

Benchmarking the Effectiveness of Associating Chains of Links for Exploratory Semantic Search

Laurens De Vocht¹, Selver Softic², Ruben Verborgh¹,
Erik Mannens¹, Martin Ebner², and Rik Van de Walle¹

¹ Ghent University, iMinds - Multimedia Lab
Gaston Crommenlaan 8 bus 201, 9050 Ghent, Belgium
{firstname.lastname}@ugent.be

² Graz University of Technology, IICM - Social Learning Group
Inffeldgasse 16c, 8010 Graz, Austria
{firstname.lastname}@tugraz.at

Abstract. Linked Data offers an entity-based infrastructure to resolve indirect relations between resources, expressed as chains of links. If we could benchmark how effective retrieving chains of links from these sources is, we can motivate why they are a reliable addition for exploratory search interfaces. A vast number of applications could reap the benefits from encouraging insights in this field. Especially all kinds of knowledge discovery tasks related for instance to ad-hoc decision support and digital assistance systems. In this paper, we explain a benchmark model for evaluating the effectiveness of associating chains of links with keyword-based queries. We illustrate the benchmark model with an example case using academic library and conference metadata where we measured precision involving targeted expert users and directed it towards search effectiveness. This kind of typical semantic search engine evaluation focusing on information retrieval metrics such as precision is typically biased towards the final result only. However, in an exploratory search scenario, the dynamics of the intermediary links that could lead to potentially relevant discoveries are not to be neglected.

1 Introduction

Resolving precise relations within chains of links is not only a matter interesting for semantic search in fact-based knowledge repositories or digital archives. The usage of such systems with linked data is getting wide spread nowadays in a variety of topical domains [12].

The semantic relations in Linked Data between a single chain of links (or nodes) define how two concepts are related to each other. For example, in DBpedia¹, we can find associations being a direct link such as **Paris** is the *capital of* **France**; but also longer chains such as **Paris** has *mayor* **Bertrand Delanoë**, which has *religion* **Catholic Church**, which is the *religion of* **Joe Biden**, the *vicepresident of* **Barack Obama**. Any single chain of links, direct or longer, preserves the value of well-assigned information fitting to a context and concepts of the underlying graph. Linked data needs

¹ <http://dbpedia.org>

precise search and exploration algorithms in order to serve the objective of qualitative informational retrieval and knowledge discovery.

Exploratory search is an serendipitous activity and represents "a shift from the analytic approach of query-document matching toward direct guidance at all stages of the information-seeking process" [22], where users can at all stages see immediate impact of their decisions. By following hyperlinks, they can better state and precise their information problem, and bring it closer to resolution. Exploratory search can describe either the context that motivates the search or the process by which the search is conducted [8]. This means that the users start from a vague but still goal-oriented defined information need and are able to refine their need upon the availability of new information to address it [3].

Led through the considerations presented, hereby we design a benchmark which addresses the main research questions addressed in this paper:

- *How to benchmark tools or search engines for Linked Data exploration?*
- *How effective can a search engine reveal initially hidden associations, as chains of links between interlinked resources, to facilitate users explore the underlying data?*

Such a benchmark is necessary because current methods used to evaluate or benchmark semantic search are biased towards evaluating how relevant the final results are presented to users. This aspect by itself is less crucial for exploratory semantic search, as it does not take into consideration the context, and the in-between steps, that motivate the final search results. Moreover regular, information retrieval, benchmark frameworks are focused on measuring precision and recall of the final retrieved values and do not capture iterative refinements during the user's search process or take into account chains of links that build up semantic associations.

Our proposed benchmark model intends to draw the focus of evaluation and capturing the quality of search on revision of the partial steps of search, on in-between results that serve as decision stations and input for further queries of search process, towards the final result achieved through exploratory actions with the search engine. It elaborates on the hypothesis that associating Linked Data where the information is retrieved based on crawling and minimal cost optimizing chains of links. It compares – and could lead to improvements of – cases of related semantic search solutions implemented for the same purpose, in particular shortest path SPARQL transitive paths. We focus on the precision of the chains of links as the main quality measure for search. The delivered results are intended to be expandable and contain underlying associations. The expansion of results happens both manually via the user-interface and automatically in back-end. The expansion stops when a reasonable extent fitting the needs of user is reached. Such specific set-up requires an extended benchmark. As a motivating example to demonstrate the application of the benchmark model we used ResXplorer² as reference. It is a visual search interface to explore publication archives, authors and events related to them, which supports step-wise extension of search focus, either manually or through back-end algorithms which detect the chain of links as useful extension of selected search term or set of terms [19].

² <http://www.resexplorer.org>

The remainder of this paper focuses on the introduction of the benchmark model and illustrating where and how it can be applied using the motivating example. First, we outline related work in Section 2. Then, we discuss where and how to apply the benchmark in Section 3. We illustrate with an example in Section 4 where we explain the features the example implements in Section 4.1, how we selected the test queries (Section 4.2) and explain the used datasets (Section 4.3). Finally, we introduce the preliminary results (Section 4.4) and we argue what we can already learn from the results of applying the benchmark to the example as well as propose next steps to affirm the initial findings and indicate how the benchmark results can be further improved and generalized in Section 5.

2 Related Work

Existing benchmarks for semantic search, SPARQL queries, and Linked Data retrieval cover only the “bottom layer”, the machine interface, of our needs for evaluation. Some components of our model use SPARQL queries, thus we considered the use of *SP2Bench* [13] benchmark and others alike, but they do not cover all aspects of search functionality we implemented. Besides barely evaluating the machine interface and measuring performance, the information and results space that existing benchmarks for semantic search cover, exploratory semantic search aims at retrieval tasks richer and more complex than the one from e.g. SP2bench.

Efforts to define benchmarks for semantic search are evolving [1,6], in terms of features, workflow coverage and/or service support discovery, although they deliver mainly single-experience recommendations so far. These experiences focus on specific data sets and measuring machine related performance. The evaluation is mostly driven based on information retrieval measures and considers the direct outputs of queries. The results of exploratory search tasks also consider the expansion dynamics and intermediary steps of these and as such they introduce additional requirements for evaluation of effectiveness. The selection of search evaluation tasks, is a tricky and crucial point in a benchmarking framework (i.e. query selection). One could ask a number of users to see what exploration queries they may do, or a systematic way could be conducted to identify the main tasks, following an existing task taxonomy or looking at the cognitive processes required for completing the tasks. This induces need to separate the conceptual exploratory operations users may carry out over semi-structured data from the particular interface designs used to give users access to such operations [10].

A system survey on Linked Data exploration systems [9] learned that massive use of linked data based exploratory search functionalities and systems constitutes an improvement for the evolving web search experience and this tendency is enhanced by the observation that users are getting more and more familiar with structured data in search through the major search engines. In a report on semantic search systems [18], the authors (i) reflected on their experiences on the evaluation; (ii) concluded that such evaluations are generally small scale due to lack of appropriate resources and test collection, agreed performance criteria and independent judgment of performance; and (iii) proposed for future evaluation work: “the development of extensible evaluation benchmarks and the use of logging parameters for evaluating individual components

of search systems”. An example is the Hermes search engine [16], which is exactly in the scope of the benchmark model we envision, implemented for advanced users in the scientific community. This is because Hermes retrieves results for very specific target audience and is very context driven. Therefore evaluation through an expert users seems as a suitable method to judge the results. Led by these findings and absence of adequate benchmarks that cover all facets of the exploratory semantic search approach, we necessitated to define complementary user-centered benchmark for the exploratory semantic search specific aspects.

3 Where and How to Apply the Benchmark Model

In order to evaluate the effectiveness of exploratory semantic search approaches focused on the resolution of chains of Linked Data, we extended an informational retrieval evaluation approach for the back-end. This section firstly explains what kind of semantic search engines we target and which benchmark methodology we used to measure the interaction between users, the semantic search engine and the interface. Secondly, it provides information about the datasets and finally it reports on the applied and executed benchmark results for the experimental setup we implemented for our use case.

3.1 Target Search Engines

The benchmark is designed (see schema in Figure 1) to work with semantic search engines supporting five distinct interfaces (or actions) that form the basis for preparing the data for exploration tasks and supporting end-users in exploring the data: *load*, *update*, *lookup*, *relate* and *query*. In this particular case we are interested in the measuring the effectiveness of the *lookup*, the translation of a keyword to a Linked Data representation, and the *relate* functionality. The *query* functionality comprises SPARQL queries for which there are existing benchmarks and therefore we consider it out of scope of this benchmark model. As the benchmark focuses on interactive exploration

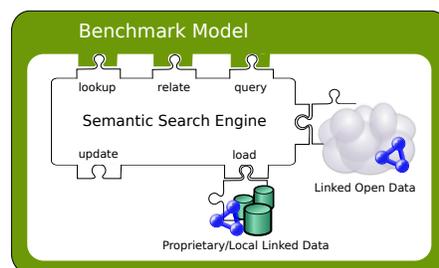


Fig. 1. Benchmark model for evaluation exploratory semantic search engines that deal with partly proprietary or local linked data which is aligned with Linked Open Data. The lookup, relate and query functionality ‘are plugged in’ for evaluation of the effectiveness.

by users, the benchmark requires input from users to define the queries required to

measure the parameters. The role of social data is to include a timely and personalized context to the search. Beyond that, social data adds additional relationships between users and resources that are contained in the more static data and potentially introduce additional references to other Linked Open Data. Most importantly including social data, will allow creating links between the users and the data they are exploring.

3.2 Specification

Parameters The benchmark consists of variable parameters for input: a set of test queries Q and an experimental setup $X = \langle O, V, I \rangle$. X contains the semantic search engine under test O , an interactive search interface V and indexes static datasets S and a dynamic dataset, for example containing links to social media, D , in a search index I , so $I = \langle S, D \rangle$.

Baseline As a baseline for the query engine O we recommend using SPARQL transitive paths or property paths, or basically any algorithm giving the shortest possible chain of links connecting two entities.

Queries A set of n queries $Q = \{q_1, \dots, q_n\}$ are identified by observing queries asked by at least N test-users in a controlled experiment, to guarantee a ‘varied’ mix consisting of distinct query patterns. Each query q_i consists of a number of keywords n_i fitted by selecting examples w_{k_i} in the query patterns the test-users were interested in: $q_i = \{w_1, \dots, w_{n_i}\}$.

Datasets It is important that both indexed datasets in the index I , the static S and the dynamic D , have sufficient links between them. If all test-users can start a personalized search, find out how several of their preferred keywords are relate to their user profile. Each test-user profile is expressed as a set of triples in S .

Measurements The main parameter under test is the engine O . The test-user interacts with the data through the interface V and the engine O is the bridging component between V and the datasets in the index I . All intermediary interfaces are optimized according to the semantic model for the selected datasets. In semantic search it is important to assess the *effectiveness* to obtain insight in how well the system performs and its individual components interact. Each of these measures indicates a different aspect of the search engine.

The effectiveness E indicates the overall perception of the results by the users taking into account expert-user feedback. This is expressed as the search precision P [11].

$$P = \frac{\# \text{ relevant objects}}{\# \text{ retrieved objects}} \quad (1)$$

In regular search and information retrieval benchmarks it is common to examine both precision and recall. In the case of exploratory search it is meaningful to measure

only precision and not recall because computing relevant results for the entire dataset is complex due to its size and dynamic nature (D). Each query q_i in Q delivers a different number of relevant results, which makes the usage of mean average precision (MAP) an important measure. The aim of this averaging technique is to summarize the effectiveness of a specific ranking algorithm over the collection of queries Q .

$$AvP(q_i) = \frac{\sum_{k=1}^{A_i} P(k) \cdot rel(k)}{\# \text{ relevant objects}} \quad (2)$$

where A_i is the number of actions taken by the user when resolving the query q_i and $P(k)$ is the precision in the result set after user action a_k in search iteration $k - 1$ via the interface V and $rel(k)$ equals to 1 if there are relevant documents after a_k and 0 otherwise. As a result, the items contained in $P(k)$ are k (where $k > 0$) steps away from the matched keyword search context items $P(0)$.

$$MAP = \frac{\sum_{q_i \in Q} AvP(q)}{|Q|} \quad (3)$$

3.3 Application

This benchmark can be applied by instantiating all generic parameters outlined in the specification: when an exploratory search engine is developed, it needs five interfaces (APIs) supporting the core functions as explained in section 3.1. The benchmark itself will make use of the implementation of the *lookup*, *relate* and *query* functions to compute the results. It remains free how the data is modeled, but RDF and SPARQL result sets are evident choices. The *update* interface can be used to feed changing data such as for example data from a user context provided as for example annotated social media. The *load* function is intended to index the static data that the user is interested in, typically domain specific and kept locally. Section 4 describes an example application where we applied this benchmark model to a use case where researchers want to explore scientific conferences and publications in online digital archives.

4 Motivating Example

When researchers search for information related to their work, they typically define search queries as a set of keywords, for example using Google Scholar or digital archives such as DBLP or PubMed. We created ResXplorer for search and exploration of the underlying Linked Data Knowledge Base as an example of a Research 2.0 (adapting Web 2.0 for scientific research) aggregated interface [17]. The way exploratory semantic search is applied to Research 2.0, is based on discovery of resources in scientific Linked Data repositories using our earlier developed EiCE (Everything is connected Engine) [2] search engine for retrieving chains of Linked Data.

4.1 Features

The front-end of ResXplorer, which we use as example semantic search engine to apply the benchmark, uses real-time keyword disambiguation to guide researchers in

Table 1. Selected query instances based on query patterns consisting of the normalized keywords by test-users. Keywords matching to a loaded social user profiles are underlined.

Query	Keywords
Q_1	LDOW, Bizer
Q_2	ISWC2012, Lyon, France,
Q_3	ISWC2008, Linked Data, Germany
Q_4	Linked Data, WWW2012
Q_5	<u>Selver Softic</u> , Semantic Web, Michael Hausenblas
Q_6	<u>Selver Softic</u> , Linked Data, Information Retrieval
Q_7	<u>Laurens De Vocht</u> , <u>Selver Softic</u>
Q_8	<u>Laurens De Vocht</u> , <u>Selver Softic</u> , 2011
Q_9	<u>Laurens De Vocht</u> , Linked Data, WWW2013
Q_{10}	Chris Bizer, WWW2013, ISWC2010

expressing an accurate query context, thereby implementing a *lookup* function. The back-end retrieves the researcher’s resources and ranks them. At the same time, it fetches neighbour links which match the selected query context as well. As a result, the selection of various resources is presented to the researchers. It resolves queries consisting of one or more research concepts by being able to resolve them with refined entities out of Linked Data sets. Furthermore it takes into account all relevant contributions from researchers, user context, to improve the ranking of found resources related to a search context, the *relate* function.

4.2 Queries

We collected for the evaluation typical keyword query patterns that were asked against the system by the target group of the use case ($N = 36$ users) [5], both researchers and innovation policy makers – all in the field of information and communication science. For the evaluation, we restricted our tests to 10 instances based on the query-patterns which are answerable by the data sets we indexed. These are shown in Table 1. These queries representatively cover some of the commonly used search terms within a researcher context: Search for an event ($Q_{1,2,3,4,9,10}$), a person, author or group of authors ($Q_{1,5,6,7,8,9,10}$) or scientific resources ($Q_{1,2,3,6,9,10}$). Each search runs through the following scenario: users enter the first keyword and select the matching result that is resolving their search focus at least one step forward. The users view selected results and can expand them at any time except when the research selects the suggestions from a typeahead interface. Parallel with this selecting and narrowing down the scope, our engine finds relations between the resources and reflects the context. Additionally neighbours which match the selection are found.

In case users logged in via their Twitter and/or Mendeley accounts, their research profiles personalize the boundaries of the search space. This is done by influencing the heuristic that is used to determine the selection order of potentially relevant nodes. As soon as a user context is present, it extends the search context. In particular the strength

Table 2. Linked Data used within the search benchmark.

Dataset	# triples	# instances	# literals
DBpedia	332,089,989	27,127,750	161,710,008
DBLP (L3S)	95,263,081	13,173,372	17,564,126
COLINDA	143,535	15,788	70,334
Social LD	41,438	7,344	15,350

of the associations of the entities representing keywords and the entities in the user context are taken into account. The effect is that other resources might be shown to the user than in the case when there is no user context present.

4.3 Datasets

The datasets used in our experiment align some existing Linked Open Data sets, namely:

- DBpedia
- DBLP³
- GeoNames⁴
- Conference Linked Data (COLINDA)⁵ [14]
- a social linked data set containing:
 - Information about attended or followed conferences
 - Social profiles of the researchers from Twitter and Mendeley
 - User generated data

More specifically the user generated data contains a snapshot in time from all their tweets and library reading list of their Mendeley profile, the users choose when they when to synchronize their data. For the evaluation we use COLINDA to resolve the connections between GeoNames, DBpedia and DBLP since it has links to these three Linked Datasets. Further it serves as a conference entity resolver for social data used with the profiles of users from Twitter and Mendeley. Table 2 highlights statistics on the used datasets.

The total time for building all indices for all the data sources is about 6 hours (throughout the benchmark, we use a 8-core single-machine server with 16GB RAM running Ubuntu 12.04 LTS). The properties *type* and *label* are indexed separately, because they are required for each Linked Data entity described in RDF⁶ and allow retrieving entities by label and disambiguating them by type. The indices contain a special type of field *ntriple* that makes use of the SIREn Lucene/Solr plugin that allows executing star-shaped queries on the resulting Linked Data [4]. Star shaped queries are essential to immediately find neighbouring entities for each entity and to ultimately find

³ <http://dblp.uni-trier.de>

⁴ <http://geonames.org>

⁵ <http://colinda.org>

⁶ <http://www.w3.org/2009/12/rdf-ws/papers/ws17>

paths between non adjacent nodes. This type of function is essential because it supports two essential aspects for the user interface V that allow interacting with chains of links: namely expanding nodes and support finding longer chains of paths between nodes to reveal hidden associations in longer chains of links for exploration and discovery. The better the search index I supports this, the faster and the more (precise) results in potentially generates with each search iteration.

To ensure maximal scalability and optimally use available resources, we primarily use simple, but effective measures based on topical and structural features of the entities in the search engine. Relations are only computed between pairs in a subgraph of the larger dataset. Every resulting relation as a path between entities are examined for ranking. Only entities belonging to a specific search context are requested. Since the result set of entities might be very large, this “targeted” exploration of relations is essential for the efficiency and scalability.

4.4 Preliminary Results

To evaluate the benchmark we illustrated with an example benchmarking the effectiveness of a semantic engine for academic library and conference metadata In this case we tested the retrieval quality of our system with the test queries shown in Table 1. To assess the effectiveness of chains of links associations, we measured the precision and the mean average precision over all queries to evaluate that the search algorithm used in our search engine returns enough high quality relevant results for researchers to achieve their research goals effectively.

The queries were run one term at a time, to simulate exploration. After each iteration the precision was evaluated for all the results that the user would see. For visualization purposes there is a cut-off at the first then results, both when looking up a keyword as finding neighbours or a chains of links for further related resources.

Proposed Engine Effectiveness. To assess the proposed engine we measured the precision of the results for the queries in Table 1. To determine the relevance of each resource we relied on expert judgment and we verified the engine’s output against expected results. We defined what the expected outcome scenario was based on familiarizing with each of the visualized keyword searches and than having an expert compare the output of the system against the predefined scenario by checking each visualized item one by one after each expansion. Additionally the available personalized data generated a user profile which we used to project the expected search results. This extension is specifically important in the case of the queries with *Selver Softic* and *Laurens De Vocht*, where we loaded the profiles of these users as the user context.

The results in Figure 2, where **P** denote precision and **AvP** average precision per query. The mean average precision, **MAP**, over all queries is **0.60**. We judged very precisely each result to enable a more accurate evaluation of the context driven aspect of our search approach. Personalized queries Q_5 - Q_9 have been especially strictly evaluated. Each found resource which has no direct relation to the persons, events or topics specified by a keyword, even if it is relevant in a wider context, are a non-relevant result (for instance a co-author that corresponds to the person but does not fit to the specified event).

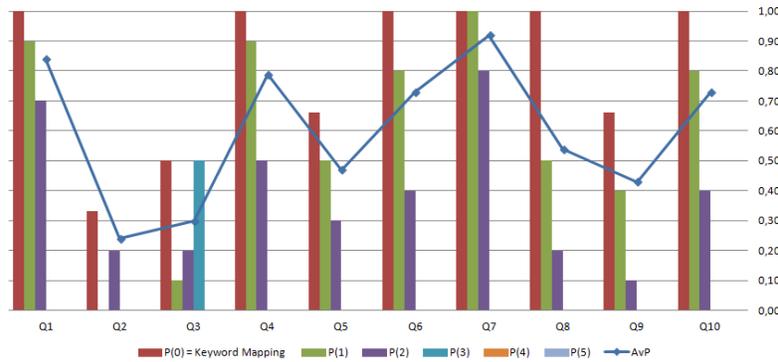


Fig. 2. Precision of the proposed engine for the test queries over different path lengths

The line chart in Figure 2 shows the average precision over queries. With exception of Q_1 , Q_4 and Q_{10} , queries with preloaded profile data (Q_5 - Q_9) deliver more precise results than anonymous queries. This difference is because the main focus of queries Q_5 - Q_9 is a person which resolves initially very good within key mapping step, thus following results keep the average precision high. Queries Q_1 , Q_4 and Q_{10} have very high precision since they have broader focus which includes more relevant results. Also the choice of keywords matches with the linked dataset instances within COLINDA and DBLP. Mean average precision overall, as expected, reaches the score of 0.60 which is high but not surprising since the resources within the linked datasets are well-connected and all used datasets properly interlinked.

The bar charts in Figure 2 indicate the precision respectively queries by distinguishing the path lengths. As expected the precision decreases with the length of paths. As the path finding progresses over extended links relation to the core concept is becoming weaker. Applying benchmark approach however was able to reveal us that the first step of keyword search as well the path finding results at distance one (path length = 1) deliver always the results that exceed or are around the value of mean average precision. With resolving each further step (search path length > 1) we get insights on progression of precision which indicates us implicitly how well the links are associated in longer chains of links. We can also observe how the algorithm decides whether to extend or hold the search for a given context. At least in this example, mean average precision (MAP) seems to be a good estimator for precisions of searches with path length = 1. This observation does not applies for the baseline evaluation.

Baseline Effectiveness. Based upon the recommendations and insights in the first run we reevaluated our system with specific focus on a comparison to a valid state-of-the-art technology baseline aiming at confirmation of our achieved results on retrieval.

Virtuoso is one of the most common triple stores. It has support for standard SPARQL transitive paths and has its own built-in index for text search (via the *bif:contains* property). In many projects dealing with the same amount of data(sets) as we did, it would be the de-facto choice. Therefore we consider it as a baseline for our solution.

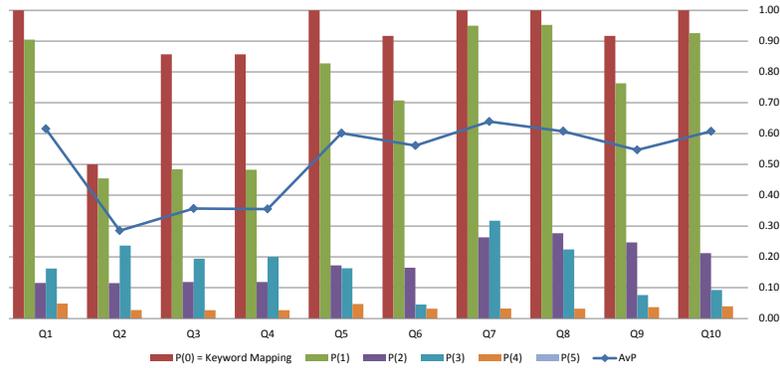


Fig. 3. Precision results of the baseline for the test queries

For the benchmarks we used version *6.1.3127*. We compare this way, executing the same ‘underlying’ queries and the keyword queries. The quality of exploratory search depends on quality and diversity of delivered top results and their connectedness to other relevant links. This is why the search is usually canceled so far new or more deeper aspects are not revealed with new steps along the link chain. The direction of the search does not aim on one single result, rather on the context of search item and the chosen transition along the chain of links in a path. Since Virtuoso supports transitive paths resolution and indexing of them for search it offers a solid comparison baseline to the ‘Everything is Connected’ engine behind the ResXplorer, our reference system, which searches the results along the chain of links or along link expansion actions[2]. The mean average precision, **MAP** for the baseline is **0.52**.

The results delivered by baseline approach confirmed our assumption about very solid retrieval responsiveness. The **MAP** we measured in first run is 8% higher than in baseline case. This first impression strengthens our first evaluation and brings us closer to the confirmation of the aforementioned hypothesis. However to explain the deviation between the results an additional detailed comparative analysis has to be done. As Figure 3 shows we detect a dip in the precision for Q_2 , Q_3 , Q_8 and Q_9 that is also found in the case of our proposed engine.

Summary. The benchmark, developed for testing implementations of exploratory semantic search approaches, was able to indicate in this example case that the proposed engine pinpoints resources for researchers effectively. The processing of queries and mapping of keywords occurs with an average precision above 50%. This is promising because it is at least equally effective as the baseline reference and given the dynamic nature of exploratory search that does not focus purely on the final presented result but also intermediary results in the chain.

5 Discussion

We delivered reproducible scripts to generate results for the queries and tested it by computing the precision measures for 10 test queries for an example semantic search engine and analyzed the results. The results in the end are by themselves not enough to generally indicate that associating links of resources facilitates exploratory semantic search but we reached the limits of what traditional information retrieval metrics can show. This leaves some open questions considering the validation of the benchmark, which we discuss here in this section and classify them under limitations and improvements of the benchmark.

5.1 Limitations

The results the benchmark produces for a proposed engine only indicate how well it performs when it is applied to an example use case. However, the goal is to take advantage of such use cases and use the benchmark as a leverage to compare different approaches. The way the evaluation was executed did not maximally demonstrate the aspects that make an exploratory semantic search approach, comparable to traditional semantic search systems. The results remain distinct but already do contribute to the field of exploratory semantic search.

Furthermore, the presented results on effectiveness are not immediately interpretable without the user context for this purpose, and without proper explanation, they even can not be properly generalized. While in fact evaluating an exploratory search system is not substantially different from evaluating any other interactive search system [21], the benchmark model as it is now, contributes to obtaining a full picture beyond the raw search effectiveness. Relevant subjective measures such as user satisfaction [20], information novelty [15], or task outcomes [7,22] would further enhance this.

The motivating example results showing a mean average precision figure (in this case around 60%) is an indicator in respect to the initial assumption that search over well contextualized repositories may compete with regular search approaches, but without any fundamental baseline approach this is a first step, but not the complete answer with sufficient evidence to support the hypothesis. This means that additional measures are required to indicate how generically applicable the approach is in the example, which exposes hidden associations as chains of links during exploratory search.

5.2 Improvements

Additional measures that will improve the benchmark results can be obtained by increasing the number of expert-user reviews and by computing the inter-rater agreement for both the exploratory semantic search engine and the baseline. Additional expert-user reviews will put the results in perspective by indicating the nuances among the different expert user ratings accurately defined. In particular we will use them to explain why there are inconsistencies in the results or why expert users disagree in some cases about an engine's effectiveness.

Some methodological improvements will facilitate to generalize the preliminary indications after the presented example results in this paper. Firstly, analyze and compare

the results of the baseline against a proposed semantic search engine in detail; and secondly, explain and verify that the approach is generic and can be applied to other search contexts with different data and use cases.

6 Conclusions and Future Work

We introduced a user-centered benchmark model to measure the effectiveness of associating chains of links because of absence of adequate benchmarks and recommendations in earlier research. The initial results we got for an illustrative case with academic library and conference metadata showed some indicative figures for a proposed search engine gave but did not provide sufficient argumentation because the lack of a baseline reference for perspective and missing nuance about the expert user ratings.

After adding the baseline which makes use of SPARQL transitive paths, reapplying the benchmark showed two things: (i) which cases the proposed engine outperforms the baseline; and (ii) when that the proposed engine is relatively more precise for query contexts that are accurately defined, i.e. consist of keywords in which the meaning is unambiguous, for example when a specific conference, author or publication are combined in a search. On the other hand when there are inconsistencies or vague terms, such as topics or years, even mismatches in the query context, expert users disagreed about this effectiveness.

In future work we want to explore the bias by adding increasing the number of reviewers: measure their level of agreement and compare the baseline more in detail versus the proposed search engine, ‘engine under test’. Furthermore we want to refine and verify the generality of the benchmark by testing it with additional users and apply it to different data and more use cases such as for example drug discovery.

Acknowledgments

The research activities that have been described in this paper were funded by Ghent University, iMinds (Interdisciplinary institute for Technology) a research institute founded by the Flemish Government, Graz University of Technology, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

References

1. Cabral, L., Toma, I.: Evaluating semantic web service tools using the seals platform. In: International Workshop on Evaluation of Semantic Technologies (IWEST 2010) at ISWC 2010 (2010)
2. De Vocht, L., Coppens, S., Verborgh, R., Vander Sande, M., Mannens, E., Van de Walle, R.: Discovering meaningful connections between resources in the web of data. In: Proceedings of the 6th Workshop on Linked Data on the Web (2013)
3. De Vocht, L., Dimou, A., Breuer, J., Van Compernelle, M., Verborgh, R., Mannens, E., Mechant, P., Van de Walle, R.: A visual exploration workflow as enabler for the exploitation of linked open data. In: Proceedings of the 3rd International Workshop on Intelligent Exploration of Semantic Data (2014)
4. Delbru, R., Campinas, S., Tummarello, G.: Searching web data: An entity retrieval and high-performance indexing model. *Web Semantics: Science, Services and Agents on the World Wide Web* 10, 33–58 (2012)
5. Dimou, A., De Vocht, L., Van Compernelle, M., Mannens, E., Mechant, P., Van de Walle, R.: A visual workflow to explore the web of data for scholars. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion. pp. 1171–1176. *WWW Companion '14, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland* (2014)
6. Elbedweihy, K., Wrigley, S.N., Ciravegna, F., Reinhard, D., Bernstein, A.: Evaluating semantic search systems to identify future directions of research. In: García-Castro, R., Lyndon, N., Wrigley, S.N. (eds.) *Second International Workshop on Evaluation of Semantic Technologies*. pp. 25–36. No. 843 in *CEUR Workshop Proceedings* (May 2012)
7. Kraaij, W., Post, W.: Task based evaluation of exploratory search systems. In: *SIGIR 2006 workshop, Evaluating Exploratory Search Systems* (2006)
8. Marchionini, G.: Exploratory search: from finding to understanding. *Commun. ACM* 49(4), 41–46 (Apr 2006)
9. Marie, N., Gandon, F.L.: Survey of linked data based exploration systems. In: Proceedings of the 3rd International Workshop on Intelligent Exploration of Semantic Data (IESD 2014) co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 20, 2014. (2014)
10. Nunes, T., Schwabe, D.: Exploration of semi-structured data sources. In: Proceedings of the 3rd International Workshop on Intelligent Exploration of Semantic Data (IESD 2014) co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 20, 2014. (2014)
11. Powers, D.M.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies* 2(1), 37–63 (2011)
12. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: *The Semantic Web–ISWC 2014*, pp. 245–260. Springer (2014)
13. Schmidt, M., Hornung, T., Lausen, G., Pinkel, C.: SP2Bench: A SPARQL performance benchmark. *CoRR* abs/0806.4627 (2008)
14. Softic, S., De Vocht, L., Mannens, E., Ebner, M., Van de Walle, R.: COLINDA: modeling, representing and using scientific events in the web of data. In: Proceedings of the 4th International Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2015) Co-located with the 12th Extended Semantic Web Conference (ESWC 2015), Protoroz, Slovenia, May 31, 2015. pp. 12–23 (2015)
15. Softic, S., De Vocht, L., Mannens, E., Van de Walle, R., Ebner, M.: Finding and exploring commonalities between researchers using the resexplorer. In: *Learning and Collaboration Technologies. Technology-Rich Environments for Learning and Collaboration*, pp. 486–494. Springer International Publishing (2014)

16. Tran, T., Wang, H., Haase, P.: Hermes: Dataweb search on a pay-as-you-go integration infrastructure. *Web Semantics: Science, Services and Agents on the World Wide Web* 7(3) (2009)
17. Ullmann, T.D., Wild, F., Scott, P., Duval, E., Vandeputte, B., Parra Chico, G.A., Reinhardt, W., Heinze, N., Kraker, P., Fessl, A., Lindstaedt, S., Nagel, T., Gillet, D.: Components of a research 2.0 infrastructure. In: *Lecture Notes in Computer Science*, pp. 590–595. Springer (2010)
18. Uren, V., Sabou, M., Motta, E., Fernandez, M., Lopez, V., Lei, Y.: Reflections on five years of evaluating semantic search systems. *International Journal of Metadata, Semantics and Ontologies (IJMSO)* 5(2), 87–98 (2010)
19. Vocht, L.D., Mannens, E., de Walle, R.V., Softic, S., Ebner, M.: A search interface for researchers to explore affinities in a linked data knowledge base. In: *Proceedings of the ISWC 2013 Posters & Demonstrations Track, Sydney, Australia, October 23, 2013*. pp. 21–24 (2013)
20. White, R.W., Marchionini, G.: Examining the effectiveness of real-time query expansion. *Information Processing & Management* 43(3), 685–704 (2007)
21. White, R.W., Marchionini, G., Muresan, G.: Evaluating exploratory search systems: Introduction to special topic issue of information processing and management. *Information Processing & Management* 44(2), 433–436 (2008)
22. White, R.W., Muresan, G., Marchionini, G.: Evaluating exploratory search systems. *EESS* 2006 p. 1 (2006)