Querying Linked Data Graphs using Semantic Relatedness: A Vocabulary Independent Approach

André Freitas[a], João Gabriel Oliveira[a,b], Seán O'Riain[a], João C. P. da Silva[b], Edward Curry[a]

[a] *Digital Enterprise Research Institute (DERI)*
National University of Ireland, Galway
[b] *Computer Science Department*
Mathematics Institute
Federal University of Rio de Janeiro (UFRJ)

Abstract

Linked Data brings inherent challenges in the way users and applications consume the available data. Users consuming Linked Data on the Web, should be able to search and query data spread over potentially large numbers of heterogeneous, complex and distributed datasets. Ideally, a query mechanism for Linked Data should abstract users from the representation of data. This work focuses on the investigation of a vocabulary independent natural language query mechanism for Linked Data, using an approach based on the combination of entity search, a Wikipedia-based semantic relatedness measure and spreading activation. Wikipedia-based semantic relatedness measures address existing limitations of existing works which are based on similarity measures/term expansion based on WordNet. Experimental results using the query mechanism to answer 50 natural language queries over DBpedia achieved a mean reciprocal rank of 61.4%, an average precision of 48.7% and average recall of 57.2%.

Keywords: Natural language queries, semantic relatedness, vocabulary independent queries, RDF, Linked Data

1. Introduction

Linked Data [1] has emerged as a de-facto standard for publishing data on the Web. With it comes the potential for a paradigmatic change in the scale in which users and applications reuse, consume and repurpose data. Linked

Data, however, brings inherent challenges in the way users and applications consume existing data. Users accessing Linked Data should be able to search and query data spread over a potentially large number of different datasets. The freedom, simplicity and intuitiveness provided by search engines in the Web of Documents were fundamental in the process of maximizing the value of the information available on the Web, approaching the Web to the casual user.

However the approaches used for searching the Web of Documents cannot be directly applied for searching/querying data, since from the perspective of structured/semi-structured data consumption, users expect expressive queries, where they can query and operate over the structural information in the data. In order to query Linked Data today, users need be aware of the structure and terms used in the data representation. In the Web scenario, where data is spread across multiple and highly heterogeneous datasets, the vocabulary gap between users and datasets (i.e. the difference between the terms and structure of user queries and the representation of data) becomes one of the most important issues for Linked Data consumers. At Web scale it is not feasible for users to become aware of all the vocabularies that the data can be represented in order to formulate a query. Ideally from the users' perspective, they should be abstracted away from the data representation. Additionally, from the perspective of user interaction, a query mechanism for Linked Data should be simple and intuitive for casual users. The suitability of natural language for search and query tasks was previously investigated by Kauffman [2].

This work focuses on the investigation of a fundamental type of query mechanism for Linked Data: the provision of *vocabulary independent and expressive natural language queries for Linked Data*, expanding the discussion present in [37, 39]. This type of query fills an important gap in the spectrum of search/query services for the Linked Data Web, allowing users to expressively query the contents of distributed linked datasets without the need for a prior knowledge of the vocabularies behind the datasets.

The rationale behind the proposed approach is to use large volumes of semantic information ambedded in comprehensive Web corpora such as Wikipedia to guide the semantic matching between query and dataset terms, and the navigational process over the Linked Data Web. The distributional information provides a commonsense associative quantitative semantic model which is used as a semantic relatedness measure between the user query terms (which reflects his/her information needs) and dataset terms.

This paper is structured as follows. Section 2 briefly introduces Linked Data and Entity-Attribute-Value (EAV) databases. Section 3 describes the proposed query mechanism, detailing how the three elements (*entity search*, *spreading activation* and *corpus-based semantic relatedness*) are used to build the query mechanism. Section 4 covers the evaluation of the approach, followed by section 5 which provides a discussion on the proposed query approach. Section 6 describes related work in the area and finally, section 7 provides a conclusion and future work.

2. Entity-Attribute-Value (EAV) Databases, RDF(S) & Linked Data

The Linked Data Web uses the Resource Description Framework (RDF) to provide a structured way of publishing information describing entities and its relations on the Web. RDF allows the definition of names for entities using Universal Resource Identifiers (URIs). RDF triples, following the *subject*, *predicate*, *object* structure, allows the grouping of entities into named classes, the definition of named relations between entities, and the definition of named attributes of entities using literal values.

Linked Data consists in a set of principles [1] for publishing structured data as RDF on the Web, allowing RDF datasets to be interlinked across the Web. For an RDF dataset published as Linked Data, each URI describing a concept in the dataset (e.g. *Paris*), can have its specific RDF description (the set of triples containing information about Paris, e.g. *Paris isLocatedIn France*) fetched through HTTP (this process is called dereferenciation). Linked Data allows users to navigate across datasets on the Web, by following the links (triples) connecting entities in different datasets.

The core elements of the semantic model of the typed labeled graph behind RDF(S) for Linked Data assume that: (i) instances can be organized into sets (classes); (ii) classes can be elements of other classes (taxonomic structure) and (iii) instances can have associated properties connecting them with other instances or values. These core elements do not include blank nodes, reification and containers, valid elements in RDF which are not recommended for the publication of RDF as Linked Data.

Entity attribute value (EAV) is a data model to describe entities where the number of attributes (properties, parameters) that can be used to describe them is potentially vast, but the number that will actually apply to a given entity is relatively modest [3]. The EAV model can be defined by a sparse matrix and is associated with open/dynamic schema databases. EAV can be

seen as the more abstract data model behind RDF(S). An EAV data model is composed of three core elements:
- *entity:* the element being described. In RDF(S) the entity maps to an *instance* or *class*.
- *attribute:* the attribute definition. In RDF(S) an attribute maps to a *predicate*.
- *value:* The value assigned to an attribute. In RDF(S) it maps to an object which can be a *resource* or *literal value*.

On the top of the EAV abstraction, RDF(S) also defines a canonical ordering of a triple (s, p, o) and a *rdfs:type* relation which is used to assert that a given instance in the *rdfs:domain* of type is in a set defined by a class in its range, allowing the definition of a taxonomic structure. An EAV model with the characteristics above is named a EAV/CR (EAV with Classes and Relationships). The fact that EAV is an entity-centric model allows an entity-based data integration process which, together with the Web URI, HTTP and RDF standards defines a de-facto data integration model for the Linked Data Web. In the context of this work, Linked Data and EAV/CR are used interchangeably.

Open schema and entity-centric data integration are core characteristics of the EAV model which are central to Web data, but also on new data environments with high data heterogeneity, complexity and variability. In these scenarios the traditional structured query mechanisms which demands users to understand the schema or vocabulary in the dataset, limits the ability of data consumers to query the data.

3. Query Processing Approach

3.1. Motivation

The central motivation behind this work is to propose a Linked Data query mechanism for casual users maximizing *flexibility* (vocabulary independency), *expressivity* (ability to query structures in the data), *usability* (provided by natural language queries) and ability to query distributed data. In order to address these requirements, this paper proposes the construction of a query mechanism based on the combination of *entity search*, a *corpus-based semantic relatedness measure* and *spreading activation*. Our contention is that the combination of these elements provides the support for the construction of a natural language query mechanism for Linked Data with an additional level of *vocabulary independency*.

The query types covered in this work concentrate on the following query graph patterns:

[1]. $instance \rightarrow predicate \rightarrow [instance|value]$
[2].$instance_0 \rightarrow predicate_0 ... predicate_n \rightarrow [instance_n|value]$
[3].$class \rightarrow type \rightarrow instance$

These graph patterns are at the core of the RDF data model. Most of the corresponding SPARQL graph patterns matching the natural language queries from the Question Answering over Linked Data test collection [14] include these patterns. Query patterns 1 and 2 can be combined with query pattern 3 in a single query. Queries with aggregation and conditional operators are not covered in the scope of this work.

The remainder of this section describes the proposed approach and its main components.

3.2. Approach & Outline

The query mechanism proposed in this work receives as an input a *natural language query* and outputs a set of *triple paths*, which are the triples corresponding to answers merged into a connected graph. Figure 1 organizes the components of the approach into a high level architecture.

The query processing approach starts by determining the *key entities* present in the natural language query. *Key entities* are entities which can be potentially mapped to *instances* or *classes* in the Linked Data Web. After detected, the key entities are sent to the entity search engine which resolve them to *pivot entities* in the Linked Data Web. A *pivot entity* is an URI which represents an entry point for the search in the Linked Data Web (Figure 2). The process of determining the *key and pivot entities* are covered in section 3.3.

After the key entities and pivots are determined, the user natural language query is analyzed in the *query parsing* component. The output of this component is a structure called *partial ordered dependency structure* (PODS), which is a reduced representation of the query targeted towards maximizing the matching between the structure of the terms present in the query and the *subject, predicate, object* structure of RDF. The partial ordered dependency structure is generated by applying a dependency parsing step [5] over the natural language query and by transforming the generated Stanford dependency structure into a PODS (section 3.4).

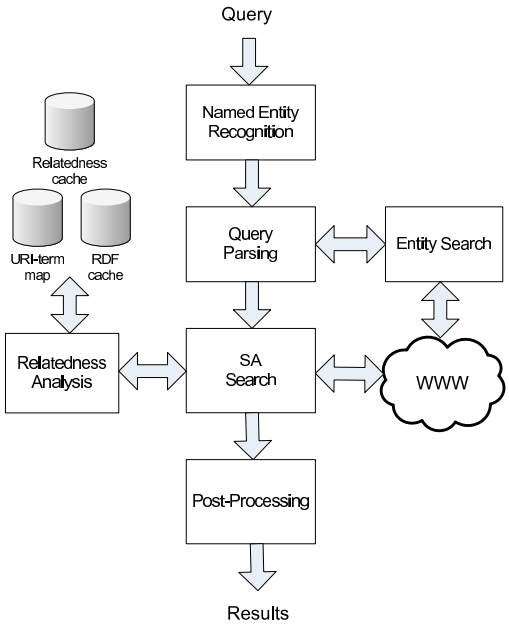

Figure 1: High-level architecture of the proposed approach.

Taking as an input the list of URIs of the pivot entities and the partial ordered dependency structure, the query processing algorithm follows a *spreading activation search* (section 3.6) where nodes in the Linked Data Web are explored by using a *measure of semantic relatedness* (section 3.5) to match the query terms present in the PODS to terms representing Linked Data entities (classes, properties and instances). Starting from a pivot entity, the node exploration process in the spreading activation search is done by computing the semantic relatedness measure between the query terms and terms corresponding to dataset elements in the Linked Data Web. The semantic relatedness measure, combined with a statistical threshold, which determines the discrimination of the winning relatedness scores, works as a spreading activation function which will determine the nodes which will be explored in the Linked Data Web.

Figure 2 depicts the core part of the query processing for the example query *'From which university did the wife of Barack Obama graduate?'*. After parsing the natural language query into a partial ordered dependency structure (PODS) (light gray nodes), and after the pivot is determined (*dbpedia: Barack_Obama*) by the entity search step, the algorithm follows computing the semantic relatedness between the next query term (*'wife'*) and all

the properties, associated types and instance labels linked to the node dbpedia:Barack_Obama (*dbpedia-owl:spouse*, *dbpedia-owl:writer*, *dbpedia-owl:child*, ...). Nodes above a certain relatedness threshold are further explored (dereferenced). The matching process continues until all query terms are covered.

In the example, after the matching between *wife* and *dbpedia-owl: spouse* is defined (2), the object pointed by the matched property (*dbpedia: Michelle _Obama*) is dereferenced (3), and the RDF of the resource is retrieved. The next node in the PODS is *graduate*, which is mapped to both *dbpedia-owl:University* and *dbpedia-owl:Educational_Institution* (4) specified in the types. The algorithm then navigates to the last node of the PODS, *university*, dereferencing *dbpedia:Princeton_University* and *dbpedia:Harvard_Law_School* (5), matching for the second time with their type class (6). Since the semantic relatedness between the terms is high, the terms are matched and the algorithm stops, returning the subgraph containing the triples which maximize the relatedness measure between the query terms and the vocabulary terms. The proposed algorithm works as a *semantic best-effort query approach*, where the semantic relatedness measure provides a *semantic ranking* of returned triples.

The output of the algorithm is a list of ranked triple paths, triples following from the pivot entity to the final resource representing the answer, ranked by the average of the relatedness scores in the path. Answers are displayed to users using a list of triple paths and a graph which is built by merging the triple paths on a simple post-processing phase. This graph can be used by the user for further complementary exploration by navigation over the answer set. The query mechanism described above was implemented in a prototype named *Treo*, the word for *direction* in Irish.

3.3. Entity Recognition and Entity Search

The query processing approach starts by determining the set of key entities (pivot candidates) that will be used in the generation of the partial ordered dependency structure and in the determination of the final pivot entity. The process of generating a pivot candidate starts by detecting named entities in the query. The named entity recognition (NER) approach used is based on Conditional Random Fields sequence models [6] trained in the CoNLL 2003 English training dataset [8], covering people, organizations and locations. Named entities are likely to be mapped to the URIs of individuals in the Linked Data Web. After the named entities are identified, the query is tagged by a part-of-speech (POS) tagger, which assigns grammatical classes

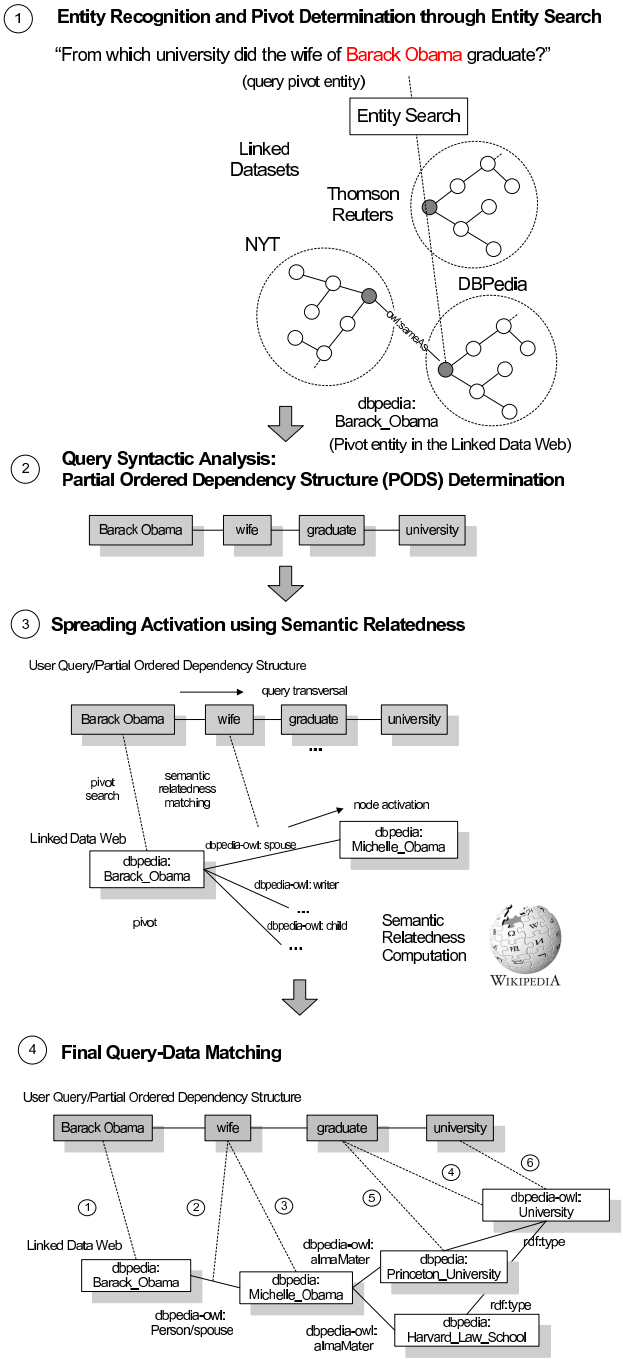

Figure 2: The execution of the semantic relatedness spreading activation algorithm for the question *'From which university did the wife of Barack Obama graduate?'*.

to the query terms. The POS tags are used to determine pivot candidates which are not named entities (typically mapping to classes). The POS tagger used is a log-linear POS tagger [9]. For the example query, the named entity *Barack Obama* is recognized as the main entity.

The terms corresponding to the *pivot entity candidates* are sent to an *entity centric-search engine* that will resolve the terms into the final pivot entities URIs in the Linked Data Web. Entity-centric search engines for Linked Data are search engines where the search is targeted towards the retrieval of individual instances, classes and properties in the Linked Data Web. The entity index uses a simple TF/IDF weighting scheme [7] over labels parsed from *instances* and *classes* URIs. The query mechanism prioritizes named entities as pivots. In case the query has more than one entity, both entity candidates are sent to the entity search engine and the entity cardinality (number of properties connected to the entity) together with a string similarity score over the result set (dice coeficient) is used to determine the final pivot entities. In the example query, the named entity *Barack Obama* is mapped to a list of URIs representing the entity Barack Obama in different datasets (e.g. *http://dbpedia.org/resource/Barack_Obama*).

3.4. Query Parsing

The semantic relatedness spreading activation algorithm takes as one of its inputs a *partial ordered dependency structure* (PODS) which is a directed acyclic graph connecting a subset of the original terms present in the query. The idea behind PODSs is to provide a representation of the natural language input which could be easily mapped to the *(subject, predicate, object)* structure of an RDF representation. Partial ordered dependency structures are derived from Stanford typed dependencies [5] which represents a set of bilexical relations between each term of a sentence, providing grammatical relation labels over the dependencies. Additional details covering Stanford dependencies can be found in [5].

The query parsing module builds PODSs by taking as inputs both Stanford dependencies and the detected named entities/pivots and by applying a set of operations over the original Stanford dependencies. These operations produce a reduced and ordered version of the original elements of the query. The pivots and named entities combined with the original dependency structure determine the ordering of the elements in the structure.

Definition 1. Let $T(V, E)$ be a typed Stanford dependency structure over

the question Q. The partial ordered dependency structure $D(V, E)$ of Q is defined by applying the following operations over T:

- *merge* adjacent nodes V_K and $V_{K+1} \in T$ where $E_{K,K+1} \in \{$nn, advmod, amod$\}$.
- *eliminate* the set of nodes V_K and edges $E_K \in T$ where $E_K \in \{$advcl, aux, auxpass, ccomp, complm, det$\}$.
- *replicate* the triples where $E_K \in \{$cc, conj, preconj$\}$.

where the edge labels *advmod, amod, etc* represent the specific dependency relations (see [5] for a the complete list of dependencies).

In the definition above, the *merge* operation consists in collapsing adjacent nodes into a single node for the purpose of merging multi-word expressions, in complement with the entity recognition output. The *eliminate* operation is defined by the pruning of a node-edge pair and eliminates concepts which are not semantically relevant or covered in the representation of data in RDF. The *replicate* operation consists in copying the remaining elements in the PODS for each coordination or conjunctive construction.

The traversal sequence should maximize the likelihood of the isomorphism between the *partial ordered dependency structure* and the *subgraph of the Linked Data Web*. The traversal sequence is defined by taking the pivot entity as the root of the partial ordered dependency structure and following the dependencies until the end of the structure is reached. In case there is a cycle involving one pivot entity and a secondary named entity, the path with the largest length is considered. For the example query the partial ordered dependency structure returned by the query parser is: *Barack Obama* $\rightarrow$ *wife* $\rightarrow$ *graduate* $\rightarrow$ *university*.

3.5. Semantic Relatedness

After the natural language query is parsed into the PODS and after the determination of the list of entity pivots, the spreading activation search process in the Linked Data Web starts. The center of the spreading activation search proposed in this work is the use of a semantic relatedness measure as the activation function in the matching process between query terms and vocabulary terms. The proposed approach is highly dependent on the quality of the semantic relatedness measure. This section briefly describes the basic concepts behind the semantic relatedness measures used in the algorithm.

The problem of measuring the semantic relatedness and similarity of two words can be stated as follows: given two words A and B, determine a measure f(A,B) which expresses the semantic proximity between words A and B. The notion of semantic *similarity* is associated with taxonomic (is-a) relations between words, while semantic *relatedness* represents more general classes of relations [10]. Since the problem of matching natural language terms to concepts present in Linked Data vocabularies can cross taxonomic boundaries, the generic concept of semantic relatedness is more suitable to the task of semantic matching for queries over the Linked Data Web. In the example query, the relation between *'graduate'* and *'University'* is non-taxonomic and a purely similarity analysis would not detect appropriately the semantic proximity between these two words. In the context of semantic query by spreading activation, it is necessary to use a relatedness measure that: (i) can cope with terms crossing part-of-speech boundaries (e.g. between verbs and nouns); (ii) measure relatedness among multi-word expressions; (iii) are based on a comprehensive knowledge base.

Existing approaches for semantically querying and searching Semantic Web/ Linked Data knowledge bases are mostly based on WordNet similarity measures or on exploring ontological relations in the query dataset (see Related Work section). WordNet-based similarity measures [11] are highly dependent on the structure and scope of the WordNet model, not addressing the requirements above. Distributional relatedness measures [16, 11] are able to meet the previous requirements, providing approaches to build semantic relatedness measures based on word co-occurrence patterns present in large Web corpora. Complementarily, recent approaches propose a better balance between the cost associated in the construction of the relatedness measure and the accuracy provided, by using the link structure present in the corpora. Examples of this class of measures is the Wikipedia Link-based Measure (WLM) [17], and Explicit Semantic Analysis (ESA) [16] which achieved high correlation with human assessments.

The following sections introduce the basic principles behind distributional semantics and the two semantic relatedness measures used in this work.

3.5.1. Corpus-based Semantic Relatedness & Distributional Semantics

Distributional semantics is built upon the assumption that the context surrounding a given word in a text provides important information about its meaning [49]. A rephrasing of the *distributional hypothesis* states that words that co-occur in similar contexts tend to have similar meaning [49].

Distributional semantics focuses on the construction of a semantic representation of a word based on the statistical distribution of word co-occurrence in texts. The availability of high volume and comprehensive Web corpora brought distributional semantic models as a promising approach to build and represent meaning. Distributional semantic models are naturally represented by Vector Space Models (VSMs), where the meaning of a word is represented by a weighted concept vector.

However, the proper use of the simplified model of meaning provided by distributional semantics implies understanding its characteristics and limitations. The distributional view on meaning does not refer to extra-linguistic representations of the object related to the word and it is inherently differential [50]: the differences of meaning are mediated by differences of the word distribution in the corpus. As a consequence, distributional semantic models allow the quantification of the amount of difference in meaning between words, allowing a comparative quantification on the semantic proximity between two words. This differential analysis can be used to determine the semantic relatedness between words. Therefore, the applications of the meaning defined by distributional semantics should focus on a space where its differential nature is suitable. The computation of semantic relatedness and similarity measures between pair of words is one instance in which the strength of distributional models and methods is empirically supported [16]. This work focuses on the use of semantic models derived from large corpora in the computation of semantic relatedness measures and as a key element to address the level of semantic flexibility necessary for the provision of vocabulary independent queries.

3.5.2. Explicit Semantic Analysis (ESA)

Explicit Semantic Analysis (ESA) [16] is a distributional semantic approach built from Wikipedia corpora. ESA provides an approach which can be used to compute an explicit semantic interpretation of a given word as a set of weighted concept vectors. In the case of ESA, the set of returned weighted concept vectors associated with a word is represented by titles of Wikipedia articles. A *universal ESA space* is created by building a vector space of Wikipedia articles using the traditional term frequency/inverse document frequency (TF/IDF) weighting scheme. In this space each article is represented as a vector where each component is a weighted term present in the article. Once the space is built, a keyword query over the ESA space returns the list of ranked articles titles, which define a concept vector associ-

ated with the query terms (where each vector component receives a relevance score). The approach also allows the interpretation of text fragments where the final concept vector is the centroid of the concept vectors representing the set of individual terms. This procedure allows the approach to partially perform word sense disambiguation [16]. The ESA semantic relatedness measure between two terms or text fragments is calculated by computing the cosine similarity between the concept vectors representing the interpretation of the two terms or text fragments.

3.5.3. Wikipedia Link Measure (WLM)

The WLM measure is built based on the link structure between Wikipedia articles. The process of creation of the WLM relatedness measure starts by computing weights for each link, where the significance of each link receives a score. This procedure is equivalent to the computation of the TF/IDF weighting scheme for the links in the place of terms: links pointing to popular target articles (receiving links from many other articles) are considered less significant from the perspective of semantic relatedness computation. The weight score is defined below:

$$w(s \rightarrow t) = log\left(\frac{\mid W \mid}{\mid T \mid}\right), \text{if s} \in T, 0 \text{ otherwise} \tag{1}$$

where s and t represent the source and target articles, W is the total number of articles in Wikipedia and T is the number of articles that link to t. The semantic relatedness measure is defined by an adaptation over the Normalized Google Distance (NGD)[17] [31].

$$r(a,b) = \frac{log(max(\mid A \mid, \mid B \mid)) - log(\mid A \cap B \mid)}{log(\mid W \mid) - log(min(\mid A \mid, \mid B \mid))} \tag{2}$$

where a and b are the two terms that the semantic relatedness is being measured, A and B are the respective articles that are linked to a and b and W is the set of all Wikipedia articles. The final relatedness measure uses a combination of the two measures. The reader is directed to [17] for additional details on the construction of the relatedness measure.

3.6. The Semantic Relatedness Spreading Activation Algorithm

3.6.1. Introduction

Spreading activation is a search technique used in graphs and, particularly in semantic or associative networks, based on the idea of using an activation

function as a threshold for the node exploration process. Spreading activation techniques have a long history in cognitive psychology, artificial intelligence and, more recently, on information retrieval.

The idea of spreading activation has its origins associated with the modeling of the human semantic memory in cognitive psychology. The spreading activation theory of human semantic processing was first proposed by Quillian [24] and later extended by Collins & Loftus [25]. The spreading activation model introduced by Quillian [24] was proposed in the context of semantic networks, a graph of interlinked concepts which contains some of the elements under the RDF(S) specifications [4].

The processing technique of spreading activation is simple [26]: it consists of one or more propagated pulses and a termination check. Additionally, the model can implement propagation decays and constraints on the spreading activation process. Types of constraints include distance constraints, fan-out, path and activation constraints. The difference among different spreading activation models reside in the characteristics of the constraints adopted. The reader is referred the *Related Work* section for applications of spreading activation in information retrieval.

3.6.2. The Spreading Activation Algorithm

The semantic relatedness spreading activation algorithm takes as an input a partial ordered dependency structure D(V, E) and searches for paths in the Linked Data Web graph W(V, E) which maximizes the semantic relatedness between D and W taking into account the ordering of both structures. The first element in the partial ordered dependency structure is the pivot entity which defines the first node to be dereferenced in the graph W. After the pivot element is dereferenced, the algorithm computes the semantic relatedness measure between the next term in the PODS and the *properties*, *associated types* and *instance terms* in the Linked Data Web. *Associated types* represent the types associated to an instance through the *rdfs:type* relation. While properties and ranges are defined in the terminological level, type terms require an a priori instance dereferenciation to collect the associated types. The relatedness computation process between the next query term k and a neighboring node n takes the maximum of the relatedness score between properties p, types c and instance terms i:

$$r_{k,n} = max(r(k,p), r(k,i), \underset{\forall c \in \mathrm{C}}{max}(r(k,c))) \tag{3}$$

Nodes above the semantic relatedness score threshold determine the node URIs which will be activated (dereferenced). The *activation function* is given by an adaptive discriminative relatedness threshold which is based on the set of relatedness scores between the pairs of query and dataset terms. The adaptive threshold has the objective of selecting the relatedness scores with higher discrimination and it is defined as a function of the standard deviation σ of the relatedness scores. The activation threshold *a(I)* of a node I is defined as:

$$a(I) = \mu(r) + \alpha \times \sigma(r) \tag{4}$$

where I is the current node instance, $\mu(r)$ is the mean of the relatedness values associated with each node instance, $\sigma(r)$ is the standard deviation of the relatedness values and α is a empirically determined constant. The value of α was determined by calculating the difference in terms of $\sigma(r)$ of the relatedness value of the correct node instances and the average relatedness value for a random 50% sample of the nodes instances involved in the spreading activation process in the query dataset. The empirical value found for α was 2.185. In case no node is activated for the first value of α, the original value decays by an exponential factor of 0.9 until it finds a candidate node above *a(I)*.

In case the algorithm finds a node with high semantic relatedness which has a literal value as an object (non-dereferenceable), the value of the node can be re-submitted to the entity search engine. In the case an URI is mapped, the search continues from the same point in the partial ordered dependency structure, in a different pivot in the Linked Data Web (working as an entity reconciliation step). The stop condition for the spreading activation is defined by the end of the query.

The final semantic relatedness spreading activation algorithm is defined below:

$D(V_G, E_G)$: partial ordered dependency graph
$W(V_W, E_W)$: Linked Data graph
$A(V_A, E_A)$: answer graph
$pivots$: set of pivots URI's
for all p in $pivots$ **do**
  $initialize(A, p)$
  while $hasUnvisitedNodes(V_A)$ **do**
    $v \leftarrow nextNode(V_A)$

$dereference(v, W)$
for all $nextT$ in $getNextDependencyNodes(D)$ **do**
$best \leftarrow bestNodes(v, nextT, W)$
$update(V_A, best)$
$update(E_A, best)$
end for
end while
end for

From a performance perspective the use of type verification in the node exploration process can bring high latencies in the node exploration process. In order to be effective, the algorithm should rely on mechanisms to reduce the number of unnecessary HTTP requests associated with the dynamic dereferenciation process, unecessary URI parsing or label checking and unnecessary semantic relatedness computation. The initial *Treo* prototype has three local caches implemented: one for RDF, the second for semantic relatedness measures pairs and the third for URI/label-term mapping. Another important practical aspect which constitutes one of the strengths of the approach is the fact that it is both highly and easily parallelizable in the process of semantic relatedness computation and on the dereferenciation of URIs.

4. Evaluation

The focus of the evaluation is to determine the relevance of the results provided by the query mechanism. With this objective in mind the query mechanism was evaluated by measuring *average precision*, *average recall* and *mean reciprocal rank* for natural language queries using DBpedia [13], a dataset in the Linked Data Web. DBpedia 3.6 (February 2011) contains 3.5 million entities, where 1.67 million are classified in a consistent ontology. The use of DBpedia as a dataset allows the evaluation of the system under a realistic scenario. The set of natural language queries annotated with answers were provided by the training test collection released for the Question Answering for Linked Data (QALD 2011) workshop [14] containing queries over DBpedia 3.6. From the original query set, 5 queries were highly dependent on comparative operations (e.g. *'What is the highest mountain?'*). The removed queries were substituted with 5 additional queries exploring more challenging cases of query-vocabulary semantic gap matching. The reader can find additional details on the data used in the evaluation and the associated results

in [15].

In the scope of this evaluation an answer is a set of ranked triple paths. Different from a SPARQL query, the algorithm works as a *semantic best-effort* approach, where the semantic relatedness activation function works both as a ranking and as a cut-off function and the final result is merged into a collapsed subgraph containing the triple paths (Figure 3). For the determination of *precision* we considered a correct answer a triple path containing the URI for the answer. For the example query used in this article, the triple path containing the answer *Barack Obama* $\rightarrow$ *spouse* $\rightarrow$ *Michelle Obama* $\rightarrow$ *alma mater* $\rightarrow$ *Princeton University* and *Harvard Law School* is the answer provided by the algorithm instead of just *Princeton University* and *Harvard Law School*. To determine both precision and recall, triple paths strongly semantically supporting answers are also considered. For the query *'Is Natalie Portman an actress?'*, the expected result is the set of nodes which highly supports the answer for this query, including the triples stating that Natalie Portman is an actress and that she acted on different movies (this criteria is used for both precision and recall). The QALD dataset contains aggregate queries which were included in the evaluation. However, since the post-processing phase does not cover aggregation operators we considered as a correct answer, triples supporting the aggregate answer.

Table 1 shows the relevance metrics collected for the evaluation for each query. For the Wikipedia Link Measure (WLM) semantic relatedness, the final approach achieved a *mean reciprocal rank* = **0.614**, *average precision* = **0.487**, *average recall* = **0.57** and **70%** of the queries were either completely or partially answered ($recall > 0$).

To analyze the results, queries with errors were classified according to 5 different categories, based on the components of the query approach. The first category, *PODS error*, contains errors which were determined by a difference between the PODS structure and the data representation which led the algorithm to an incorrect search path (Q35). In this case, the flexibility provided by semantic relatedness and spreading activation was unable to cope with this difference. The second error category, *Pivot Error*, includes errors in the determination of the correct pivot. This category includes queries with non-dereferenceable pivots (i.e. pivots which are based on literal resources) or errors in the pivot determination process (Q5, Q27, Q30, Q44). Some of the errors in the pivot determination process were related to overloading classes with complex types (e.g. for the query Q30 the associated pivot is a class *yago:HostCitiesOfTheSummerOlympicGames*).

Relatedness Error includes queries which were not addressed due to errors in the relatedness computation process, leading to an incorrect matching and the elimination of the correct answer (Q11, Q12). The fourth category, *Excessive Dereferenciation Timeout Error* covers queries which demanded a large number of dereferenciations to be answered (Q31, Q40). In the query Q40, the algorithm uses the entity *California* as a pivot and follows each associated *Organization* to find its associated type. This is the most challenging category to address, putting in evidence a limitation of the approach based on dynamic de-referenciation. The last categories cover small errors outside previous categories or combined errors in one query (Q32, Q39, Q43). The relatedness measure was able to cope with non-taxonomic variations between query and vocabulary terms, showing high average discrimination in the node selection process (average difference between the relatedness value of answer nodes and the relatedness mean is 2.81 $\sigma(r)$).

The evaluation on this paper concentrates on the use of a link-based semantic model/relatedness measure (WLM). The reader is directed to [36, 35] for results involving the use of a a term based distributional model (Explicit Semantic Analysis).

From the perspective of *query execution time* an experiment was run using an Intel Centrino 2 computer with 4 GB RAM. No parallelization or indexing mechanism outside entity search was implemented in the query mechanism. The average query execution time for the set of queries which were answered was **635s** with no caching and **203s** with active caches.

5. Discussion: The Semantic Search Pattern

5.1. Motivation

The problem of abstracting users from the vocabularies of knowledge bases (where these vocabularies can be database schemas or logical models) is a recurrent problem in computer science. The Treo approach can be abstracted into a semantic search pattern which can be applied to scenarios beyond natural language queries over Linked Data. In this section we isolate the basic principles and assumptions behind the Treo semantic search approach, generalizing Treo as a *semantic search pattern*.

5.2. Semantic Model

The query mechanism assumes a double-layered semantic model where the *labeled typed graph model layer* is complemented by a *distributional se-*

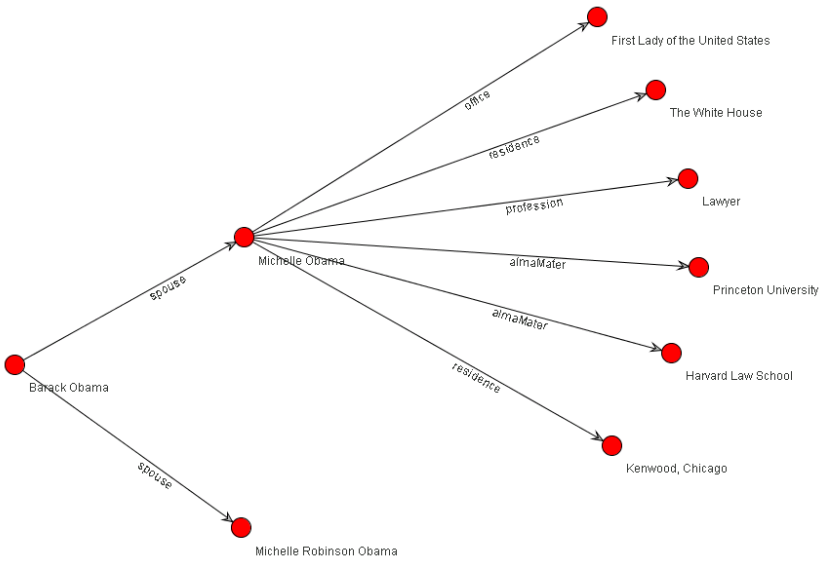

Figure 3: Output of the Treo mechanism for the example query.

mantic model layer.

5.2.1. Entity-Attribute-Value (EAV) data model layer

In the Entity-Attribute-Value (EAV) data layer the core semantic assumptions are:

i Assumption of a minimal *entity-relation-entity* or *entity-property-value* model, where *entity* represents instances (named entities) or classes (category names), and *relations* and *properties* represent an attribute associated with an entity.

ii Assumption of labels with explicit lexical representation, where labels for both instance and terminology-level elements can be resolved in the datasets, instead of relying on numerical identifiers' aliases or truncated labels.

5.2.2. Corpus-based semantic model layer

Consists in the construction of a large-scale commonsense corpus-based semantic layer which is used to semantically interpret the elements in the

EAV data layer. By allowing the computation of semantic relatedness over elements in the graph layer, this layer provides a quantitative semantic model. The semantic enrichment allows a flexible interpretation of the data elements according to a reference corpus. The corpus-based knowledge can be used to semantically match user queries to the data, providing a comprehensive semantic matching solution. Figure 4 depicts the two-layered semantic model of the approach.

Despite having as motivation the introduction of a lightweight semantic relatedness measure, the use of link structures to determine the semantic relatedness limits the transportability of the model to different domains. *Distributional semantic models*, such as ESA [16], can provide a more comprehensive and transportable semantic model, since they are not dependent on the existence and quality of a link-structure in the source corpus.

Queries over the graph layer can use distributional semantic relatedness measures as a ranking function. An analysis of the suitability of semantic relatedness as a semantic ranking function is provided in [38].

Freitas et al. [36],[35] generalizes the basic elements present in the Treo approach to build a distributional structured vector space model (VSM) named T-Space. The construction of an approach based on a vector space model provides a distributional structured semantic index for semantic search over data graphs, targeting queries with additional vocabulary independency. The construction of a structured index instead of using navigational queries also provides an improvement on the performance of the query mechanism [40].

5.3. The Core Pattern

The key characteristic and strength of the proposed approach lies on the fact that by using a simplified, comprehensive and automatically built semantic model, that supports a comprehensive semantic relatedness computation and enables a higher level of vocabulary independency between users' query terms and dataset terms. In this section we revisit the key elements of the approach, analyzing how each part of the query processing approach contributes to the creation of semantic search/matching pattern which can be applied over databases that can be transformed into an EAV model.

1. *Entity Recognition & Entity Search:* The performance of the semantic relatedness-based search step is dependent on the *entity recognition and search* step, which impacts the search algorithm in two ways. The

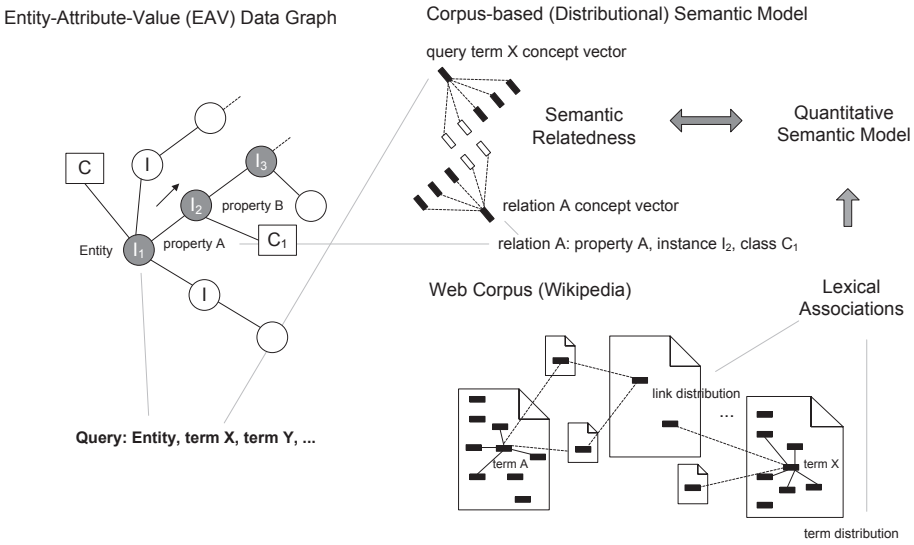

Figure 4: Depiction of the two layered semantic model behind the approach.

first level of impact is the prioritization of the resolution of the less ambiguous/polysemic part of the query (in most cases), since named entities in the query usually map to instances, which are less subject to vocabulary variation. The second type of impact is the drastic reduction of the search space by prioritizing the resolution of instances or classes.

2. *Corpus-based (Distributional) Semantic Relatedness:* A corpus-based semantic model provides an automatic solution to allow a comprehensive semantic matching in the context of entity relations. It is important to emphasize that the semantic model can (and in most of the cases should) be defined by resources outside the dataset, where the comprehensiveness of the model is given by the fact that an arbitrarily large corpora can be used in the creation of the semantic model.
3. *Determine a Threshold/Filter Unrelated Results:* A threshold should be defined to filter unrelated results. Threshold determination strategies can vary for each reference corpus and for each semantic relatedness measure being used.

6. Related Work

We approach four main categories of related work: *natural language interfaces & QA systems for the Semantic Web*, *search engines for structured data*, *spreading activation applied to information retrieval* and *query mechanisms with structural approximations*. For a complementary discussion of exiting approaches for querying Linked Data the reader is referred to [34].

6.1. Natural Language Interfaces & QA Systems

Different natural language query approaches for Semantic Web/Linked Data datasets have been proposed in the literature. Most of the existing query approaches using semantic approximations are based on WordNet and on the use of ontological/taxonomic information present in the datasets.

PowerAqua [18] is a question answering system focused on natural language questions over Semantic Web/Linked Data datasets. PowerAqua uses PowerMap to match query terms to vocabulary terms. According to Lopez et al. [22], *PowerMap is a hybrid matching algorithm comprising terminological and structural schema matching techniques with the assistance of large scale ontological or lexical resources*. PowerMap uses WordNet-based similarity approaches as a semantic approximation strategy.

NLP-Reduce [20] approaches the problem from the perspective of a lightweight natural language approach, where the natural language input query is not analyzed at the syntax level. The matching process between the query terms and the ontology terms present in NLP-Reduce is based on a WordNet expansion of synonymic terms in the ontology and on matching at the morphological level.

The matching process of another approach, Querix [21], is also based on the expansion of synonyms based on WordNet. Querix, however, uses syntax level analysis over the input natural language query, using this additional structure information to build the corresponding query skeleton of the query. Ginseng [19] follows a controlled vocabulary approach: the terms and the structure of the ontologies generate the lexicon and the grammar for the allowed queries in the system. Ginseng ontologies can be manually enriched with synonyms.

ORAKEL [41] is a natural language interface focusing on the portability problem across different domains. For this purpose, ORAKEL implements a lexicon engineering functionality, which allows the creation of explicit frame mappings. Instead of allowing automatic approximations, ORAKEL focuses

on a precise manually engineered model. Exploring user interaction techniques, FREyA [23] is a QA system that employs feedback and clarification dialogs to resolve ambiguities and improve the domain lexicon with users help. FREyA delegates part of the semantic matching and disambiguation process to users. User feedback enriches the semantic matching process by allowing manual entries of query-vocabulary mappings.

Compared to existing *natural language interfaces* approaches, *Treo* provides a query mechanism which explores a more comprehensive and automatic semantic approximation technique which can cope with the variability of the query-vocabulary matching on the heterogeneous environment of Linked Data on the Web. Additionally, its design supports querying dynamic and distributed Linked Data. The proposed approach also follows a different query strategy, by following sequences of dereferenciations and avoiding the construction of a SPARQL query and by focusing on a semantic best-effort/semantic approximate approach.

6.2. Search Engines

Information Retrieval based approaches for Linked Data range from search mechanisms focusing on providing an entity-centric search functionality to mechanisms aiming towards more flexible query mechanisms. In the entity-centric search space Sindice [12] is a search engine for Linked Data which is focused on entity-centric search and applies a variation of the traditional VSM and TF-IDF ranking scheme to index and rank entities. Sindice ranks entities according to the incidence of keywords associated with them. It uses a node-labeled tree model to represent the relationship between datasets, entities, attributes, and values. Queries vary from keyword-based to hybrid queries (mixing keyword search with structural elements in the database). Sindice does not apply a more principled technique for addressing the vocabulary problem.

Semplore [42] is a search engine for Linked Data which uses a hybrid query formalism, combining keyword search with structured queries. The Semplore approach consists in indexing entities of the Linked Data Web (instances, classes, properties) using the associated tokens and sub/superclasses as indexing terms. In addition to entity indexing, Semplore focuses on indexing relations using a position-based index approach to index relations and join triples. In the approach, relation names are indexed as terms, subjects are stored as documents and the objects of a relation are stored in the position lists. Based on the proposed index, Semplore reuses the IR engine's

merge-sort based Boolean query evaluation method and extends it to answer unary tree-shaped queries.

Dong & Halevy [43] propose an approach for indexing triples allowing queries that combine keywords and structure. The index structure is designed to cope with two query types: predicate queries and neighborhood keyword queries. The first type of queries covers conjunctions of predicates and associated keywords. Dong & Halevy propose four structured index types which are based on the introduction of additional structure information as concatenated terms in the inverted lists. Taxonomy terms are introduced in the index using the same strategy. Schema-level synonyms are handled using synonyms tables. Both approaches [42, 43] provide limited semantic matching strategies and are built upon minor variations over existing inverted index structures. By avoiding major changes over existing search paradigms, these approaches can inherit the implementation of optimized structures used in the construction of traditional indexes.

Compared to existing search approaches Treo focuses on providing a solution focused on the increase of vocabulary independency while keeping the query expressivity.

6.3. Spreading Activation

More recently the interest in spreading activation models had migrated to information retrieval, mostly in the context of associative retrieval. The idea behind associative retrieval is that relevant information could be retrieved by using associated information provided by an existing knowledge base, previous search activities or current query context. In the context of associative retrieval, associations among information items are usually represented as a graph and the search mechanism in this graph is based on spreading activation. According to Crestani [26], there was a decreasing interest in the application of associative retrieval techniques in the information retrieval area due to the difficulty in building associative graphs. The recent emergence of the Linked Data Web is likely to motivate new investigations in associative spreading activation techniques.

In Information Retrieval, the work of [27] is one of the first to approach associative search by using spreading activation. Shoval [28] focused on query expansion based on spreading activation over a semantic network based on a thesaurus. Later, Cohen & Kjeldsen [29] developed the GRANT system, which applied a constrained spreading activation approach for information retrieval. More recently, Rocha et al. [30] applied a spreading activation

approach to find related concepts in the domain ontology given an initial set of returned results from a traditional search. The ranking of each result returned from the traditional search is used to calibrate the spreading activation process. A key element in the spreading activation process is the use of a set of weights which are associated with the ontology relations by incorporating information from instances. The set of constraints used concept-type constraints (the activation does not propagate through nodes of a specific concept type), fan-out constraint (the activation is not propagated to highly connected nodes) and distance constraints. Katifori et al. [32] describes a spreading activation based approach over ontologies for task and activity oriented Personal Information Management. Schumacher et al. [33] describes a semantic search approach combining fact retrieval and document retrieval with spreading activation in the context of Semantic Desktops.

Differently from existing approaches, the spreading activation model proposed in the Treo approach uses a corpus-based semantic relatedness measure defined over query-data term pairs as an activation function. An additional difference is the entity search step included in the spreading activation process (as the entry point, and then as an entity reconciliation step).

6.4. Query Mechanisms with Structural Approximations

A set of approaches focuses on the problem of introducing approximations on SPARQL-like conjunctive queries for Semantic Web/Linked Data datasets. In this scope, Stuckenschmidt & van Harmelen [44] proposes a model for approximating conjunctive queries based on the relaxation and progressive restoration of query constraints. Similarly, Oren et al. [45] introduces an evolutionary approach over RDF data where query constraints are relaxed and further restored and refined by the application of an evolutionary algorithm. Hurtado et al. [46] introduce a logic-based approximate query mechanism where RDFS constraints are used to logically relax the schema-level constraints.

Existing approaches under this category mainly apply semantic approximation strategies based on ontological information or by relaxing structural constraints in the query. Comparatively Treo uses external corpus knowledge to increase vocabulary-independency while keeping query expressivity.

7. Conclusion & Future Work

This paper proposes a natural language query mechanism for Linked Data focusing on improving the vocabulary independency and on addressing the trade-off between expressivity and usability for queries over Linked Data. To address this problem, a novel combination for querying Linked Data is proposed, based on entity search, spreading activation and a Wikipedia-based semantic relatedness. The approach was implemented in the *Treo* prototype and was evaluated with an extended version of the QALD query dataset containing 50 natural language queries over the DBpedia dataset, achieving an overall *mean reciprocal rank* of 0.614, *average precision* of 0.487 and *average recall* of 0.572. Additionally, a set of short-term addressable limitations of the approach were identified. The result shows that corpus-based semantic relatedness can provide a comprehensive semantic matching mechanism to support higher levels of vocabulary independency. The proposed approach was designed for querying live distributed Linked Data but it can be extended to databases which can be mapped to an Entity-Attribute-Value (EAV) data model. Directions for future investigations include addressing the set of limitations identified during the experiments, the incorporation of a more sophisticated post-processing mechanism and the investigation of performance optimizations for the approach.

Acknowledgments.

The work presented in this paper has been funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

[1] T. Berners-Lee, Linked Data Design Issues, (2009) http://www.w3.org/DesignIssues/LinkedData.html.

[2] E. Kaufmann, A. Bernstein, Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases, J. Web Semantics: Science, Services and Agents on the World Wide Web 8 (2010) 393-377.

[3] P. Nadkarni, L. Marenco, R. Chen, E. Skoufos, G. Shepherd, P. Miller, Organization of heterogeneous scientific data using the EAV/CR representation, J Am Med Inform Assoc. (1999) 478-93.

[4] Resource Description Framework (RDF) Model and Syntax Specification, (2013) http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

[5] M. Marneffe, B. MacCartney, C. D. Manning, Generating Typed Dependency Parses from Phrase Structure Parses. In LREC 2006, (2006).

[6] J.R. Finkel, T. Grenager, C.D. Manning, Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling, In: Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), 2005, pp. 363-370.

[7] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Procesing and Management (1988) pp. 513-523.

[8] F. Sang, F. Meulder, Introduction to the CoNLL-2003 shared task: language-independent named entity recognition, In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL (2003).

[9] K. Toutanova, D. Klein, C.D. Manning, Y. Singer, Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network, In: Proceedings of HLT-NAACL 2003, 2003, pp. 252-259.

[10] P. Resnik, Using Information Content to Evaluate Semantic Similarity in a Taxonomy, In: Proceedings of the International Joint Conference for Artificial Intelligence (IJCAI-95), 1995, pp. 448-453.

[11] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, A. Soroa, A study on similarity and relatedness using distributional and wordnet-based approaches. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009, pp.19-27.

[12] R. Delbru, N. Toupikov, M. Catasta, G. Tummarello, S. Decker, A Node Indexing Scheme for Web Entity Retrieval, In: Proceedings of the 7th Extended Semantic Web Conference (ESWC), 2010.

[13] C. Bizer, J. Lehmann, G. Kobilarov, S. R. Auer, C. Becker, R. Cyganiak, S. Hellmann, DBpedia - A crystallization point for the Web of Data, J. Web Semantics: Science, Services and Agents on the World Wide Web (2009).

[14] 1st Workshop on Question Answering over Linked Data (QALD-1), 2011, http://www.sc.cit-ec.uni-bielefeld.de/qald-1.

[15] Evaluation Dataset, 2011, http://treo.deri.ie/results/dke2013.htm.

[16] E. Gabrilovich, S. Markovitch, Computing semantic relatedness using Wikipedia-based explicit semantic analysis, In: Proceedings of the International Joint Conference On Artificial Intelligence, 2007.

[17] D. Milne, I.H. Witten, An effective, low-cost measure of semantic relatedness obtained from Wikipedia links, In: Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08), 2008, Chicago, I.L..

[18] V. Lopez, E. Motta, V. Uren, PowerAqua: Fishing the Semantic Web, In: Proc 3rd European Semantic Web Conference ESWC, Vol. 4011. Springer, 2004, pp. 393-410.

[19] A. Bernstein, E. Kaufmann, C. Kaiser, C. Kiefer, Ginseng A Guided Input Natural Language Search Engine for Querying Ontologies,In: Jena User Conference, 2006.

[20] E. Kaufmann, A. Bernstein, L. Fischer, NLP-Reduce: A naive but Domain-independent Natural Language Interface for Querying Ontologies,In: 4th European Semantic Web Conference ESWC, 2007, pp. 1-2.

[21] E. Kaufmann, A. Bernstein, R. Zumstein, Querix: A Natural Language Interface to Query Ontologies Based on Clarification Dialogs, In: 5th International Semantic Web Conference (ISWC), Springer, 2006, pp. 980-981.

[22] V. Lopez, M. Sabou, E. Motta, PowerMap: Mapping the Real Semantic Web on the Fly, In: International Semantic Web Conference, Vol. 4273, Springer, 2006, pp. 5-9.

[23] D. Damljanovic, M. Agatonovic, H. Cunningham, FREyA: An Interactive Way of Querying Linked Data Using Natural Language, In: Proc. 1st Workshop on Question Answering over Linked Data (QALD-1), Collocated with the 8th Extended Semantic Web Conf. (ESWC 11), 2011.

[24] M.R. Quillian, Semantic Memory, Semantic Information Processing, MIT Press, 1968, pp. 227-270.

[25] A. M. Collins, E. F. Loftus, A spreading-activation theory of semantic processing, (1975) Psychological Review.

[26] F. Crestani, Application of Spreading Activation Techniques in Information Retrieval, Artificial Intelligence Review, 11 (6) (1997) 453-453.

[27] S.E., Preece, A spreading activation network model for information retrieval, University of Illinois at Urbana-Champaign, 1981.

[28] P. Shoval, Expert/consultation system for a retrieval data-base with semantic network of concepts, ACM SIGIR Forum, 16 (1) (1981) 145-149.

[29] P. Cohen, Information retrieval by constrained spreading activation in semantic networks, Information Processing & Management, 23 (4) (1987) 255-268.

[30] C. Rocha, D. Schwabe, M.P. Aragao, A hybrid approach for searching in the semantic web, In: International World Wide Web Conference, 2004, pp. 374-374.

[31] R. Cilibrasi, P. M. B. Vitnyi: The Google Similarity Distance, In IEEE Trans. Knowl. Data Eng, (2007) 19(3): 370-383.

[32] A. Katifori, C. Vassilakis, A. Dix, Ontologies and the brain: Using spreading activation through ontologies to support personal interaction, Cognitive Systems Research, 11 (1) (2010) 25-41.

[33] K. Schumacher, M. Sintek, L. Sauermann, Combining fact and document retrieval with spreading activation for semantic desktop search, In: Proceedings of the 5th European semantic web conference (ESWC'08), 2008, pp. 569-583.

[34] A. Freitas, E. Curry, J.G. Oliveira, S. O'Riain, Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches and Trends, IEEE Internet Computing, Special Issue on Internet-Scale Data, (2012) 24-33.

[35] A. Freitas, E. Curry, J.G. Oliveira, S. O'Riain, A Distributional Structured Semantic Space for Querying RDF Graph Data, International Journal of Semantic Computing (IJSC), (2012).

[36] A. Freitas, J. G. Oliveira, E. Curry, S. O'Riain, A Multidimensional Semantic Space for Data Model Independent Queries over RDF Data. In Proceedings of the 5th International Conference on Semantic Computing (ICSC), (2011).

[37] A. Freitas, J.G. Oliveira, S. O'Riain, E. Curry, J.C.P. da Silva, Querying Linked Data using Semantic Relatedness: A Vocabulary Independent Approach, In: Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB) (2011).

[38] A. Freitas, E. Curry, S. O'Riain, A Distributional Approach for Terminological Semantic Search on the Linked Data Web, In: Proceedings of the 27th ACM Symposium On Applied Computing (SAC), Semantic Web and Applications (SWA), (2012).

[39] A. Freitas, J.G. Oliveira, S. O'Riain, E. Curry, J.C.P. da Silva, Treo: Combining Entity-Search, Spreading Activation and Semantic Relatedness for Querying Linked Data, In: 1st Workshop on Question Answering over Linked Data (QALD-1) Workshop at 8th Extended Semantic Web Conference (ESWC), (2011).

[40] A. Freitas, F de Faria, E. Curry, S. O'Riain, Answering Natural Language Queries over Linked Data Graphs: A Distributional Semantics Approach, In: Proceedings of the 36th Annual ACM SIGIR Conference, (2013).

[41] P. Cimiano, P. Haase, J. Heizmann, M. Mantel, R. Studer, Towards portable natural language interfaces to knowledge bases: The Case of the ORAKEL system, Data Knowledge Engineering (DKE), 65 (2) (2008) 325-354.

[42] H. Wang, Q. Liu, T. Penin, L. Fu, L. Zhang, T. Tran, Y. Yu, Y. Pan, Semplore: A scalable IR approach to search the Web of Data, Web Semantics: Science, Services and Agents on the World Wide Web, 7 (3) 2009 177-188.

[43] X. Dong, A. Halevy, Indexing dataspaces, In: Proc. of the 2007 ACM SIGMOD international conference on Management of Data, 2007.

[44] H. Stuckenschmidt, F.van Harmelen, Approximating Terminological Queries, In: Proceedings of the 5th International Conference on Flexible Query Answering Systems (FQAS '02), 2002, pp. 329-343.

[45] E. Oren, C. Guaret, S. Schlobach, Anytime Query Answering in RDF through Evolutionary Algorithms, In: Proceeding of the 7th International Conference on The Semantic Web (ISWC '08), 2008, pp.98-113.

[46] C.A. Hurtado, A. Poulovassilis, P.T. Wood, Ranking Approximate Answers to Semantic Web Queries, In: Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications (ESWC 2009), 2009, pp. 263-263.

[47] K.J.Kochut, M. Janik, SPARQLeR: Extended SPARQL for semantic association discovery, In: Proceedings of the 4th European conference on The Semantic Web: Research and Applications (ESWC 2007), 2007, pp. 145-159.

[48] C. Kiefer, A. Bernstein, M. Stocker, The fundamentals of iSPARQL: A virtual triple approach for similarity-based semantic web tasks, In: Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference (ISWC'07/ASWC'07), 2007, pp.295-309.

[49] P.D. Turney, P. Pantel, From Frequency to Meaning: Vector Space Models of Semantics, Journal of Artificial Intelligence Research 3 (2010) 141-188.

[50] M. Sahlgren, The Distributional Hypothesis. From context to meaning: Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), Rivista di Linguistica, 20 (1) 2008.

#	query	rr	precision	recall
1	From which university the wife of Barack Obama graduate?	0.25	0.333	0.5
2	Give me all actors starring in Batman Begins.	1	1	1
3	Give me all albums of Metallica.	1	0.611	0.611
4	Give me all European Capitals!	1	1	1
5	Give me all female German chancellors!	0	0	0
6	Give me all films produced by Hal Roach?	1	1	1
7	Give me all films with Tom Cruise.	1	0.865	1
8	Give me all soccer clubs in the Premier League.	1	0.956	1
9	How many 747 were built?	0.5	0.667	1
10	How many films did Leonardo DiCaprio star in?	0.5	0.733	0.956
11	In which films did Julia Roberts as well as Richard Gere play?	0	0	0
12	In which programming language is GIMP written?	0	0	0
13	Is Albert Einstein from Germany?	0.5	0.5	1
14	Is Christian Bale starring in Batman Begins?	0.125	0.071	1
15	Is Einstein a PHD?	1	1	1
16	Is Natalie Portman an actress?	1	0.818	0.273
17	Is there a video game called Battle Chess?	1	1	0.023
18	List all episodes of the first season of the HBO television series The Sopranos!	0.333	0.090	1
19	Name the presidents of Russia.	1	1	0.167
20	Since when is DBpedia online?	1	0.667	1
21	What is the band of Lennon and McCartney?	0	0	0
22	What is the capital of Turkey?	1	1	1
23	What is the official website of Tom Hanks?	1	0.333	1
24	What languages are spoken in Estonia?	1	1	0.875
25	Which actors were born in Germany?	1	0.033	0.017
26	Which American presidents were actors?	1	0.048	1
27	Which birds are there in the United States?	0	0	0
28	Which books did Barack Obama publish?	1	0.5	1
29	Which books were written by Danielle Steel?	1	1	1
30	Which capitals in Europe were host cities of the summer olympic games?	0	0	0
31	Which companies are located in California, USA?	0	0	0
32	Which companies work in the health area as well as in the insurances area?	0	0	0
33	Which country does the Airedale Terrier come from?	1	1	0.25
34	Which genre does the website DBpedia belong to?	1	0.333	1
35	Which music albums contain the song Last Christmas?	0	0	0
36	Which organizations were founded in 1950?	0	0	0
37	Which people have as their given name Jimmy?	0	0	0
38	Which people were born in Heraklion?	1	1	1
39	Which presidents were born in 1945?	0	0	0
40	Which software has been developed by organizations in California?	0	0	0
41	Who created English Wikipedia?	1	1	1
42	Who developed the video game World Warcraft?	1	0.8	1
43	Who has been the 5th president of the United states?	0	0	0
44	Who is called Dana?	0	0	0
45	Who is the wife of Barack Obama?	1	1	1
46	Who owns Aldi?	0.5	1	0.667
47	Who produced films starring Natalie Portman?	1	0.3	0.810
48	Who was the wife of President Lincoln?	1	1	1
49	Who was Tom Hanks married to ?	1	0.214	1
50	Who wrote the book The pillars of the Earth?	1	0.5	0.5

Table 1: Relevance results for the query set with the associated reciprocal rank, precision and recall using Wikipedia Link Measure (WLM).

Error Type	% of Queries
PODS Error	2%
Pivot Error	10%
Relatedness Error	4%
Excessive Dereferenciation Timeout Error	6%
Combined Error	8%

Table 2: Error types and distribution