A. Ferrara, and S. Montanelli. Structured Data Clouding across Multiple Webs. Technical report, Università degli Studi di Milano, 2011.

[4] S. Castano and G. Varese. *Next Generation Data Technologies for Collective Computational Intelligence*, chapter Building Collective Intelligence through Folksonomy Coordination, pages 87–112. Springer, 2011.

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

[6] B. Everitt. *Cluster Analysis*. Edward Arnold, London, UK, 3rd edition, 1993.

[7] W. Halb, Y. Raimond, and M. Hausenblas. Building Linked Data for both Humans and Machines. In *Proc. of the WWW Int. Workshop on Linked Data on the Web (LDOW 2008)*, Beijing, China, 2008.

[8] C. Hirsch et al. Interactive Visualization Tools for Exploring the Semantic Graph of Large Knowledge Spaces. In *Proc. of the IUI Int. Workshop on Visual Interfaces to the Social and the Semantic Web*, Sanibel Island, USA, 2009.

[9] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the Pedantic Web. In *Proc. of the WWW Int. Workshop on Linked Data on the Web (LDOW 2010)*, Raleigh, NC, USA, 2010.

[10] A. Leclercq. he perceptual evaluation of information systems using the construct of user satisfaction: case study of a large french group. *ACM SIGMIS Database*, 38(2):27–60, 2007.

[11] G. Marchionini. Exploratory Search: from Finding to Understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[12] R. Mirizzi, A. Ragone, T. Di Noia, and E. Di Sciascio. Semantic Wonder Cloud: Exploratory Search in DBpedia. In *Proc. of the ICWE 2nd Int. Workshop on Semantic Web Information Management (SWIM 2010)*, pages 138–149, Vienna, Austria, 2010.

[13] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. *Nature*, 435:814–818, 2005.

[14] D. Petrelli, S. Mazumdar, A. Dadzie, and F. Ciravegna. Multi Visualization and Dynamic Query for Effective Exploration of Semantic Data. In *Proc. of the 8th Int. Semantic Web Conference*, pages 505–520, Chantilly, VA, USA, 2009.

[15] S. Sorrentino et al. Schema Normalization for Improving Schema Matching. In *Proc. of the 28th Int. ER Conference*, pages 280–293, Gramado, Brazil, 2009.

[16] G. Tummarello et al. Sig. ma: Live Views on the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4):355–364, 2010.