Exploratory search on topics through different perspectives with DBpedia

Nicolas Marie, Fabien Gandon, Alain Giboin, Emilie Palagi

Exploratory search on topics through different perspectives with DBpedia

Nicolas MARIE
Alcatel-Lucent Bell labs
INRIA Sophia-Antipolis
France
nicolas.marie@inria.fr

Fabien Gandon, Alain Giboin, Émilie Palagi
INRIA Sophia-Antipolis
France
firstname.lastname@inria.fr

ABSTRACT

A promising scenario for combining linked data and search is exploratory search. During exploratory search, the search objective is ill-defined favorable to discovery. A common limit of the existing linked data based exploratory search systems is that they constrain the exploration through single results selection and ranking schemes. The users can not influence the results to reveal specific aspects of knowledge that interest them. The models and algorithms we propose unveil such *knowledge nuances* by allowing the exploration of topics through several perspectives. The users adjust important computation parameters through three operations that help retrieving desired exploration perspectives: specification of interest criteria about the topic explored, controlled randomness injection to reveal unexpected knowledge and choice of the processed knowledge source(s). This paper describes the corresponding models, algorithms and the Discovery Hub implementation. It focuses on the three mentioned perspective-operations and their evaluations.

Categories and Subject Descriptors

[**Graph Theory**]: Graph algorithms; [**HCI design and evaluation methods**]: User studies

General Terms

Algorithms, experimentation

Keywords

Exploratory search, Multi-perspectives exploration, DBpedia, Discovery engine, Discovery Hub

1. INTRODUCTION

As stated by White in [16]: "*search is only a partially solved problem*". *Exploratory search* refers to cognitively consuming search tasks like learning or investigation [9]. The term learning is employed in the broad sense and can concern educational, professional and personal contexts. During exploratory search the search objective is ill-defined, favorable to discovery. Actual popular search engines are evolving in the right direction but are still not supporting exploratory search efficiently today. This is notably due to their keyword-search paradigm and lack of assistance during the results consultation. First it is impossible to capture complex information needs in few keywords, especially for vague ones. Second the users have to synthesize an important amount of information without support from the system. They have to rely on their own search strategies leading to a considerable cognitive load. It increases the information integration work needed to understand and to use the information collected: *"the human user's brain is the fundamental platform for information integration"* [3]. There is a need to complete the actual widely-used solutions by designing and popularizing systems optimized for exploratory search tasks. This challenge requires contributions from several research fields including in particular information retrieval and interaction design[1].

Linked data [1] and DBpedia [8] have been extensively described in the scientific literature and respectively correspond to an approach for publishing and linking data on the web and its application to data extracted from Wikipedia[2]. The improvement of search through incorporation of semantics is referred to as *semantic search*. Semantic search has been the subject of numerous researches targeting a wide range of search objectives [14]. Today important initiatives emerge from major players including the knowledge graphs-based search engines functionalities (based on Bing Satori[3], Google[4] and Yahoo Knowledge Graphs [2]), Apple Siri[5] and the Facebook Graph Search[6]. Such technologies and functionalities open the door to a better support of exploratory search tasks ("*explore your search*", "*help you research a topic more in depth than before*"[4]). Even if some of these graphs are partially built from open data sources they are not publicly accessible and are consequently not part of the linked open data cloud.

Linked data are promising for supporting exploratory search. Their richness and structured aspect allow proposing new al-

SEM '14, September 04 - 05 2014, Leipzig, AA, Germany

http://dx.doi.org/10.1145/2660517.2660518.

[1]https://sites.google.com/site/hcirworkshop/
[2]http://www.wikipedia.org/
[3]http://www.bing.com/blogs/site_blogs/b/search/archive/2013/03/21/satorii.aspx
[4]google.com/insidesearch/features/search/knowledge.html
[5]http://www.apple.com/ios/siri/
[6]https://www.facebook.com/about/graphsearch

gorithms and interaction models optimized for exploration purposes. The research in this field still faces several challenges. One limit is that existing linked data based systems often offer only one exploration perspective i.e. the users can not or hardly influence the queries' results in a *direction* of interest. This paper supports (1) the idea that a plurality of relevant exploration perspectives can be offered to the users starting from a topic of interest, (2) that some linked data datasets constitute a valuable source of knowledge for such multi-perspectives exploratory search. Indeed, the objects described in linked data datasets can be rich, complex and approached in many manners. For example, a user can be interested in a painter (e.g. Claude Monet or Mary Cassat) in many ways: works, epoch, movement, entourage, social or political contexts and more. The user may also be interested by basic information or by unexpected and unusual ones depending on his actual knowledge about the painter. He may also want to explore the topic through a specific culture or area e.g. impressionism in American or French culture. A single interest can be explored through many perspectives corresponding to different *knowledge nuances*. In the graph context of linked data these perspectives correspond to different non exclusive sets of objects and relations that are informative on a topic regarding specific aspects.

In our proposition such exploration perspectives are obtained by increasing the results relevance thanks to users' query refinement/personalization and provoking discoveries by injecting randomness during the results' computation. Such topics are not new in the general field of information retrieval but, to the best of our knowledge, no approaches were formalized, implemented and evaluated in the context of linked data based exploratory search. This flexibility in the query processing is reached thanks to a framework that computes the results at query time and that is loosely coupled to the data source queried.

2. STATE-OF-THE-ART

In this state-of-the-art we consider exploratory search systems (ESS) and recommenders. As mentioned in [11] despite their differences recommenders and exploratory search systems show interesting intersections. They share the objective of assisting the users in information or resource discovery in a collection. However they achieve it in very different ways. The recommenders provide direct suggestions that require no or minimal users interactions whereas the exploratory search systems lead to discoveries through users' engagement and high interactivity. Moreover some linked data-based exploratory search systems integrate recommendation functionalities e.g. query terms recommendation for assisting and inspiring the users in their query formulation as in LED[7]. Thus it is interesting to review linked-data based recommenders today as few exploratory search systems based on such data exist.

A list of recommenders and exploratory search systems based on linked data was presented in [10]. In the present paper we focus on applications having a user interface allowing the users to define the perspective they want to explore about topics of interest. By performing this state-of-the-art, we wanted in particular to review how existing recommenders' and exploratory search systems implemented such a *perspective-setting*, and to determine how this perspective-setting could be improved. The systems presented hereafter are summarized in Table 1. We observed that the following limits are recurrent:

- The results selection process is generally fixed.
- The ranking scheme is generally fixed.
- The data source(s) used to process the query is fixed.

LED (Lookup Explore Discover) [11] is an exploratory search system that suggests related query-terms starting from a user initial query. It implements the DBpedia Ranker algorithm for such recommendations. Seevl[8] is a music discovery platform that offers DBpedia-based artists' recommendations thanks to the DBrec algorithm [13]. Yovisto[9] is an academic video platform offering related-queries suggestions computed with a set of heuristics on the German and English DBpedia chapters [15]. These 3 applications do not allow the users to influence the results retrieved.

Starting from a resource of interest Aemoo[10] visually presents its direct neighborhood filtered with semantic-based patterns called Encyclopedic Knowledge Patterns (EKPs) [12]. EKPs are selections of the most informative classes regarding a specific class: *"the most relevant types of things that people use for describing other things"*. For instance the DBpedia *Actor* class EKP includes the classes *Actor*, *City*, *Film*, *TelevisionShow*, etc. Aemoo proposes a *"curiosity"* function which displays the queried resource neighborhood through an inverted EKP filtering. This function offers an exploration perspective that aims to unveil unexpected knowledge. The MORE movie recommender[11] [11] is based on a semantic adaptation of the Vector Space Model and allows the users to tune the importance of each vector/property for the recommendation e.g. *director*, *genre*, *starring*.

There is a certain lack of flexibility in the existing approaches based on linked data. In other words considering a topic, captured in the form of a resource, there is only one or few processing and consequently one or few results space(s) available. This lack of flexibility is notably due to the fact that the results are pre-computed and stored for later retrieving. The users retrieve the pre-computed results and are consequently not able to influence the computation parameters. We propose a flexible framework offering several exploration perspectives on the users' topics of interest. Contrary to other approaches the results are computed at query-time allowing the users to have the hand on several parameters through the interface of an application implementing the framework.

3. RELEVANCE AND SURPRISE AS USER-CENTRIC CRITERIA

3.1 Framework basis

The framework we propose for linked data based exploratory search was described in [10], its general architecture is presented on Figure 1. It is based on a semantic-sensitive spreading activation algorithm that is coupled to an incremental and live graph sampling technique. Spreading activation is a class of algorithms that iteratively diffuse a value

[7]http://sisinflab.poliba.it/led/

[8]http://seevl.fm

[9]http://www.yovisto.com/

[10]http://wit.istc.cnr.it/aemoo/

[11]https://apps.facebook.com/new-more

Table 1: Summarization of closest systems

Application	Aemoo	LED	MORE	Seevl	Yovisto
Purpose	ESS	ESS	Recommender	Recommender	Video exploration
Data	DBpedia subset	DBpedia subset	DBpedia subset	DBpedia subset	DBpedias subset
Data choice	No	No	No	No	No
Algorithm	EKP-based view	DBpedia Ranker	sVSM	DBrec	Heuristics
Computation	Offline	Offline	Offline	Offline	Offline
Perspect.-setting	Yes	No	Yes	No	No
Perspect. number	2	1	n	1	1
Perspectives types	Core, curiosity	/	Property weighting	/	/

initially associated to one or several nodes representing the user's interest [4]. The value distribution and the stop conditions depend on the implementation objective. In our case the activation spreads only to nodes belonging to a subset of classes identified as relevant ($CPD(o)$ below) and favors the nodes similar to the initial node of interest thanks to a triple-based similarity measure ($ctriple(i,o)$ below). Concerning the architecture the algorithm is applied on a small sub-graph incrementally imported at query-time. The imported triples are stored in a local KGRAM[12] triple store using $INSERT$ queries[13] sent to a targeted SPARQL endpoint. The full algorithm formalization is published in [10].

Definition 1: semantic spreading activation

$$a(i,n+1,o) = s(i,n,o) + w(i,o) * \sum_{j \in Neighbors(i)} \frac{a(j,n,o)}{degree_j}$$

where:

- $a(i,n+1,o)$ is the activation value of node i at iteration $n+1$ for an initial stimulation at o (the starting point of the spreading activation). o being the user's query e.g. the DBpedia resource *Claude Monet*[14];
- $s(i,n,o)$ is the external stimulation value of the node i at iteration n for an initial stimulation at o;
- $Neighbors(i)$ is the set of neighbors of the node i in the linked data graph:
  $Neighbors(i) = \{x; (i,p,x) \in KB \ \vee \ (x,p,i) \in KB \wedge p \neq rdf:type \wedge x \in \cup B\}$
- $degree_j$ is the number of edges involving the node j:
  $degree_j = |(j,p,x) \in KB \cup (x,p,j) \in KB|$
- where $w(i,o)$ is a semantic-based pattern that constrains the propagation to a subset of resources' types:

$$w(i,o) = \begin{cases} 0 \ \ if \ \nexists t \in Types(i) \ ; \ t \in CPD(o) \\ 1 + |ctriple(i,o)| \ \ otherwise \end{cases}$$

where:

- t is a type of i;
- $Types(i)$ is the set of types of i;
- $CPD(o)$ is the Class Propagation Domain i.e. a subset of classes that are relevant to the query and that constrains the propagation. These classes are selected among the types of neighbors of the initially stimulated node o. The idea behind it is that informative types regarding a topic of interest must be present in its direct neighborhood. It aims to augment the quality and lower the cost of the spreading activation by applying it only to nodes belonging to this subset of types, see [10] for more information.
- $ctriple(i,o)$ (common triples) is a semantic similarity function that calculates the amount of triples having the activated node i or the origin o as objects and similar property (p) and value (v).
  $ctriple(i,o) = \{ \ (i,p,v) \in KB; \ \exists \ (o,p,v) \in KB; \ \}$
- KB is the set of all the triples (subject,predicate,object) in the triple store.

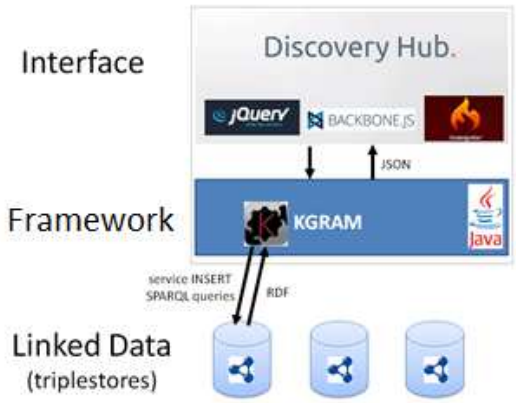

Figure 1: Discovery Hub general architecture

The objective of the spreading activation step is to identify a set of relevant results related to the initial topic of interest that will be explored by the user. They constitute an exploration perspective. The live graph importation and computation allow to adjust several important computation parameters before the processing. The changes in the algorithm configuration lead to substantial variations in the results list retrieved (composition and ranking).

DBpedia was chosen for our implementation, called Discovery Hub, because it captures a rich knowledge about many domains (arts, politics, history, etc). We also wanted to leverage the advantage that DBpedia international chapters now exist in 15 languages[15] e.g. the French, German,

[12] http://wimmics.inria.fr/corese

[13] http://www.w3.org/TR/sparql11-update/#deleteInsert

[14] *Claude Monet* is a recurrent example for Discovery Hub. It was initially chosen for a screencast presented during the Semanticpedia presentation which was strongly related to culture. *Claude Monet* was the first art-related query entered by a user in Discovery Hub

[15] http://dbpedia.org/Internationalization

Italian and Spanish ones. The Discovery Hub exploratory search engine is accessible online[16] and was showcased in several screencasts[17], see also Figure 2. A video of an exploration performed with Discovery Hub and related commentaries are available online[18]. The details of the implementation including its main parameters and the graph sampling functioning are published in [10]. Discovery Hub offers functionalities that helps exploration such as faceted browsing and results explanations features. It presents the results sorted by types (corresponding to $CPD(o)$, used as facets, to ease the results understanding and browsing e.g. artist, museums for a painter. The users have the possibility to import their Facebook *likes* in order to query them. In the context of long-lasting interests they can also gather the results of their searches in *collections*.

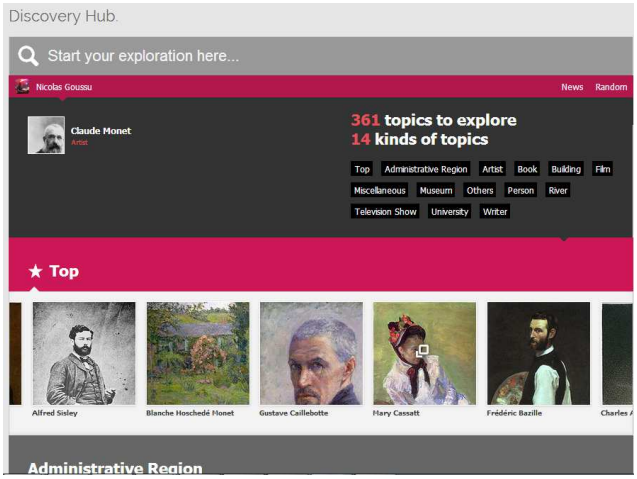

Figure 2: The Discovery Hub exploratory search engine interface

An important part of the DBpedia knowledge is captured by the hierarchy of categories. In Wikipedia the articles are classified into hierarchically organized categories that appear at the bottom of the pages. For instance Claude Monet[19] belongs to, among others: *artists from Paris, French impressionists painters, alumni of the école des Beaux-Arts*. In DBpedia the resources are related to their categories with the *dcterms:subject* property[20]. The value of using the DBpedia categories to compute recommendations was shown in [5]. To demonstrate the impact of categories in our case we reused the results of the previous experimentation detailed in [10]. During this experimentation we asked 15 participants to rate the results retrieved by our framework. The results lists were constituted of 20 films that originated from 5 queries having a film as input. These films were randomly selected in the *50 films to see before you die* list[21]: *2001 a space odyssey, Erin Brokovitch, Terminator 2, Princess Mononoke and Fight Club* (mentioned later as *query-films*). The participants rated the results on their relevance and discovery aspects.

[16]http://discoveryhub.co
[17]https://www.youtube.com/user/wearediscoveryhub
[18]https://www.youtube.com/watch?v=MUK01T-n1Ks
[19]http://en.wikipedia.org/wiki/Claude_Monet
[20]http://purl.org/dc/terms/subject
[21]http://www.listal.com/list/film-4-50-films-see

We observed the categories consideration influence on the results lists rankings. We constituted 2 *golden truth* results lists for each 5 query-films by ordering them by their relevancy and unexpectedness in decreasing order. Then we compared our framework results with ($ctriple(i,o)$ with $p = dcterms : subject$) and without ($ctriple(i,o)$ always equal to 0) categories consideration to these *golden truth* lists. The lists were compared to the golden truth ones thanks to a Kendall-Tau rank correlation coefficient [6] where 1 corresponds to a perfect rank correlation, see the results on Figure 3. It is observable that on average the ranking correlation with the golden truth is better when the categories are considered for both the relevance and the discovery for this set of movies. This goes in the sense of [5].

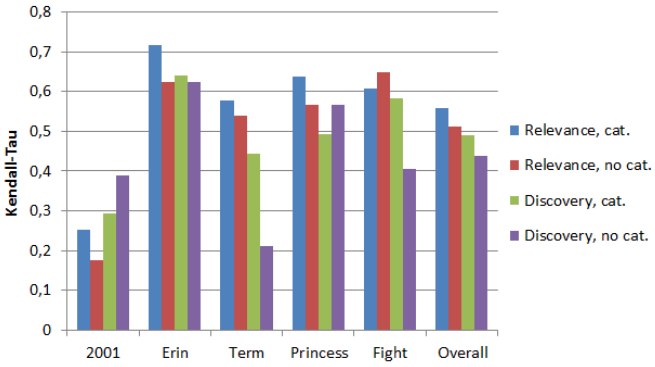

Figure 3: Kendall-Tau rank correlation coefficient between results with and without categories consideration and *golden truth* results list

3.2 Specification of Interest Criteria

Using Discovery Hub the users can specify criteria of interest and disinterest that are used by the framework during the sample importation and its computation, see Figure 4. The objective is to guide the spread and modify the activation values in order to retrieve results that are more specifically related to aspects that interest the users. The criteria specification function modifies the definition 1 as follows:

Definition 2: semantic spreading activation, criteria specification variant

$$a(i,n+1,o,V) = s(i,n,o) + w_{crits}(i,o,V) * \sum_j \frac{a(j,n,o)}{degree_j}$$

where:

$$w(i,o,V) = \begin{cases} 0 \;\; if \;\; \nexists \, t \in Types(i);\; t \;\in CPD(o) \\ 1 + |ctriple_{crits}(i,o,V)| \;\; else \end{cases}$$

where:

- $ctriple_{crits}(i,o,V) = \{\; p \in P,\; v \in V,\; (i,p,v) \;\in\; KB,\; \exists\; (o,p,v) \;\in\; KB;\; \}$
- P is the set of properties used for the triple-based similarity measure,
- V is the set of criteria of interest specified by the user.

In our implementation the DBpedia categories are used as criteria of interest/disinterest i.e. p = *http://purl.org/dc/terms/subject*. We made this choice because as shown in the

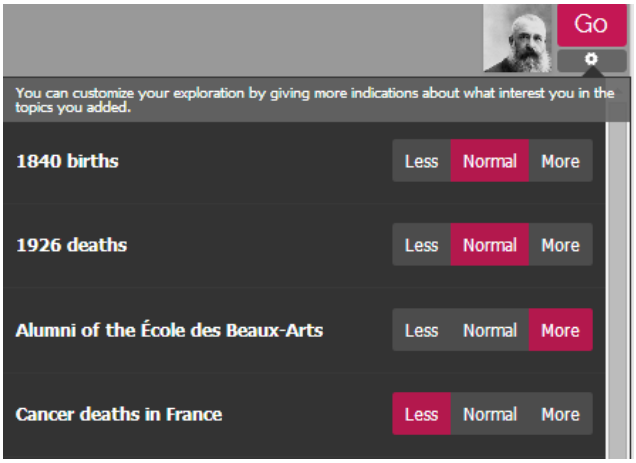

Figure 4: Criteria of interest specification through the Discovery Hub interface

previous section the categories consideration has influence on the results. With the criteria specification variant the categories are finely considered and not used in an undifferentiated manner as in the basis algorithm (definition 1). For instance by declaring such criteria the user can perform a query specifying that he is interested in Claude Monet because he is an impressionist but not because he is French. Examples of such queries and corresponding top results are presented in Table 2. The results presented are the *Artist* facet results. The DBpedia categories often reflect pieces of information associated to a class e.g. movement, origin for artists. The influence of the criterions selection is consequently easily observable on this facet (but have also influence on the others). The top 10 results lists presented in Table 2 are all related to Claude Monet but constitute different perspectives:

- The first query where no category of interest are specified retrieves artists that are strongly related to France and impressionism: 9 on 10 are French, 8 on 10 are impressionists.
- The second query where all the categories related to France were declared as uninteresting and the category impressionist painters declared of interest retrieves 9 non-French (and 1 French) impressionists painters. American artists are well represented.
- The third query where all the categories related to France were declared interesting and the category impressionist painters declared uninteresting retrieve 10 French painters where 5 are not impressionists (realist, fauvist, romantic), 4 are impressionist but not only (fauvist, cubist, modern artist) and only 1 who is only declared as impressionist and post-impressionist.

3.3 Randomness injection

It is possible to inject randomness into the propagation process in order to modify the ranking scheme and expose more unexpected results. It changes the algorithm behavior that was originally designed to retrieve the results that are the most related regarding the activation's origin i.e. the most obvious ones. This randomizing operation is particularly interesting for experts who want to retrieve unusual information in order to deepen their peripheral knowledge on a topic. To avoid to confront quickly the user with too surprising results the algorithm is different if the chosen level of randomness is inferior or superior to 0.5 (minimum 0, maximum 1). If the value is inferior or equal to 0.5 the results are randomized only at the last iteration. In other words the spreading activation occurs normally till the last iteration. If the desired serendipity level is superior to 0.5 the randomization occurs at each iteration influencing strongly the spreading activation algorithm and consequently the results list.

Definition 3: semantic spreading activation, controlled randomness variant

$$a_{rand}(i,n,o,r) =$$
$$\begin{cases} (1-r) * a(i,n,o) + r * random() \; if \; r > 0.5 \\ | \begin{cases} a(i,n,o) \;\; if \;\; n < maxPulse \\ (1-r) * a(i,n,o) + r * random() \; else \end{cases} \quad | \; else \end{cases}$$

where:

- r is the level of randomness specified by the user, comprised between 0 and 1;
- $random()$ retrieves a random value between 0 and 1;
- $maxPulse$ is the maximum number of spreading activation iterations.

3.4 Data source selection

The data source used to process the query can be easily changed with the proposed framework. It is especially interesting in the distributed context of the LOD where many datasets capturing different knowledge are accessible through their public SPARQL endpoints[22]. In the case of DBpedia it enables the use of the local DBpedia version SPARQL endpoints. Today 15 chapters are accessible online[23]. The DBpedia chapters vary significantly in what they describe and how they describe it[24]. In this publication we focus on the 5 DBpedia chapters that propose more than 100 millions triples: the English, French, German, Italian and Spanish ones.

The differences between the results when using different SPARQL endpoints are substantial. They act as a cultural prism, quantitatively studied in the next section of this paper. Continuing with the example of Claude Monet we executed the query on the 5 previously mentioned SPARQL endpoints. Regarding the Artist and Museum facets, both interesting because strongly associated to a country and culture, we observed that:

- Using the English DBpedia chapter, 4 artists and 5 museums in the tops 10 are from English-speaking countries (as officially recognized) i.e. United Kingdom and the United States. Contrary to other languages the *Art Institute of Chicago* is ranked as the first museum (instead of the *Orsay Museum*).
- Using the French DBpedia chapter 9 artists are French in the top 10 artists and 9 museums are situated in French-speaking countries (as officially recognized): 1 in Switzerland and 8 in France.

[22] http://www.w3.org/wiki/SparqlEndpoints
[23] wiki.dbpedia.org/Internationalization/Chapters?v=190k
[24] dbpedia.org/Datasets39/CrossLanguageOverlapStatistics

Table 2: Results of three queries about Claude Monet using the criteria specification

Query	Claude Monet (1)	Claude Monet (2)	Claude Monet (3)
Criteria	None	Impressionist painters + Artists from Paris - People from Le Havre - Alumni of the École des Beaux-Arts - French painters -	Impressionist painters - Artists from Paris + People from Le Havre + Alumni of the École des Beaux-Arts + French painters +
Results			
1	Pierre-Auguste Renoir	Theodore Robinson	Pierre-Auguste Renoir
2	Alfred Sisley	Édouard Manet	Gustave Courbet
3	Édouard Manet	Alfred Sisley	Edgar Degas
4	Mary Cassatt	Władysław Podkowiński	Jacques-Louis David
5	Camille Pissarro	Leslie Hunter	Jean-Baptiste-Camille Corot
6	Edgar Degas	Theodore Earl Butler	Jean-François Millet
7	Charles Angrand	Lilla Cabot Perry	Paul Cézanne
8	Gustave Courbet	Frank Weston Benson	Marc Chagall
9	Berthe Morisot	Childe Hassam	Camille Pissarro
10	J.-Baptiste-Camille Corot	Edward Willis Redfield	Édouard Manet

- The *Kunsthalle Bremen, Alte Nationalgalerie, Museum Folkwang, Wallraf-Richartz-Museum* and the *Fondation Corboud*, situated in Germany as well as the German artist *Max Liebermann* appear only in the German chapter results.
- The *Galleria nazionale d'arte moderna e contemporanea*, situated in Italy, appears only in the Italian chapter results.
- The *Botero* museum, situated in Columbia, appears only in the Spanish chapter results.

4. EVALUATIONS

One of the toughest difficulty encountered by the exploratory search community is the evaluation of the systems [7]. In fact there is no standard metric for this purpose. The high human engagement in the search process necessitates adapted evaluation protocols that both evaluate the IR and the HCI aspects of the systems. The traditional information retrieval evaluations metrics such as recall and precision are not sufficient to evaluate this human engagement. More user-centric metrics are necessary. In this article we decided to focus on the quality of the retrieved results, on their relevance and discovery potential, rather than on the interactions.

4.1 Specification of Interest Criteria

The positive impact of criteria of interest specification was briefly studied in [10] where the top 10 films results of *Fight Club* were evaluated a second time with 3 criteria of interest specified by each participant. The average relevance rating rose from 1.42 to 1.94 (with 0 corresponding to *not interesting at all* and 3 corresponding to *very interesting*). In order to confirm this positive influence on the results we built another protocol. This protocol also covers the evaluation of the randomness injection influence, discussed in the following sub-section. We wanted to confirm that the model we propose for capturing topic-related criteria of interest is relevant and modifies efficiently the results retrieved by the algorithm. For this we formulated the following hypotheses:

- **Hypothesis 1**: Users who specify their criteria (categories) of interest about a topic before launching the search, find the results of the search more relevant than users who did not specify their criteria.
- **Hypothesis 2**: Users who specify their criteria (categories) of interest about a topic do not find the results of the search less novel than users who do not specify their criteria. In other words, there is no loss of discovery power due to the specification of interest criteria.

First we randomly choose a set of 20 queries - i.e. DBpedia resources - from the query log of Discovery Hub. These queries or resources are hereafter referred to as *exploration topics*. Second we asked 16 participants to select in this list the 4 exploration topics that were the most interesting to them. We retained the 2 topics selected by the largest number of participants: information visualization and the singer Serge Gainsbourg. This selection is interesting regarding exploratory search as it is composed of a topic mainly related to a professional interest (information visualization) and an exploration topic related to a personal interest (Serge Gainsbourg). Then we asked each participant to specify the categories associated to each topic they considered either as interesting or uninteresting. 5 categories were available for information visualization and 19 for Serge Gainsbourg. Finally a list of results was generated with 4 algorithm configurations: with the basis formula (definition 1), with categories consideration (personalized per participant, definition 2), with a random level of 0.5 and 1 (definition 3). All these results were randomized in a single list. The participants evaluated them with the Discovery Hub application notably using the explanation features. An evaluator was present during the test to help them and to collect their impressions for further research. The participants were composed of 6 females and 10 males of 31 years old on average.

Our experimentation aimed to evaluate the interest and the surprise of the users regarding the perspectives they can explore. The framework notably proposes operations that constrain (criteria specification) or free (randomness injection) the spread over the data graph in order to increase the users' interest or surprise. We wanted to measure the influence of such variants on both the interest and surprise. For a precise evaluation we proposed the following definitions to the participants. A result is surprising if:

- You discovered an unknown resource or relation.
- You discovered something unexpected.

A result is interesting if:

- You think it is similar to the topic explored.
- You think you will remind or reuse it.

The users were invited to evaluate the interestingness and surprisingness of each result by indicating their degree of agreement or disagreement about the four following statements, presented in the form of a 4-point Likert scale:

- S1: This result in itself is surprising: *Not agree at all 1-2-3-4 Totally agree*;
- S2: This relation between the topic searched and the result is surprising: *Not agree at all 1-2-3-4 Totally agree*;
- S3: This result is interesting: *Not agree at all 1-2-3-4 Totally agree*;
- S4: This result is too distant from the topic searched: *Very close 1-2-3-4 Too distant.*

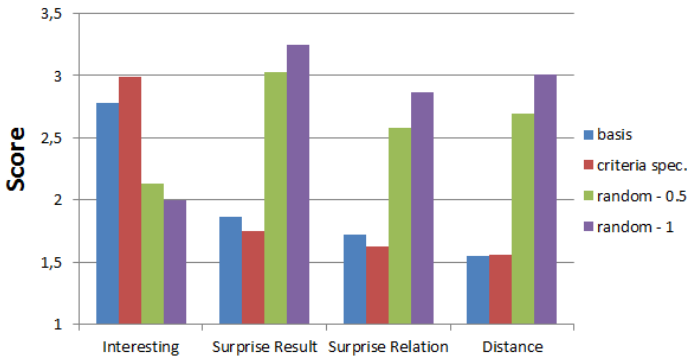

Figure 5: Interest, surprise and perceived distance of results according to 4 algorithm configurations

The first interesting observation is that the selection of categories was very diverse. Only 2 criteria selections on 16 were doubloons for information visualization and 1 for Serge Gainsbourg. It confirms that regarding a topic the users are interested in different aspects and that allowing the exploration of topics through different perspectives might be relevant. Looking at Figure 5 we observe that the results generated by the algorithm considering the users' categories are judged more interesting than the results generated by the other algorithms thus the hypothesis 1 is validated. Conversely, we observe that these results are judged a bit less surprising thus the hypothesis 2 is not validated. Otherwise the loss in term of surprise is minor and does not require in our sense a modification of the algorithm. It is certainly due to the prior knowledge of the users' about the criteria of interests they specified. Concerning the agreement the standard deviation was of 0.54 on average for all the different metrics and algorithm variants. The maximum average standard deviation was 0.68 (surprisingness of the relation, 0.5 randomized variant) and the minimum was 0.37 (perceived distance, basis formula).

4.2 Randomness injection

We formulated 2 hypotheses related to the controlled randomness injection functionality.

- **Hypothesis 3**: The stronger is the level of randomness the more surprising the results are for the users.
- **Hypothesis 4**: Even if the level of surprise is high, the majority of the top results are still relevant to the users.

Looking again at Figure 5 we observe that the results with a randomness set at 1 are judged more surprising than the ones with a randomness set at 0.5. Thus the hypothesis 3 is validated. We also observe that a majority are judged irrelevant (>2.5). Thus the hypothesis 4 is not validated. The intersection of the results evaluated as very interesting and very surprising is in favor of the lower randomness value (0.5). Indeed, their percentage reaches only 3.3% for the randomness value of 1 versus 7.5% for the 0.5 value (4.5% for the other algorithm variants). Thus we might use lower levels of randomness in the future to obtain a better trade-off between relevance and surprise.

4.3 Data source selection

The following hypothesis was formulated about the selection of the DBpedia chapter used to process the query.

- **Hypothesis 5**: Significant results variations exist when using different DBpedia chapters for the same query. In others words, the results reflect the knowledge variation present in the DBpedia chapters.

It is especially hard to evaluate the result relevance according to cultural criteria as it is profoundly subjective. Thus we decided to evaluate quantitatively the difference between the results lists obtained from different DBpedia local chapters. We filtered the whole list of distinct queries entered in Discovery Hub (2302)[25] to keep only the entities that were described in the 5 biggest DBpedia chapters: the English, French, German, Italian and Spanish ones. The amount of query-entities that was described in all this 5 chapters was 739 (32%)[26]. Then we processed these queries with the SPARQL endpoints, which expose the localized DBpedia versions. We compared the French, German, Italian and Spanish results with the English chapter ones. We chose to compare them with the English chapter results because the vast majority of existing applications uses it and only it today. The results are shown on Figure 6. It is interesting to notice that the top 100 shared results are relatively low and that a consequent proportion of results does not exist in the English DBpedia chapter. Thus the hypothesis 5 is verified. The average execution time on each chapter was few seconds (maximum 5 seconds for the English chapter and minimum 3 seconds on average for the Spanish one). It proves that the framework is adapted in the distributed context of the LOD where many knowledge sources emerge. Our configuration for tests was:

- Application server: 8 processors Intel Xeon CPU E5540 @2.53GHz 48 Go RAM
- SPARQL endpoint: 2 cores Intel Xeon CPU X7550 @2.00GHz 16Go RAM

[25]http://discoveryhub.co/querylog-DH.txt
[26]http://discoveryhub.co/querylog-DH-multilingual.txt

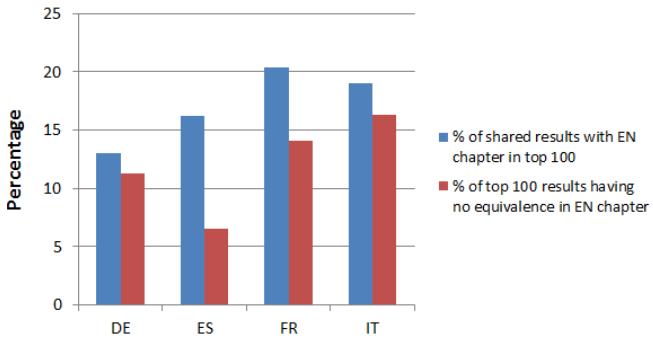

Figure 6: Percentage of shared results with top 100 English chapter results and percentage of top 100 results that are specific to the chapter

5. CONCLUSION AND PERSPECTIVES

This paper was motivated by the idea that the linked data richness allows to explore topics of interest through several perspectives. The contributions of this paper are the formalization, implementation and evaluation of such search approach over a linked data source: DBpedia. Through the interactions with such perspectives the users unveil *knowledge nuances* about the topic explored. This multi-perspectives exploratory search functionality over DBpedia was achieved thanks to a framework that computes the results at query-time. It allows the users to have the hand on important computation parameters that change the processing and consequently the results list according to specific aspects of interest. At the time of writing 3 main operations are available within the framework: the specification of criteria of interest, the randomness injection and the choice of the data source.

During our analysis we observed that the interest criteria specification with DBpedia categories leads to more interesting results according to the users. Concerning the randomness injection we observed that the level of randomness is correlated with the level of surprise. We also observed that the gain of surprise is at the cost of a consequent loss of relevance. The execution of a set of queries on the English, French, German, Italian and Spanish SPARQL endpoints and their comparison showed considerable differences between the results lists. It is also noticeable that an important part of the *non-English* results does not exist in the English DBpedia chapter. It is not a surprise as it is known that the knowledge captured by the DBpedia chapters is different but it is the first time that such variations in search results were quantified. Thus it is important to pursue the efforts of DBpedia's internationalization both in quantity and quality to avoid an over-use or misuse of the DBpedia English chapter.

The randomness injection and the choice of the knowledge source will be integrated in Discovery Hub as components of the advanced search functionality. The current implementation already integrates the criteria of interest specification. We will propose a randomness value between 0 and 0.5 in order to get a better tradeoff between relevance and surprise than the one we obtained during our experimentation.

Several linked data based functionalities are relevant to enhance the concept of multi-perspectives exploratory search. A promising one for Discovery Hub is to allow the users to tune the criteria of interest when they examine the results in order to re-rank them without launching another query. The classification of the results, leveraged by faceted browsing, could also be improved by integrating the users' specified criteria of interest. The introduction of such functionalities will also require new evaluations of the users' interface. To conclude an evaluation of the whole application and the interactions it proposes is necessary and planned.

6. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International journal on semantic web and information systems*, 2009.

[2] R. Blanco, B. B. Cambazoglu, P. Mika, and N. Torzec. Entity recommendations in web search. In *The Semantic Web–ISWC 2013*. Springer, 2013.

[3] M. Brambilla and S. Ceri. Designing exploratory search applications upon web data sources. In *Semantic Search over the Web*. 2012.

[4] P. R. Cohen and R. Kjeldsen. Information retrieval by constrained spreading activation in semantic networks. *Information processing & management*, 1987.

[5] T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker. Linked open data to support content-based recommender systems. In *International Conference on Semantic Systems*. ACM, 2012.

[6] M. G. Kendall. Rank correlation methods. 1948.

[7] B. Kules, R. Capra, M. Banta, and T. Sierra. What do exploratory searchers look at in a faceted search interface? In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2009.

[8] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, et al. Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2013.

[9] G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 2006.

[10] N. Marie, F. Gandon, M. Ribière, and F. Rodio. Discovery hub: on-the-fly linked data exploratory search. In *Proceedings of the 9th International Conference on Semantic Systems*. ACM, 2013.

[11] D. S. E. Mirizzi, Roberto and T. Di Noia. Exploratory search and recommender systems in the semantic web.

[12] A. Musetti, A. G. Nuzzolese, F. Draicchio, V. Presutti, E. Blomqvist, A. Gangemi, and P. Ciancarini. Aemoo: Exploratory search based on knowledge patterns over the semantic web. *Semantic Web Challenge*, 2012.

[13] A. Passant. Dbrec-music recommendations using dbpedia. In *The Semantic Web–ISWC 2010*, pages 209–224. Springer, 2010.

[14] T. Tran and P. Mika. A survey of semantic search approaches.

[15] J. Waitelonis and H. Sack. Augmenting video search with linked open data. In *Proc. of int. conf. on semantic systems*, volume 2009, 2009.

[16] R. W. White, B. Kules, S. M. Drucker, et al. Supporting exploratory search, introduction, special issue, communications of the acm. *Communications of the ACM*, 49(4):36–39, 2006.