Linked Data and Entity Search: A Brief History and Some Ways Ahead

Fahad Alahmari[1] **Liam Magee**[2]

[1] Collage of Computer Science
King Kahlid University
Abha, Saudi Arabia
Email: fahad@kku.edu.sa
[2] Institute for Culture and Society
University of Western Sydney
Sydney, Australia
Email: l.magee@uws.edu.au, URL: http://www.uws.edu.au/ics/people/researchers/liam_magee

Abstract

This paper presents a brief survey of the current efforts in the Semantic Web literature with respect to the three research topics of Linked Data, semantic search and search interfaces. First we give a basic overview about Linked Data, discussing the concepts, principles and examples of relevant initiatives such as Linked Open Data (LOD). Second we review current and relevant work on semantic search, with a particular focus on entity search from different perspectives: entity type identification, entity query suggestions, and entity attributes ranking. Finally we highlight some approaches to semantic search interfaces from two perspectives: the design of search interfaces, and the evaluation of search interfaces. Our survey offers new ways of understanding the evolving connections between Linked Data, semantic search and search interfaces.
Keywords: Semantic Search, Linked Data, Web of Data, Entity Search, User Interface

1 Introduction

Guided by the vision of Linked Data, much useful information about individual entities such as people, places or organisations has already been published and connected together over the web [25, 26]. Further information continues to be published and added; however such data is not always linked explicitly to other datasets. In response, semantic browsers have been developed to support the exploration and finding of data and relationships across diverse Linked Data sources. Increasingly, in many cases users are also searching for information about specific entities. A recent study by Pound et al. [68] shows that more than 50% of Web search queries target a specific entity or entity type. This type of Web search, also termed *entity search* [10, 11], has been investigated in a number of recent studies, some of which we review in this paper.

The recent advent of Linked Data offers powerful possibilities to improve entity search. However, there are two potential challenges associated with using entity search. The first challenge is the quality of integrated entity search results. This is due to the problems of query ambiguity and redundant attributes when searching for an entity. The second challenge concerns the limitations in providing users with engaging experiences when navigating entities over Linked Data. This is part due to the complexity of Semantic Web technologies and data structures, as well as the current immaturity of state-of-the-art tools for navigating this complexity. Considerable further work needs to be doen to build intuitive and enjoyable interface designs for browsing Linked Data.

In this paper, we provide a general exploration of the literature that discuss these challenges, specifically relating to areas of entity search and user interface design. The aim is to summarise the latest research results, and to suggest some ways forward.

2 Overview to Linked Data

The term *Linked Data* was first introduced by Berners-Lee [17, 20]. Linked Data challenges and technical principles were then discussed by Bizer et al. [26], presenting a research agenda for linking various data sources. Efforts by several others, including Heath and Bizer [51], Bizer et al. [25], Wood [91] and Wood [92], have focused on and contributed to the emerging vision of Linked Data. In addition these developments have helped established “Linked Data” as a set of established practices for publishing, linking and consuming structured data on the Web.

Broadly, this initiative makes the World Wide Web useful for sharing and interlinking *data*. Just as the value of *documents* increases dramatically when they are linked to other documents, Linked Data enhances the value of *data* by providing standardised mechanisms for describing and linking them to other datasets. Specifically, it enables developers to build Web applications that manipulate and combine data from multiple sources dynamically and simultaneously. This makes possible a network effect, where data becomes more useful the more it is related to other datasets. In short the adoption of Linked Data makes possible the creation of a universal data space for various domains such as people, music, books and organisations. Though it borrows extensively from Semantic Web standards, critically it aims ot lower the barrier for providers to publish data, and for end users to consumer data, through augmented publishing and browsing interfaces.

From a technical point of view, Linked Data data sets consist of typed links between related data from different resources. This can be achieved by adhering to the following two tenets:

1. use the RDF data model to publish structured data on the Web; and

2. use RDF links for connecting data from various data sources.

Applying these tenets enables users to share structured data on the Semantic Web as easily as they can share on the traditional Web. Linked Data uses RDF to represent resources in a triple form: subject (entity), predicate (property or attributes) and object (value or another entitiy). Each element of an RDF triple can be a URI. Table 1 illustrates an example of a RDF triple.

On the Web of Data, resources are identified using Uniform Resource Identifiers (URIs). These are used to specify a globally unique name for a resource. Resources that are identified using URIs can be further distinguished either as resources that describe real-world objects, such as people, places and cars; or as resources that describe other resources, such as documents, images and video. Identifying resources with URIs provides a simple way to describe the semantics of these resources. Consistent use of RDF and URIs produces a powerful network of machine-processable information, making way for the next generation of the Web [17]. Distributed on a global scale, Linked Data in turn can be used by machines to generate new information and knowledge.

Although terms such as Semantic Web, Linked Data and the Web of Data are often used interchangeably, it is necessary here to clarify what is meant by these terms.

The vision of the Semantic Web was expressed by Berners-Lee et al. [18] as a "Web of Data that can be processed directly and indirectly by machines". The Semantic Web extends the capabilities of documents to data published on the web. In more ambitious terms, the Semantic Web aims to create a Web of Data, where data is discoverable, accessible and processable by agents utilising the general web architecture, and the particular forms of data representation specified by the OWL and RDF standards. As with Linked Data, URIs are used to establish unique addresses for resources, and semantic metadata, in the form of ontologies, enable machines to understand and reason over the content [21]. Applying this vision leads to the creation of a common framework that allows data to be shared and reused.

When defining the relationship between the Semantic Web and Linked Data, it is evident that Linked Data focusses more on the specific mechanisms of publishing and consuming data, of connecting items in different datasets, and of retrieving resources. Indeed, this relationship has been defined on the Linked Data community's website[1], which explains it as follows:

> Opinions on this topic do differ somewhat, however a widely held view is that the Semantic Web is made up of Linked Data; i.e. the Semantic Web is the whole, while Linked Data is the parts"

In another account, Linked Data provides the road map to the Semantic Web. As described by Bizer et al. [26]:

> "while the Semantic Web, or Web of Data, is the goal or the end result of this process, Linked Data provides the means to reach that goal"

[1] http://linkeddata.org/faq

Linked Data therefore is imagined as a more pragmatic and practical form of the Semantic Web: the *means* or the *parts* of the more ambitious project that is the Semantic Web.

Linked Data is also often referred to as the *Web of Linked Data* or just *Web of Data*. This description emphasises the close association and shared common practices with the traditional *Web of Documents*, such as browsing and navigation interactions. Researchers have also used the more generic term of *Web of Data* to include data that are open and accessible, but are not yet linked. In spite of their commonalities—both the Web of Documents and the Web of Data allow users to search for information, for example—there are important technical distinctions between them. One crucial distinction is that interlinked hypertext documents can be accessed via HTTP, and an HTML-compliant web browser can be used to view the contents and navigate between documents via hyperlinks. In the Web of Data, the links are instead *typed*, to describe different relationships between datasets. This implies, and arguably necessitates, a range of different mechanisms for discovering, navigating and browsing linked data—unlike the Web of Documents, where a single application category, the web browser, suffices for most browsing purposes. As discussed below, it is not yet clear whether an anologous generic and generalisable Linked Data browser, covering all use cases, is possible or desirable. Table 2 shows a simple comparison of Web of Documents with Web of Data.

2.1 The Linked Data Rules

In addition Berners-Lee [17] has articulated several key rules for publishing data on the Web in order to establish it as Linked Data. Later known as the "Linked Data principles", these rules are:

1. Use of Uniform Resource Identifiers (URIs) as names for things.

2. Use of hypertext transfer protocol (HTTP) for URIs, as the key protocol for agents to resolve URIs.

3. Use of RDF and SPARQL [2], for representing and querying Semantic Web data respectively, to simplify how data can be retrieved and processed.

4. Links to other URIs must be embedded within RDF datasets, in order to allow for both human and computer clients to discover additional information.

These principles ensure data becomes part of the Linked Data, and is available for third-party tools to discover, connect and process data from heterogenous sources. This in turn facilitates the development of smarter, more flexible and more intuitive user applications for working with Linked Data.

Linked Data also includes the idea of levels of compliance. Clearly, data published on the Web can be expressed in an enormous range of formats and schemas. To distinguish levels of data compliance with Linked Data principles, Berners-Lee developed a 5-star rate scheme, ranging from zero (poor compliance) through to five (high compliance). The scheme is designed to encourage data owners in various domain areas such as government, healthcare and multimedia, to publish their data sets under Linked Data-friendly terms. According to Heath and Bizer [51], the 5-star system is as follows:

[2] http://www.w3.org/TR/rdf-sparql-query

	Machine-readable	Human-readable
Entity	http://dbpedia.org/resource/Melbourne	Melbourne City
Property	http://dbpedia.org/ontology/populationTotal	has a population of
Value	4169103	4.1 million

Table 1: Example of an RDF triple with machine and human representations

Features	Web of Documents	Web of Data
Basic unit	document	data
Consumers	human	human and machines
Links	Un-typed hyperlinks	RDF typed links
Implementation	HTML, URL	HTTP, RDF, URIs and SPARQL

Table 2: Basic comparison of Web of Documents with Web of Data

1 Star: "data is available on the Web (whatever format), but with an open license"

2 Stars: "data is available as machine-readable structured data (e.g., Microsoft Excel instead of a scanned image of a table)"

3 Stars: "data is available as (2) but in a non-proprietary format (e.g., CSV instead of Excel)"

4 Stars: "data is available according to all the above, plus the use of open standards from the W3C (RDF and SPARQL) to identify things, so that people can link to it"

5 Stars: "data is available according to all the above, plus outgoing links to other people's data to provide context"

2.2 The Linking Open Data Cloud

The LOD cloud contains data from a range of different domains, with high representation from the domains of media, government, the life sciences and geography [26]. Significantly, these datasets are linked together using terms expressed in W3C base vocabularies such as RDFS and OWL, in order to make Linked Data machine-processable.

The following briefly details the recent history of this project:

- **In 2006**: Berners-Lee [17] introduced the concept of Linked Data as a more practical and realisable form of the Semantic Web vision.
- **In 2007**: the project announced a limited number of linked datasets, 12, in May 2007; the number had increased to 28 datasets in November of the same year.
- **In 2008**: the number of datasets increased to 45.
- **In 2009**: the number of datasets increased to 95. In this year also, Berners-Lee gave his talk on "The next Web of Open, Linked Data" on TED [20], in which he defined what Linked Data is, and discussed the benefits of exposing 'raw' data in the Web.
- **In 2010**: there were 203 datasets including 26 billion RDF triples interlinked by 395 billion links[3].
- **In 2011**: the project had grown to 295 datasets, which included nearly 30 billion triples interlinked by 471 million links[4].
- **In 2012**: there were more than 300 datasets according to LOD stats [5]. However, no attempt was made to visualise the LOD cloud in this year, which may be indicative of the large number of datasets.
- **In 2014**: there are 558 data sources according to the latest statistics [72].

DBpedia - Core Datasets for LOD. DBpedia is a community project that aims to automatically generate information from Wikipedia and make it available [4]. Unsurprisingly it has become one of the major Linked Data knowledge bases on the Web. Hence, Cyganiak and Jentzsch put it in the centre dataset of their diagram of LOD[5]. It includes a large coverage of resources that describe real entities in the world such as people, places, events, activities, and movies.

DBpedia extracts information from Wikipedia infoboxes and normalises them to a set of ontology types and properties [27]. For example: if a Wikipedia page has the property *Population*, then the DBpedia version becomes *dbpedia-owl:populationTotal*. So all the infobox properties are mapped to OWL properties at a high level of normalisation. Wikipedia also provides values for each *infobox* property; DBpedia takes these into consideration by assigning them to the matched data formats: numeric (e.g.: integer), metrics(e.g.: length), temporal (e.g.: time) or plain text.

Table 3 shows several DBpedia statistics and the changes between the 3.8 and 3.9 English versions. Updated continuously, clearly DBpedia is a massive collection of useful information. Its very success has however has brought to light some issues regarding this collection. One critical issue is the prevalence of noisy and redundant information, which must be decreased in order to improve the quality of searching information. This motivates in part our discussion of entity search and attribute ranking below.

3 Overview to Semantic Search

With the rapid growth of LOD, an open question remains as to how best to produce semantic search results from different data sources, in order to ensure that typical Web users can easily query and search this wide range of semantic data without redundant or irrelevant information. This is a challenging task in

[3] http://wifo5-03.informatik.uni-mannheim.de/lodcloud/state/2010-10_index.html

[4] http://www4.wiwiss.fu-berlin.de/lodcloud/state/

[5] http://lod-cloud.net/

		v3.8	**v3.9**
Entities		3.77 million	4.0 million
Properties		1772	2333
Types		359	529
Persons		764k	832k
Places		573k	639k
	Populated places	387k	427k
Work		333k	372k
	Music	112k	116k
	Movies	72k	78k
	Video games	18k	18500
Organisations		192k	209k
	companies	45k	49k
	educational institutions	42k	45k
Species		202k	226k
Diseases		5500	5600

Table 3: DBpedia statistics for 3.8 and 3.9 English versions

both the Information Retrieval (IR) and the Semantic Web communities [7, 85]. The objective is to provide a search environment that captures the meaning beyond users' search intentions and the contents on the Web. This search process is referred to as *semantic search*.

Semantic search is a process that exploits semantic techniques such as ontology matching, information extraction, and inference and reasoning to improve search results. These semantic techniques can be merged with some existing techniques in IR, such as keyword querying, crawling, and indexing, to enable a higher level of semantic resolution to queries [30]. That in turn can be used to improve different tasks in semantic search in order to help end users to satisfy information needs.

Generally, three different forms of search can be used to access semantic data over Linked Data:

1. structured queries, where users can use the SPARQL query language for manipulating data sources;
2. keyword-based queries, where users can input an entity query as a free text keyword search; and
3. exploratory browsing, where users can browse from one entity to another.

In recent years, there has been an increasing amount of literature on each of these forms. In what follows, we summarise the main studies for each category.

3.1 Structured Queries

SPARQL is a query language for RDF data. It exploits the powerful graph-based nature of RDF, and has appropriate operators (unions, joins, selections and projections) for creating queries to extract RDF information from one or moreWeb sources. Listing 1 is an example of a SPARQL query over DBpedia.

With the rise of SPARQL, there are several approaches that address its efficiency, and more general query processing capabilities over Linked Data [69]. These approaches develop benchmarks of datasets and queries to evaluate query processing performance across multiple data sources [44] [73] [23] [45]. Such studies cover different aspects of datasets and different optimisation techniques for SPARQL. LUBM [44] represents an early work in this area. More recently, SP2Bench [73] and BSBM [23] have been designed to compare the performance of different RDF stores and their architectures. Another benchmark, FBench [45], uses two scalable datasets that reflect generic and specific domains. The benchmark queries reflect the performance of federated query approaches. All these benchmarks are different in some aspects: datasets size, data domain and the number of queries. However, the only benchmark that provides Linked Data support is FBench.

Listing 1: SPARQL query Example

```
PREFIX owl: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/resource/>
PREFIX foaf:<http://xmlns.com/foaf/0.1/>
SELECT ?name ?dob ?person WHERE {
        ?person owl:birthPlace dbp:Melbourne .
        ?person owl:birthDate ?dob .
        ?person foaf:name ?name}
Limit 10
```

In other work, Langegger [60] has reiterated the call to use of Semantic Web technology, and SPARQL particularly, to provide access to distributed data sources. In their approach, they retrieve data with the use of SPARQL queries. They then introduce a system, called Semantic Web Integrator and Query Engine (SemWIQ), to allow users to execute queries across distributed data sources. They use a combination of matching optional, multiple basic graph and alternative patterns, within an overall platform that uses a pipelined query processing workflow.

Hartig et al. [48] have introduced an approach for executing SPARQL queries over the Web of Linked Data. They present a pipelining approach as an alternative to the problem of HTTP request latency triggering blocking in classical IR architectures. With this approach, query execution and implementation involves the use of an iterator-based pipeline which enables a parallelized and efficient query execution. They implemented the approach using the Semantic Web Client Library (SWClLib) [24]. This highly optimised approach enables applications to query Linked Data on the Web with greater efficiency, maximising the potential of the data.

Other approaches have also been developed based on structured queries; for example, Swoogle [43], ONTOSEARCH2 [81] and NAGA [57, 58].

3.2 Keyword-Based Query

In semantic search, a structured query using a language like SPARQL is often required to retrieve information about an entity. However, most users have only limited knowledge of SPARQL, RDF and ontologies, and would prefer to specify their search intentions with keyword queries. It is therefore useful to be able to apply IR techniques to solve this problem, allowing users to conduct semantic searches to retrieve information from Linked Data sources. One recent approach that combines IR and LOD was proposed by Tonon et al. [82] and employed two techniques: keyword search and inverted index over LOD. Herzig and Tran [52] have also developed a hybrid approach combining structured query and keyword query for data sources integration. Examples of existing search systems that support the keyword-based approach are SWSE [47], Falcons [36], Sindice [86], and Sig.ma [87].

Generally, in a semantic search context, keyword-based queries take forms that are different from general Web searches of largely unstructured or semi-structured document content. As shown by Pound et al. [68], more than half of Web queries target entities. In semantic search, we can further classify queries into: entity name query, attributes query, or relationships query. Section 4 specifically focuses on entity search.

3.3 Exploratory Browsing

Exploration helps users to acquire knowledge and discover more information about their queries. A large body of literature has covered information exploration [89, 64, 88]. However, Linked Data has some additional requirements and considerations that must be taken into account when designing a browser or a user interface. Several studies have discussed these requirements [39, 41]. Also, some studies have specifically focused on entity exploratory search [35, 32].

Unsurprisingly exploratory browsing requires some kind of user interface application. Many end-user tools and browsers have been developed to explore semantic data; for example, Marbles [16], Piggy Bank [54], Sig.ma [87], URIburner[6], SWSE [53], DBpedia Mobile [16, 27], OpenLink Data Explorer (ODE)[7], RDF Gravity[8], RelationshipFinder [66, 61] and Tabulator [19].

While specific patterns of interaction differ widely, an innovative example of addressing the challenges of Linked Data browsing can be found in the template-based visualisation approach proposed by Dadzie et al. [40]. It allows users to discover information according to a *rdf:type*. They use two browsing mechanisms to view information: a graph view to highlight the relations between resources based on the objects properties, and a details view to present information based on properties of datatypes. Their research involved a user study to evaluate the usability of their approach. They found that the graph view is more effective and usable for browsing Linked Data. They further suggested that interaction can be improved once the search path between graphs has been highlighted, or if the browsing starts from a user's history. Unlike the world of browsers in the Web of Documents, however, no consensus or common patterns of interaction yet exist for browing the Web of Data.

4 Entity Search

One of the most significant current discussions in both Information Retrieval and Semantic Web communities concerns *Entity Search* [9, 11, 38, 33]. Entity search involves finding information about individual entities on the Web, such as persons, places, organisations and documents. The Semantic Web provides particularly meaningful description about entities, and the relationships between them, due to the nature of the highly structured knowledge representation standards in OWL and RDF [4, 9, 28]. Linked Data [6, 26], in particular, offers powerful possibilities to improve entity search, because it includes both technical standards and strong social endorsement to the idea of publishing and, as importantly, linking entities.

Several works have tackled the problems of entity retrieval and entity linking [42, 31, 76, 74]. They investigate several issues associated with entity search such as entity ranking, similarity and extraction. Accordingly, we can group all the works for enhancing entity search into three main categories as follows:

1. entity attributes ranking;
2. entity query suggestions;
3. entity type identification.

In the following, we summarise the main approaches for each category.

4.1 Entity Attributes Ranking

As discussed above, the Web of Data—in general—contains a large and growing body of heterogeneous information, sourced from Wikipedia and elsewhere. Search results may, accordingly, offer users a high diversity of attributes and values. If these attributes are not filtered based on users' information needs, then the search process can be time-consuming and frustrating. The key problem described in the literature concerns the retrieval model of entity search, which in part aims to address this problem of noisy, irrelevant or falsely positive results. Some preliminary approaches to addressing the relevance of semantic features include those of Sartori and Lombardi [71], Lombardi and Sartori [62]. In particular, Sartori and Lombardi [71] focus on indexing the level of importance of each feature to a specific concept. The study proposed two measures which can weight semantic features: dominance and distinctiveness. Dominance is a local measure for scoring a feature of the given concept. Distinctiveness is a global measure for scoring a feature across all the other concepts. The other study by Lombardi and Sartori [62] extended this approach by proposing an alternative weighting scheme. They called this scheme FF-ICF. Formally, this scheme can be explained as follows:

- **FF:** means the feature frequency, for a given feature f_i within a specific concept c_j is defined as the number of occurrence of that feature in the c_j concept:
$$ff_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$
where $n_{i,j}$ is the number of occurrences of feature f_i in c_j.
- **ICF:** the inverse concept frequency, defined as:
$$icf_i = \log \frac{|C|}{|\{c : f_i \in c\}|}$$
where $|C|$ is the total number of concepts in the dataset and $|\{c : f_i \in c\}|$ is the number of concepts with feature f_i.

[6] http://linkeddata.uriburner.com/fct
[7] http://ode.openlinksw.com/
[8] http://semweb.salzburgresearch.at/apps/rdf-gravity/

- **FF-ICF** is defined as:

$$fficf_{i,j} = ff_{i,j} * icf_i$$

This scheme is similar to the well-known weighting scheme (TF-IDF) term frequency—inverse document frequency.

In other work, Bazzanella et al. [13] examined a general model for entity representation. The model can identify a core set of attributes in relation to the entity type or a concept. They evaluated this model by conducting a user study [14] where they asked users to nominate a set of queries for particular entities. These queries were used to estimate sets of attributes for different entity types.

More recently, over the past two years, several studies have focused on ranking RDF data. A study by Blanco et al. [29] created their index for attributes based on *BM25F*. BM25F is a model for ranking document fields in IR that uses weighting schema similar to that of TF-IDF. The study also showed that a minimal structure of attributes can effectively improve entity search. The BM25F model has been further extended by Campinas et al. [34], who introduced the MF model for ranking entity attributes. In another study which set out to rank attributes of films, Thalhammer et al. [80] used a game to determine the relevance of entity information. The study established "a game with a purpose" that was able to distinguish between two types of attributes: "interesting" attributes and "too common" attributes. Clearly, some attributes are common but not enough to represent an entity; while some attributes are not common but they are informative attributes. The study showed it was possible for users to vote for answers in a way that led to a model for ranking facts. However, it not clear whether this game could be extended to other domains beyond cinema—a domain where users might be more knowledgeable, passionate and motivated that others.

These studies offer a range of techniques and approaches for ranking and prioritising entity attributes. Outstanding questions, however, include (1) how can users' queries be mapped to related concepts and associated terms that may not be related to concepts? (2) how can a set of the related entity types for these concepts be extracted from the Linked Data? (3) how should aggregated attributes be ranked in relationship to the suggested entity types? The previously mentioned approaches do not yet to address these questions. Consequently entity search still suffers from one key limitation: the lack of a generalisable model for improving the relevance of queries by connecting entities names, types and attributes.

4.2 Entity Query Suggestions

Recently, several authors have adopted a range of techniques to generate query suggestions for semantic data retrieval [22, 65, 2]. In particular, Bhogal et al. [22] reviewed different approaches for query expansion, specifically citing studies on information retrieval with domain-independent and domain-specific ontologies. Query expansion has also been applied in several large knowledge bases such as DBpedia [4] and Yago [77, 78]. These are capable of providing effective suggestions for user queries when searching semantic data, similar to using Wikipedia with query expansion in a document retrieval context [12].

Although extensive research has been carried out on entity search [11, 10], as well as on various entity retrieval models [33, 42], there has been relatively little attention given to the area of *suggestion* for entity queries in semantic search. Query suggestion, or *expansion*, describes a process of interactively augmenting queries with possible extensions. This has long been demonstrated as a useful technique to help users develop queries; previous work in the information retrieval field has shown that the effectiveness of retrieval can be significantly improved using query suggestion [56, 63]. A further study by Qiu and Frei [70] investigated query expansion based on the use of concepts similar to the query term.

4.3 Entity Type Identification

In order to provide meaning to users, entity type identification can help to establish improved ranking for entity attributes, particularly in the case of generic entities such as "Agent" or "Person". Current approaches such as Named Entity Recognition (NER) generate top-level types (such as *Person* and *Place*) when extracting entities from documents for this purpose. One recent study by Tonon et al. [83] has investigated the relevance of entity types based on a collection of statistics to rank all identified types. A related problem is the task of identifying the hierarchical level of an entity type specified in a query, and matching that level with an existing type hierarchy such as the DBpedia ontology. Balog and Neumayer [8] produce baseline models for hierarchical type identification. Their study reveals several findings about automatically annotating queries to relevant entity types in a given ontology. Other approaches such as Paulheim and Bizer [67] and Bazzanella et al. [15] discuss how entity types can reduce noisy information and enhance the entity description in general.

5 Search Interfaces

Users often use a search interface to interact with information on the Web. The design of search interfaces is therefore important to help users to retrieve relevant and meaningful data. A number of published studies are focused on the design of search interfaces [50, 75, 90]. Here we highlight some approaches to search interfaces from two perspectives: the design of search interfaces and the evaluation of search interfaces. Both should be considered when designing an entity search interface.

5.1 Search Interfaces Design

There are two types of design for search interfaces:

Text-based interfaces: these interfaces use textual structures such as tables and lists to present information, properties and relationships. Some also use advanced features such as faceted browsing to allow for more intuitive rendering and navigation of data. In entity search, examples of such interfaces include Sig.ma [87], URIburner[9], SView [37] and SWSE [53].

Visual interfaces: these interfaces use primarily visual or graphic structures such as images, maps, graphs and timelines (individually and in combinations) to represent information. In entity search, examples include VisiNav [46], DBpedia Mobile [16, 27], IsaViz[10], RDF Gravity[11], and RelationshipFinder [66, 61].

[9] http://linkeddata.uriburner.com/fct
[10] http://www.w3.org/2001/11/IsaViz/
[11] http://semweb.salzburgresearch.at/apps/rdf-gravity/

All these search interfaces target the RDF data model for exploration or visualisation, and Alahmari et al. [1] provides an extensive review and comparisons of those listed and others.

5.2 Evaluating Search Interfaces

Many studies have examined general user behaviour on the Web [59, 55], typically employing variants of user-based evaluation methods. Kellar et al. [59], for instance, have examined how users interact with Web browsers based on some information-seeking tasks. In an IR context, modelling user query and navigation behaviour has been examined by Sutcliffe and Ennis [79].

Due to the highly structured and aggregated nature of the data in the Web of Data, it is likely that variations on general search interface evaluation are required. As one example, Tran et al. [84] have carried out a user study regarding complex information needs using Semantic Web data. However there is not yet a common methodology for evaluating the specific forms of search discussed here, such as entity search. It is likely that some consideration of the different kinds of users will be important for the development of such a methodology. Following [39], we suggest three kinds:

Technical users with expertise in the Semantic Web and Linked Data. Such users might use semantic browsers for data retrieval, integration and analysis (so-called "mash-ups"), using advanced filtering and querying services.

Lay users with little or no understanding of underlying semantic technologies. Such users might use semantic browsers for exploring large data sets or finding particular facts of general interest (on DBpedia for example).

Domain experts with expertise in a specific domain, but who may not be familiar with particular Semantic Web and Linked Data technologies. For example, medical researchers might use semantic browsers for advanced domain-specific queries and ontology reasoning.

We note these kinds of users broadly correlate with the three categores of semantic search we discuss above. Technical users are likely to want to use structured queries, while lay users would prefer to use keyword-based queries. Domain experts are equally likely to use keyword-based queries, but might also prefer exploratory browsing, given their greater familiarity with the concepts in the domain. We note also that a more rigorous account of Linked Data users and use cases would help in the development of a common methodology.

6 Conclusion

In this paper, we summarised recent literature related to Linked Data: entity search, and entity search user interfaces. We offered an overview of Linked Data, and also discussed three main semantic search forms: structured query, keyword-based query and exploratory browsing. We also summarise recent studies on entity search and semantic search interfaces. We itemise some specific ways forward for research in this area.

There are several possible directions for further work. One direction is continued research on how entity search can be refined and improved by hiding noisy attributes. Another related direction is a greater understanding of the kinds of users and use cases for Linked Data. Finally, semantic search still lacks a common evaluation methodology. Such a methodology could improve the level of innovation in interface design, and consequently increase the uptake of Linked Data.

References

[1] Fahad Alahmari, James A. Thom, Liam Magee, and Wilson Wong. Evaluating Semantic Browsers for Consuming Linked Data. In *Australasian Database Conference*, volume 124, pages 89–98, 2012.

[2] Fahad Alahmari, James A. Thom, and Liam Magee. A model for ranking entity attributes using DBpedia. *Aslib Journal of Information Management*, 66(5), 2014.

[3] Harith Alani, Lalana Kagal, Achille Fokoue, Paul T. Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha F. Noy, Chris Welty, and Krzysztof Janowicz, editors. *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, volume 8218 of *Lecture Notes in Computer Science*, 2013. Springer.

[4] S. Auer, C Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*, volume 4825. LNCS, 2007.

[5] Sören Auer, Ivan Ermilov, Jens Lehmann, and Michael Martin. LODStats. http://stats.lod2.eu/, 2012. Accessed: 10 Dec 2013.

[6] Ricardo Baeza-Yates and Prabhakar Raghavan. Chapter 2: Next generation web search. In *Search Computing*, volume 5950, pages 11–23. LNCS, 2010.

[7] Ricardo A. Baeza-Yates, Massimiliano Ciaramita, Peter Mika, and Hugo Zaragoza. Towards semantic search. In *Natural Language and Information Systems, 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008, London, UK, June 24-27, 2008, Proceedings*, volume 5039 of *Lecture Notes in Computer Science*, pages 4–11, 2008.

[8] Krisztian Balog and Robert Neumayer. Hierarchical target type identification for entity-oriented queries. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 2391–2394. ACM, 2012.

[9] Krisztian Balog, Edgar Meij, and Maarten de Rijke. Entity search: Building bridges between two worlds. In *Proceedings of the 3rd International Semantic Search Workshop*, SEMSEARCH '10. ACM, 2010.

[10] Krisztian Balog, Arjen P. de Vries, Pavel Serdyukov, and Ji-Rong Wen. The first international workshop on entity-oriented search (eos). *SIGIR Forum*, 45(2):43–50, 2011.

[11] Krisztian Balog, David Carmel, Arjen P. de Vries, Daniel M. Herzig, Peter Mika, Haggai Roitman, Ralf Schenkel, Pavel Serdyukov, and

Thanh Tran Duc. The first joint international workshop on entity-oriented and semantic search (jiwes). *SIGIR Forum*, 46(2):87–94, 2012.

[12] Holger Bast, Debapriyo Majumdar, and Ingmar Weber. Efficient interactive query expansion with complete search. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 857–860. ACM, 2007.

[13] B. Bazzanella, T. Palpanas, and H. Stoermer. Towards a general entity representation model. In *Information Reuse Integration, IEEE International Conference on 2009*, pages 431–432. IEEE Systems, Man, and Cybernetics Society, 2009.

[14] Barbara Bazzanella, Paolo Bouquet, and Heiko Stoermer. Top level categories and attributes for entity representation. 2008. URL http://eprints.biblio.unitn.it/1467/Accessed2013.

[15] Barbara Bazzanella, Heiko Stoermer, and Paolo Bouquet. Entity type disambiguation in user queries. *Journal of Information and Knowledge Management*, 10(03):209–224, 2011.

[16] C. Becker and C. Bizer. Exploring the Geospatial Semantic Web with DBpedia Mobile. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(4):278 – 286, 2009.

[17] T. Berners-Lee. Linked Data–Design Issues, 2006. URL http://www.w3.org/DesignIssues/LinkedData.html. Access Date: 2013.08.04.

[18] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.

[19] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets. Tabulator: Exploring and analyzing linked data on the semantic web. In *In Proceedings of SWUI 2006*, 2006.

[20] Tim Berners-Lee. The next web of open, linked data: Tim berners-lee on ted.com, 2009. URL http://blog.ted.com/2009/03/13/tim_berners_lee_web/.

[21] Tim Berners-Lee and Mark Fischetti. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Harper Information, 2000. ISBN 006251587X.

[22] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information Processing and Management*, 43(4):866 – 886, 2007.

[23] C. Bizer and A. Schultz. Benchmarking the Performance of Storage Systems that expose SPARQL Endpoints. In *4th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS2008)*, 2008.

[24] C. Bizer, T. Gauß, R. Cyganiak, and O. Hartig. Semantic Web Client Library. 2008. URL http://www4.wiwiss.fu-berlin.de/bizer/ng4j/semwebclient/. Access Date: 28 Sep 2010.

[25] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked Data on the Web (LDOW2008). In *Proceeding of WWW '08*, pages 1265–1266, New York, NY, USA, 2008. ACM.

[26] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5, 2009.

[27] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165, 2009.

[28] Roi Blanco, Harry Halpin, Daniel M Herzig, Peter Mika, Jeffrey Pound, Henry S Thompson, and T Tran Duc. Entity search evaluation over structured web data. In *Proceedings of the 1st international workshop on entity-oriented search workshop (SIGIR 2011)*. ACM, 2011.

[29] Roi Blanco, Peter Mika, and Sebastiano Vigna. Effective and Efficient Entity Search in RDF Data. In *International Semantic Web Conference (ISWC)*, pages 83–97. Springer Berlin Heidelberg, 2011.

[30] Stephan Bloehdorn, Marko Grobelnik, Peter Mika, and Duc Thanh Tran, editors. *Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008), Tenerife, Spain, June 2nd, 2008*, volume 334, 2008. CEUR-WS.org.

[31] Ilaria Bordino, Yelena Mejova, and Mounia Lalmas. Penguins in sweaters, or serendipitous entity search on user-generated content. In He et al. [49], pages 109–118.

[32] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, and Davide Mazza. Exploratory search framework for web data sources. *VLDB J.*, 22 (5):641–663, 2013.

[33] S. Campinas, D. Ceccarelli, T. E. Perry, R. Delbru, K. Balog, and G. Tummarello. The Sindice-2011 dataset for entity-oriented search in the web of data. In *1st International Workshop on Entity-Oriented Search (EOS)*, July 2011.

[34] Stéphane Campinas, Renaud Delbru, and Giovanni Tummarello. Effective retrieval model for entity with multi-valued attributes: Bm25mf and beyond. In *EKAW*, volume 7603 of *Lecture Notes in Computer Science*, pages 200–215. Springer, 2012.

[35] David Carmel, Naama Zwerdling, and Sivan Yogev. Entity oriented search and exploration for cultural heritage collections: the eu cultura project. In *WWW (Companion Volume)*, pages 227–230. ACM, 2012.

[36] Gong Cheng and Yuzhong Qu. Searching linked objects with falcons: Approach, implementation and evaluation. *Int. J. Semantic Web Inf. Syst.*, 5(3):49–70, 2009.

[37] Gong Cheng, Saisai Gong, Xiangqian Li, Qingxia Liu, Liang Zheng, Jidong Jiang, Wei Hu, and Yuzhong Qu. SView: Sharing Views for Browsing Linked Data. In *Online proceedings on the Semantic Web Challenge website*, 2013.

[38] Tao Cheng, Xifeng Yan, and Kevin Chen-Chuan Chang. EntityRank: Searching Entities Directly and Holistically. In *VLDB*, pages 387–398. ACM, 2007.

[39] A. Dadzie and M. Rowe. Approaches to Visualising Linked Data: A Survey. *Semantic Web*, 2 (2):89–124, 2011.

[40] Aba-Sah Dadzie, Matthew Rowe, and Daniela Petrelli. *Hide the Stack*: Toward Usable Linked Data. In *The Semantic Web: Research and Applications*, volume 6643 of *LNCS*, pages 93–107. Springer Berlin / Heidelberg, 2011.

[41] Stephen Davies, Jesse Hatfield, Chris Donaher, and Jessica Zeitz. User interface design considerations for linked data authoring environments. In *Proceedings of the WWW2010 Workshop on Linked Data on the Web (LDOW)*, 2010.

[42] Renaud Delbru, Stéphane Campinas, and Giovanni Tummarello. Searching web data: An entity retrieval and high-performance indexing model. *J. Web Sem.*, 10:33–58, 2012.

[43] Tim Finin, Li Ding, Rong Pan, Anupam Joshi, Pranam Kolari, Akshay Java, and Yun Peng. Swoogle: Searching for knowledge on the semantic web. In *In AAAI 05 (intelligent systems demo*, pages 1682–1683. The MIT Press, 2005.

[44] Y. Guo, Z. Pan, and J. Heflin. LUBM: A benchmark for OWL knowledge base systems. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):158 – 182, 2005.

[45] P. Haase, T. Mathäß, and M. Ziller. An evaluation of approaches to federated query processing over linked data. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 5:1–5:9, New York, NY, USA, 2010. ACM.

[46] Andreas Harth. Visinav: A system for visual search and navigation on web data. *J. Web Sem.*, 8(4):348–354, 2010.

[47] Andreas Harth, Aidan Hogan, Renaud Delbru, Jürgen Umbrich, Seán O'Riain, and Stefan Decker. Swse: Answers before links! In *Semantic Web Challenge*, 2007.

[48] O. Hartig, C. Bizer, and J. Freytag. Executing SPARQL Queries over the Web of Linked Data. In *The Semantic Web - ISWC 2009*, volume 5823 of *Lecture Notes in Computer Science*, pages 293–309. Springer Berlin / Heidelberg, 2009.

[49] Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Rajeev Rastogi, editors. *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, 2013. ACM.

[50] M.A. Hearst. Search user interfaces. *Cambridge University Press*, 2009.

[51] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan and Claypool, 2011.

[52] Daniel M. Herzig and Thanh Tran. Heterogeneous web data search using relevance-based on the fly data integration. In Alain Mille, Fabien L. Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab, editors, *WWW*, pages 141–150. ACM, 2012.

[53] Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker. Searching and browsing linked data with swse: The semantic web search engine. *J. Web Sem.*, 9(4):365–401, 2011.

[54] D. Huynh, S. Mazzocchi, and D. Karger. Piggy Bank: Experience the Semantic Web Inside Your Web Browser. In *ISWC 2005*, volume 3729 of *LNCS*, pages 413–430. Springer Berlin / Heidelberg, 2005.

[55] Bernard J. Jansen. *Understanding User-Web Interactions via Web Analytics*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.

[56] Hideo Joho, Mark Sanderson, and Micheline Beaulieu. A study of user interaction with a concept-based interactive query expansion support tool. In Sharon McDonald and John Tait, editors, *ECIR*, volume 2997 of *Lecture Notes in Computer Science*, pages 42–56. Springer, 2004.

[57] G. Kasneci, F.M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. Naga: Searching and ranking knowledge. In *IEEE 24th International Conference on Data Engineering, 2008. ICDE 2008.*, pages 953–962. IEEE, 2008.

[58] Gjergji Kasneci, Fabian M. Suchanek, Maya Ramanath, and Gerhard Weikum. How naga uncoils: searching with entities and relations. In *WWW*, pages 1167–1168, 2007.

[59] Melanie Kellar, Carolyn Watters, and Michael Shepherd. A field study characterizing web-based information-seeking tasks. *Journal of the American Society for Information Science and Technology*, 58(7):999–1018, 2007.

[60] A. Langegger. Virtual data integration on the web: novel methods for accessing heterogeneous and distributed data with rich semantics. In *Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*, iiWAS '08, pages 559–562, New York, NY, USA, 2008. ACM.

[61] S. Lohmann, P. Heim, T. Stegemann, and J. Ziegler. The RelFinder User Interface: Interactive Exploration of Relationships between Objects of Interest. In *Proceedings of IUI 2010*, pages 421–422, New York, NY, USA, 2010. ACM.

[62] Luigi Lombardi and Giuseppe Sartori. Models of relevant cue integration in name retrieval. *Journal of Memory and Language*, 57(1):101 – 125, 2007.

[63] Mark Magennis and Cornelis J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proceedings of the 20Th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, pages 324–332, New York, NY, USA, 1997. ACM.

[64] Gary Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49 (4):41–46, 2006.

[65] Edgar Meij, Marc Bron, Laura Hollink, Bouke Huurnink, and Maarten de Rijke. Mapping queries to the Linking Open Data cloud: A case

study using DBpedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9 (4):418 – 433, 2011.

[66] S. Lohmann P. Heim and T. Stegemann. Interactive Relationship Discovery via the Semantic Web. In *Proceedings of ESWC 2010*, volume 6088 of *LNCS*, pages 303–317, Berlin/Heidelberg, 2010. Springer.

[67] Heiko Paulheim and Christian Bizer. Type inference on noisy rdf data. In Alani et al. [3], pages 510–525.

[68] Jeffrey Pound, Peter Mika, and Hugo Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 771–780, New York, NY, USA, 2010. ACM.

[69] E. Prud'Hommeaux, A. Seaborne, et al. SPARQL query language for RDF. *W3C working draft*, 4, 2006. URL http://www.w3.org/TR/rdf-sparql-query/. Access Date: 28 Sep 2010.

[70] Yonggang Qiu and Hans-Peter Frei. Concept based query expansion. In *SIGIR*, pages 160–169, 1993.

[71] Giuseppe Sartori and Luigi Lombardi. Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, 16(3):439–452, 2004.

[72] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the linked data best practices in different topical domains. In *13th International Semantic Web Conference (ISWC2014)*, 2014.

[73] M. Schmidt, T. Hornung, M. Meier, C. Pinkel, and G. Lausen. SP2Bench: A SPARQL Performance Benchmark. In *Semantic Web Information Management*, pages 371–393. Springer Berlin Heidelberg, 2008.

[74] Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. LINDEN: Linking Named Entities With Knowledge Base via Semantic Knowledge. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 449–458, New York, NY, USA, 2012. ACM.

[75] B. Shneiderman, C. Plaisant, M. Cohen, and S. Jacobs. *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Addison Wesley, 5th edition, March 2009.

[76] Avirup Sil and Alexander Yates. Re-ranking for joint named-entity recognition and linking. In He et al. [49], pages 2369–2374.

[77] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *WWW*, pages 697–706, 2007.

[78] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A large ontology from wikipedia and wordnet. *J. Web Sem.*, 6(3):203–217, 2008.

[79] Alistair Sutcliffe and Mark Ennis. Towards a cognitive theory of information retrieval. *Interacting with Computers*, 10(3):321 – 351, 1998.

[80] Andreas Thalhammer, Magnus Knuth, and Harald Sack. Evaluating entity summarization using a game-based ground truth. In Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Manfred Hauswirth, Josiane Xavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein, and Eva Blomqvist, editors, *International Semantic Web Conference (2)*, volume 7650 of *Lecture Notes in Computer Science*, pages 350–361. Springer, 2012.

[81] Edward Thomas, Jeff Z. Pan, and Derek H. Sleeman. Ontosearch2: Searching ontologies semantically. In *OWLED*, 2007.

[82] Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *SIGIR*, pages 125–134. ACM, 2012.

[83] Alberto Tonon, Michele Catasta, Gianluca Demartini, Philippe Cudré-Mauroux, and Karl Aberer. Trank: Ranking entity types using the web of data. In Alani et al. [3], pages 640–656.

[84] Thanh Tran, Tobias Mathäß, and Peter Haase. Usability of keyword-driven schema-agnostic search: a comparative study of keyword search, faceted search, query completion and result completion. In *Proceedings of the 7th international conference on The Semantic Web: research and Applications - Volume Part II*, ESWC'10, pages 349–364, Berlin, Heidelberg, 2010. Springer-Verlag.

[85] Thanh Tran, Daniel Herzig, and Günter Ladwig. Semsearchpro - using semantics throughout the search process. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4), 2011.

[86] G. Tummarello, R. Delbru, and E. Oren. Sindice.com: Weaving the Open Linked Data. In *The Semantic Web*, volume 4825, pages 552–565. Springer Berlin / Heidelberg, 2007.

[87] G. Tummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker. 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.

[88] Daniel Tunkelang. *Faceted Search.* Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.

[89] Ryen W. White and Resa A. Roth. *Exploratory Search: Beyond the Query-Response Paradigm.* Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.

[90] Max L. Wilson. *Search User Interface Design.* Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2011.

[91] David Wood. *Linking Enterprise Data.* Springer US, 2010.

[92] David Wood. *Linking Government Data.* Springer US, 2011.