A Multidimensional Semantic Space for Data Model Independent Queries over RDF Data

André Freitas, João Gabriel Oliveira, Edward Curry, Seán O'Riain
Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway

***Abstract*—The vision of creating a Linked Data Web brings together the challenge of allowing queries across highly heterogeneous and distributed datasets. In order to query Linked Data on the Web today, end-users need to be aware of which datasets potentially contain the data and also which data model describes these datasets. The process of allowing users to expressively query relationships in RDF while abstracting them from the underlying data model represents a fundamental problem for Web-scale Linked Data consumption. This article introduces a multidimensional semantic space model which enables data model independent natural language queries over RDF data. The center of the approach relies on the use of a distributional semantic model to address the level of semantic interpretation demanded to build the data model independent approach. The final multidimensional semantic space proved to be flexible and precise under real-world query conditions achieving *mean reciprocal rank* = 0.516, *avg. precision* = 0.482 and *avg. recall* = 0.491.**

I. Introduction

The vision behind the construction of a Linked Data Web [1] where it is possible to consume, publish and reuse data in a new granularity and scale steps into a fundamental problem in the semantic computing space. In order to query highly heterogeneous and distributed data at Web-scale, it is necessary to reformulate the current paradigm on which users interact with datasets, which is highly dependent on an a priori understanding of the data model behind the datasets. In order to query datasets today, users need to articulate their information needs in a query containing explicit representations of the relationships present in the dataset data model (i.e. the dataset 'vocabulary'). This query paradigm, deeply attached to the traditional perspective of structured queries over databases, does not suit the heterogeneity, distributiveness and the scale of the Web, where it is impractical for data consumers to have a previous understanding of the structure and location of available datasets.

Behind this problem resides a fundamental limitation of information systems today to provide a semantic interpretation approach that could bridge the semantic gap between users' intentions and the 'vocabulary' used to describe systems' objects and actions. This *semantic gap*, defined by Furnas et al. [6] as the *vocabulary problem in human-system communication*, is associated to the dependency on human language (and its intrinsic variability) in the construction of systems and information artifacts. At Web-scale, the vocabulary problem for querying existing Web data represents a fundamental barrier which ultimately limits the utility of Linked Data for data consumers.

For many years the level of semantic interpretation needed to address the vocabulary problem was associated with deep problems in the Artificial Intelligence space, such as knowledge representation and commonsense reasoning. However, the solution to these problems also depends upon some prior level of semantic interpretation, creating a self-referential dependency. More recently, promising results related to research on *distributional semantics* [9][7] are showing a possible direction to solve this conundrum by bootstrapping on the knowledge present on large volumes of Web corpora.

This work proposes a multidimensional semantic space focused on providing a data model independent query approach over RDF data. The multidimensional semantic space introduced in this paper builds upon the *Treo* query mechanism, introduced in [8]. The center of the approach relies on the use of distributional semantics and on a hybrid search strategy (entity-centric search and spreading activation search) to build the semantic space. The proposed approach generalizes the previous Treo query mechanism, introducing a new entity search strategy and a multidimensional index structure based on distributional semantics. The final semantic space, named *T-Space* (tau space), proved to be flexible and precise under real-world query conditions.

The construction of a semantic space based on the principles behind Treo (discussed in section 3) defines a search/index generalization which can be applied into different problem spaces, where data is or could be represented as labeled graph data, including graph databases and semantic-level representations of unstructured text. The description of the features present in the proposed semantic space, together with the analysis of emerging related approaches, provides the opportunity for the reader to understand fundamental trends in semantic search.

The paper is organized as follows: section 2 introduces the central concept of distributional semantics and describes one specific distributional approach, Explicit Semantic Analysis (ESA), describing how the distributional model is used to compute a semantic relatedness measure; section 3 covers the basic principles behind the semantic search approach; section 4 describes the construction of the multidimensional semantic space; section 5 covers the evaluation of the approach; section 6 describes the related work and finally section 7 provides the conclusion and future work.

II. Distributional Semantics

A. Motivation

Distributional semantics is built upon the assumption that the context surrounding a given word in a text provides important information about its meaning [9]. Another rephrasing of the *distributional hypothesis* is that words that occur in similar contexts tend to have similar meanings [9]. 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. The fact that distributional semantic models are naturally represented by Vector Space Models, where the meaning of a word is represented by a weighted concept vector, brings an additional strength to these approaches.

However, the proper use of the model of meaning provided by distributional semantics implies understanding its characteristics and limitations. As Sahlgren [7] notes, the distributional view on meaning is non-referential (does not refer to extra-linguistic representations of the object related to the word), being inherently differential: the differences of meaning are mediated by differences of distribution. As a consequence, distributional semantic models allows the quantification of the amount of difference in meaning between linguistic entities. This differential analysis can determine the semantic relatedness between words [7]. Therefore, the applications of the meaning defined by distributional semantics should be constrained to 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 [5]. This work focuses on the use of distributional semantics in the computation of semantic relatedness measures as a key element to address the level of semantic flexibility necessary for the provision of data model independent queries over RDF data. In addition, the differential nature of distributional semantics also fits into a *best-effort* query strategy which is the focus of this work.

B. Semantic Relatedness

The concept of *semantic relatedness* is described [10] as a generalization of *semantic similarity*, where semantic similarity is associated with taxonomic relations between concepts (e.g. *car* and *airplane* share *vehicle* as a common taxonomic ancestor) and semantic relatedness covers a broader range of semantic relations (e.g. *car* and *driver*). Since the problem of matching natural language terms to concepts present in datasets can easily cross taxonomic boundaries, the generic concept of semantic relatedness is more suitable to the task of semantic matching for queries over the RDF data.

Until recently WordNet, an interlinked lexical database, was the main resource used in the computation of similarity and relatedness measures. The limitations of the representation present in WordNet, including the lack of a rich representation of non-taxonomic relations, fundamental for the computation of relatedness measures and the limitation in the number of modeled concepts, motivated the construction of approaches based on distributional semantics. Additionally, the availability of large amounts of unstructured text on the Web and, in particular, the availability of Wikipedia, a comprehensive and high-quality knowledge base, motivated the creation of relatedness measures based on Web resources, focusing on addressing the limitations of WordNet-based approaches, trading structure for volume of commonsense knowledge. Comparative evaluations between WordNet-based and distributional approaches for the computation of relatedness measures have shown the strength of the distributional model, reaching a high correlation level with human assessments.

C. Explicit Semantic Analysis

The distributional approach used in this work is defined by the Explicit Semantic Analysis (ESA) semantic space [5], which is built using Wikipedia corpora. The ESA space provides a distributional model which can be used to compute an explicit semantic interpretation of a term as a set of weighted concepts. In the case of ESA, the set of returned weighted concept vectors associated with the term is represented by titles of Wikipedia articles. A *universal ESA space* is created by building a vector space of Wikipedia articles using the traditional 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 a list of ranked articles titles, which define a concept vector associated with the query terms (where each vector component receives a relevance weight). The approach also allows the interpretation of text fragments, where the final concept is the centroid of the vectors representing the set of individual terms. This procedure allows the approach to partially perform word sense disambiguation [5]. The ESA semantic relatedness measure between two terms or text fragments is calculated by comparing the concept vectors representing the interpretation of the two terms or text fragments. The use of the ESA distributional approach in the construction of the proposed semantic space is covered in the next two sections.

III. Semantic Search Approach

The multidimensional semantic space introduced in this paper generalized and improves the approach used in the *Treo* query mechanism [8]. The construction of a semantic space, based on the principles behind Treo, defines a search/index generalization which can be applied into different problem spaces, where data is represented as labeled graph data, such as RDF/Linked Data, graph databases and semantic-level representation of unstructured text. This section first introduces the strategies and principles behind the Treo search approach, followed by an instantiation of the search model for an exemplar natural language query.

A. Strategies behind the Semantic Search

In order to build the data model independent query mechanism, five main strategies are employed:

Best-effort search model: The proposed approach targets a best-effort solution for queries over Linked datasets. Instead of expecting the query mechanism to return exact results as in structured SPARQL queries, it returns an approximate and ranked answer set which can be later cognitively assessed by human users. An explicit requirement in the construction of the best-effort approach is the conciseness of the answer set, where a more selective cut-off function is defined, instead of an exhaustive ranked list of results (as in most document search engines).

Use of semantic relatedness measures to match query terms to dataset terms: Semantic relatedness and similarity measures allow the computation of a measure of semantic proximity between two natural language terms. The measure allows query terms to be semantically matched to dataset terms by their level of semantic relatedness. While semantic similarity measures are constrained to the detection of a reduced class of semantic relations and are mostly restricted to compute the similarity between terms which are nouns, semantic relatedness measures are generalized to any kind of semantic relation, being more robust to the heterogeneity of the vocabulary problem at Web-scale.

Use of a distributional semantic relatedness measure built from Wikipedia: Distributional relatedness measures are built using comprehensive knowledge bases on the Web, by taking into account the distributional statistics of a term, i.e. the co-occurrence of terms in its surrounding context. The use of comprehensive knowledge Web sources allows the creation of a high coverage distributional semantic model.

Query dependency (structure and ordering) as a fundamental carrier of semantic information: The approach builds upon the concept of using *partial ordered dependency structures* (PODS) as the query input. PODS are an intermediate form between a natural language query and a structured graph pattern that is built upon the concept of dependency grammars [11]. Dependency grammar is a syntactic formalism that has the property of abstracting over the surface word order, mirroring semantic relationships and creating an intermediate layer between syntax and semantics [11]. The idea behind the PODS query representation is to maximize the matching probability between the natural language query and triple-like (subject, predicate and object) structure present in the dataset. Additional details are covered in [8].

Two phase search process combining entity search with spreading activation search: The search process over the graph data is split into two phases. The first phase consists in searching in the datasets for instances or classes (*entity search*) which are expressed as terms in the query, defining *pivot entities* as entry points in the datasets for the semantic matching approach. The process is followed by a semantic matching phase based on a *spreading activation search based on semantic relatedness*, which matches the remaining query terms. This separation allows the search space to be pruned in the first search step by the less ambiguous part of the query (the key entity in the query), followed by a search process over the properties of the pivot entities (attributes and relations).

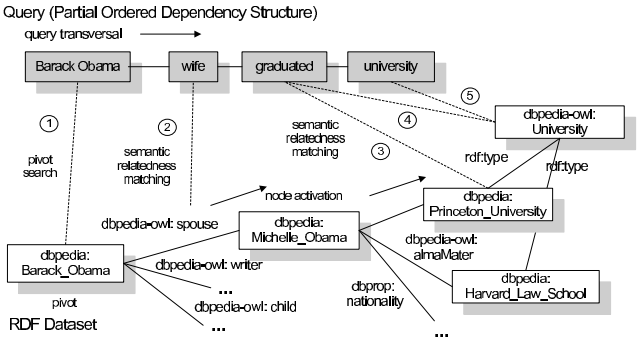

Fig. 1. The semantic relatedness based spreading activation search model for the example query

The next section details how the strategies described above are implemented in a search procedure over RDF data.

B. Semantic Search Steps

The semantic search approach assumes that the user natural language query is pre-processed into a partial ordered dependency structure (PODS), a format which is closer from the triple-like (subject, predicate and object) structure of RDF. The construction of the PODS demands an *entity recognition step*, where key entities in the query are determined by the application of named entity recognition algorithms, complemented by search over the lexicon defined by dataset instances and classes labels, followed by a *query parsing step*, where the partial ordered dependency structure is built by taking into account the dependency structure of the query, the position of the key entity and a set of transformation rules. An example of PODS for the example query *'From which university did the wife of Barack Obama graduate?'* is shown as gray nodes in figure 1. For additional details on the entity recognition and the query parsing steps the reader is directed to [8].

The semantic search process takes as input the PODS representation of the query and consists of two steps:

Entity Search and Pivot Entity Determination: The key entities represented in the PODS (which were detected in the entity recognition step) are sent to an entity-centric search engine which maps the natural language terms for the key entities into dataset entities (represented by URIs). In the entity-centric search engine, instances are indexed using TF/IDF for the terms in the labels, while classes are indexed using the ESA semantic space for its terms (see section 4). The URIs define the pivot entities in the datasets, which are the entry points for the semantic search process. In the example query, the term *Barack Obama* is mapped to the URI *http://dbpedia.org/resource/Barack_Obama* in the dataset.

Semantic Matching (Spreading Activation using Semantic Relatedness): Taking as inputs the pivot entities URIs and the PODS query representation, the semantic matching process

starts by fetching all the relations associated with the top ranked pivot entity. In the context of this work, the semantics of a relation associated with an entity is defined by taking into account the aggregation of the predicate, associated range types and object labels. Starting from the pivot node, the labels of each relation associated with the pivot node have their semantic relatedness measured against the next term in the PODS representation of the query. For the example entity *Barack Obama*, the next query term *wife* is compared against all predicates/range types/objects associated with each predicate (e.g. *spouse*, *child*, *religion*, etc). The relations with the highest relatedness measures define the neighboring nodes which will be explored in the search process. The search algorithm then navigates to the nodes with high relatedness values (in the example, *Michelle Obama*), where the same process happens for the next query term (*graduate*). The search process continues until the end of the query is reached, working as a spreading activation search over the RDF graph, where the activation function (i.e. the threshold to determine the further node exploration process) is defined by a semantic relatedness measure.

The spreading activation algorithm returns a set of *triple paths*, which are a connected set of triples defined by the spreading activation search path, starting from the pivot entities over the RDF graph. The triple paths are merged into a final graph and a visualization is generated for the end user (figure 4). The next section uses the elements of the described approach to build a multidimensional semantic space.

IV. Multidimensional Semantic Space

A. Introduction

The main elements of the approach described in the previous section are used in the construction of a multidimensional semantic space, named here a *T-Space* (tau-space). The final semantic space is targeted towards providing a vocabulary/data model independent representation of RDF datasets. This work separates the discussion between the definition of the semantic space model and the actual implementation of its corresponding index. Despite the implementation of an experimental index for evaluation purposes, this article concentrates on the definition and description of the semantic space model.

The multidimensional semantic space is composed by an entity-centric space where *instances* define vectors over this space using the TF/IDF weighting scheme and where *classes* are defined over an ESA entity space (the construction of the ESA space is detailed further). The construction strategy for the instance entity-centric space benefits a more rigid and less semantically flexible entity search for instances, where the expected search behavior is closer to a string similarity matching scenario. The rationale behind this indexing approach is that instances in RDF datasets usually represent named entities (e.g. names for people and places) and are less constrained by lexico-semantic variability issues in their dataset representation.

Classes demand a different entity indexing strategy and since they represent categories (e.g. yago:UnitedStatesSenators) they are more bound to a variability level in their representation (e.g. the class yago:UnitedStatesSenators could have been expressed as yago:AmericanSenators). In order to cope with this variability, the entity space for classes should have the property of semantically matching terms in the user queries to dataset terms. In the case of the class name *United States Senators* it is necessary to provide a semantic match with equivalent or related terms such as *American Senators* or *American Politicians*. The desired search behavior for a query in this space is to return a ranked list of semantically related class terms, where the matching is done by providing a semantic space structure which allows search based on a semantic interpretation of query and dataset terms. The key element in the construction of the semantic interpretation model is the use of distributional semantics to represent query and dataset terms. Since the desired behavior for the semantic interpretation is of a semantic relatedness ranking approach, the use of distributional semantics is aligned with the differential meaning assumption (section 2.2). Despite exemplifying with classes, the same distributional approach can be used for indexing *entity relations* which, in the scope of this work, consists of both terminological-level (properties, ranges, and associated types) and instance-level object data present in the set of relations associated with an entity.

B. Building the Semantic Space

The semantic space construction for terms present in the classes and entity relations starts by first creating a *universal Explicit Semantic Analysis (ESA) space* (step 1, figure 2). A *universal ESA space* is created by indexing Wikipedia articles using the traditional TF/IDF vector space approach. Once the space is built, a keyword query over the ESA space returns a set of ranked articles titles which define a concept vector associated with query terms (where each component of this vector is a Wikipedia article title receiving a relevance score) (figure 2). The concept vector is called the *semantic interpretation* of the term and can be used as its semantic representation.

In order to build the class and entity relation spaces, the ESA universal space is used to generate an *ESA semantic vector space* which is used to index the dataset terms (classes and relations). The construction of the ESA semantic vector space is done by taking the concept vectors (containing the TF/IDF scores associated with each vector component) of each dataset term and by creating a vector space where each defined point in the vector space represents a term being semantically indexed. This space has the desired property of returning a list of semantically related terms for a query (ordered from the most to the less semantically related). This procedure is described in the step 3 of figure 2 for the construction of the class entity space. The *final entity space* is a space with a double coordinate basis where instances are defined using a *TF/IDF term basis* and classes with an *ESA concept basis* (steps 2, 3, figure 2).

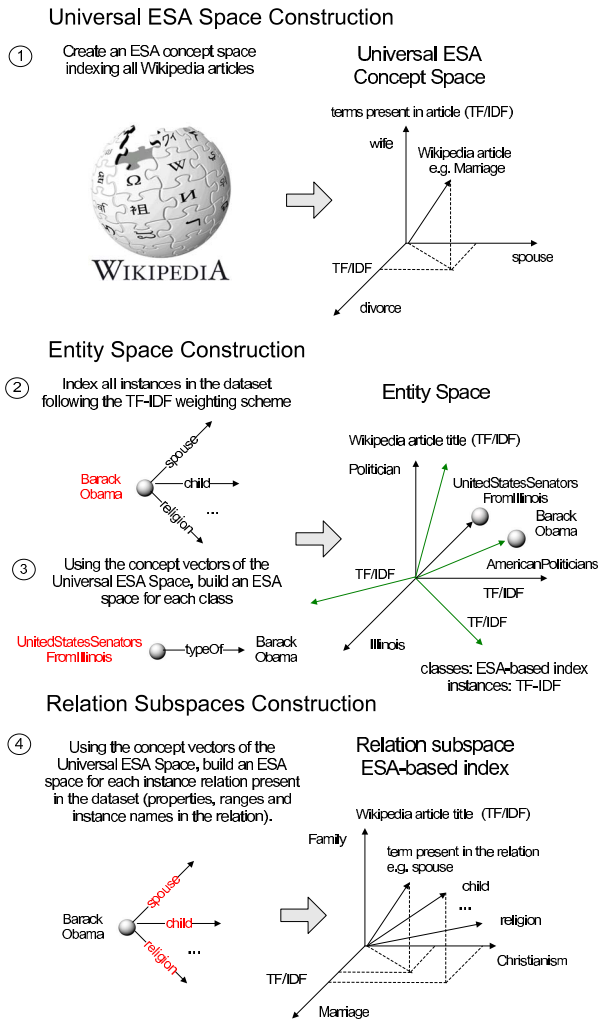

Fig. 2. Construction of the base spaces

C. Semantic Space Structure

Once the entity space is built, it is possible to assign for each point defined in the entity vector space, a linear vector space for representing the relations associated with an entity. For the example entity *Barack Obama*, a relation is defined by the set of properties, associated types and objects which are associated with this entity in its RDF description. The procedure is similar to the construction of the class space, where the terms present in the relations (properties, range types and objects) are used to create a linear vector space associated with the entity. One property of *entity relation spaces* is the fact that each entity has an independent number of dimensions, being scoped to the number of relations specific for each entity (figure 3). The property of associating a vector space for each entity reduces the associated query execution time by reducing the dimensionality of each relation space.

The final *multidimensional T-Space* has the topological structure of two linear vector spaces ($E_I^{TF/IDF}$ and E_C^{ESA}) defined for the *individuals* and *entities* respectively. Each entity defined over these spaces has an associated *vector bundle*

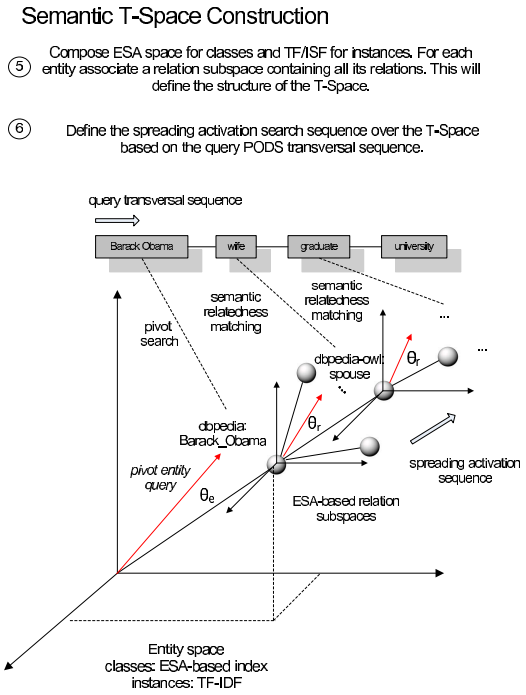

Fig. 3. Construction of the final multidimensional semantic vector space

$R^{ESA}(E)$ which is the space of relations. The spaces of relations, however, have a variable number of dimensions and a different coordinate basis. Each vector defined in $R^{ESA}(E)$ is associated with a specific object resource and has an associated tensor mapping into an instance on the $E_I^{TF/IDF}$ space. This mapping reflects the graph structure in the RDF.

With the final multidimensional semantic vector space built, it is necessary to define the search procedure over the space. The query input is a partial ordered dependency structure (PODS) with the key query entity defined. The key query entity is the first term to be searched on the entity space (it is searched in the instances entity space in case it is a named entity; otherwise it is searched over the class space). The return of the query is a set of activated URIs mapping to entities in the space (e.g. dbpedia:Barack_Obama is one example). The next term of the PODS structure sequence is taken ('wife') and it is used to query each relation subspace associated with the set of entities. The set of relations with high relatedness scores is used to activate other entities in the space (e.g. dbpedia:Michelle_Obama). The same process follows for the activated entities until the end of the query is reached. The search process returns a set of ranked triple paths where the rank score of each triple path is defined by the average of the relatedness measure. Figure 4 contains a set of merged triple paths for the example query.

In the node selection process, nodes above a relatedness score threshold determine the entities which will be activated. The activation function is given by an adaptive discriminative relatedness threshold which is defined based on the set of returned relatedness scores. The adaptive threshold has the objective of selecting the relatedness scores with higher dis-

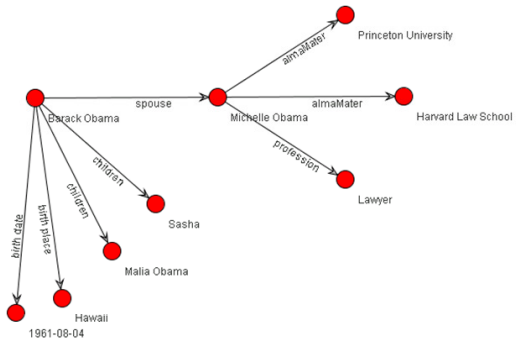

Fig. 4. Screenshot of the returned graph for the implemented prototype for the example query.

crimination. Additional details are available in [8].

The best-effort nature of the approach allows addressing *semantic tractability* problems by returning an answer set which users can quickly assess to determine the final answer. As an example consider the query '*Is Albert Einstein a PhD?*'. In the current version of DBPedia there is no explicit statement containing this information. However, the proposed approach returns an answer set containing the relation '*Albert Einstein doctoralAdvisor Alfred Kleiner*' from which users can derive the final answer.

The final multidimensional semantic space unifies into a single approach important features which are emerging as trends in the construction of new semantic and vector space models. The first feature is related to the adoption of a *distributional model of meaning* in the process of building the semantic representation of the information. The second feature is the use of *third-party available Web corpora* in the construction of the distributional model, instead of just relying on the indexed information to build the distributional semantic base. The third important feature is the inclusion of a *compositional element in the definition of the data semantics*, where the ordering given by the RDF data structure and by the PODS are used to define the semantic interpretation of the query, together with the individual distributional meaning of each word. Finally, another important feature which is a contribution of this work is the inclusion of the additional structure given by the RDF graph data into a multidimensional semantic space. This additional structure, embedded in the semantic structure of the space, defines a generic semantic space model with both distributional and compositional semantic features.

V. Evaluation & Discussion

An experimental evaluation of the proposed multidimensional semantic space was implemented to evaluate the answer quality of the approach using 50 natural language queries over DBPedia [2], defined in the QALD evaluation query set [3]. Since the final approach returns answers as triple-paths and considering that some queries requires the application of post-processing operations (e.g. such as aggregation), a definition of a correct answer for the triple path format had to be generated. In the experimental set-up a correct answer is given by a triple path containing the URI supporting the final answer. For the example query '*How many films did Leonardo DiCaprio star in?*' the triple paths containing the URIs for the films were considered as the correct answer instead of the number of movies.

Query Set Type	MRR	Avg. Precision	Avg. Recall
Full DBPedia Query Set	0.516	0.482	0.491
Partial DBPedia Query Set	0.680	0.634	0.645

TABLE I
Quality of results for the semantic space measured using 50 natural language queries over DBPedia. The first row represents the results for the full QALD query set while the second row contains a reduced query set where some classes of queries were removed.

For evaluation purposes the entity indexes corresponding to the class and instance entity spaces were generated for all DBPedia instances and classes. In order to simplify the experimental set-up, only relation vector spaces associated with entities which were effectively explored by the algorithm were generated, without any impact on the results reported on the evaluation of the approach. The final approach was able to answer 58% of the queries. The results were collected with a minimum level of post-processing. The final *mean reciprocal rank*, *avg. precision* and *avg. recall* are given on the table 1. The measurements for each query and the output data generated from the experiment can be found online [4].

In order to understand the collected measures, the errors for the set of unanswered queries were classified into 5 categories: *PODS Error* (where the final PODS query form did not match the dataset structure), *Literal Pivot Error* (queries in which the main entity was a literal instead of an object resource), *Overloaded Pivot Error* (queries in which the main entity is a class with more than 3 terms e.g. yago:HostCitiesOfTheSummerOlympicGames), *Relatedness Error* (where the relatedness measure lead to a wrong answer) and *Combined Pre/Post-Processing Error* (queries which demanded more sophisticated query interpretation and post-processing). Table 2 contains the distribution of error types. The complementary error analysis for each query can be found online [4].

The error analysis shows that the distributional approach was able to cope with the semantic variation of the dataset (low level of *Relatedness Error*). The low level of *PODS Error* also shows that PODSs provide a primary query representation suitable for the proposed query approach and for the dataset representation. Queries referring to literal objects as key query entities are currently not addressed by the approach (*Literal Pivot Error*) since only object resources are mapped into pivot entities in the entity space. This limitation can be addressed by generating object resources for data properties in the index construction. Most of the errors in the evaluation are in the *Combined Pre/Post-Processing Error* category, which concentrates errors relative to the lack of a pre/post-processing analysis necessary to cope with a natural language query scenario, such as answer type detection, more comprehensive linguistic

Error Type	% of Queries
PODS Error	8%
Literal Pivot Error	4%
Overloaded Pivot Error	8%
Relatedness Error	2%
Combined Pre/Post-Processing Error	20%

TABLE II
ERROR TYPES AND DISTRIBUTION

analysis of the query, post-processing filters, etc. Despite the relevance of evaluating the suitability of the proposed semantic space as a natural language query scenario, this error category does not reflect directly the effectiveness of the semantic representation and query approach as a supporting structure for the natural language query process. The second line in table 1 provides a comparative basis of quality measures removing the category containing errors which are considered addressable in the short term (Literal Pivot Error) and the category which does not reflect the core of the evaluation for this work (Pre/Post-Processing Error). Compared to the results using the approach described in [8] but using the full QALD dataset, there is an improvement of 5.2% over mrr, 18% over avg. precision, and 8.2% over avg. recall. The individual analysis of the entity and spreading activation queries shows that the introduction of the proposed refinements for the semantic space construction led to a quantitative improvement which might be overshadowed by errors present in the *Combined Pre/Post-Processing Error* category.

The evaluation focused on the determination of the quality of the approach. No rigorous index construction performance evaluation was considered since, to be comparatively meaningful with existing approaches, a minimum level of optimization in the index construction process was necessary. One clear strength of the approach from the index construction perspective is the fact that the intrinsic nature of the index makes its construction process straightforward to parallelize, where the creation of index entries associated with relation vector bundles can be easily distributed.

VI. Related Work

The related work section concentrates on the analysis of works proposing new vector space based models and structure indexes. The motivation for this section is both to provide to the reader a perspective over existing trends in the semantic search space [13][14] and also to provide a comparative basis with existing work for RDF structure indexes [16][17].

In [13] Clark et al. provide a formal description of a compositional model of meaning, where distributional models are unified with a compositional theory of grammatical types (using Labek's pregroup semantics [12]). The approach attempts to unify the quantitative strength of distributional approaches with the compositionality provided by symbolic approaches. The final mathematical structure uses vectors to represent word meanings, grammatical roles represent types in a pregroup and the tensor product to allow the composition of meaning and types. The work concentrates on the mathematical formalization of the abstract model and does not provide an instantiation in a more specific distributional semantic space. The model proposed by Clark et al. share with the proposed T-Space the multidimensional aspect in the representation of the distributional meaning and the compositional model used in the construction of the semantic space. However, the compositional approach used in the T-Space is defined over the order given by the RDF data and by the PODS structure. In addition, the work presented on this paper concentrates on the instantiation and evaluation of the proposed model, while [13] concentrates on the formalization of the approach.

In [14] van Rijsbergen proposes the use of the formalism behind quantum physics to provide a principled theoretical framework for the investigation of new information retrieval models. The idea behind these approaches is to use the support provided by the probabilistic, logical and geometrical aspects of Hilbert spaces to create a user-oriented search model. Different quantum-inspired IR approaches have emerged from this motivation. In [15], Piwowarski et al. proposes a quantum-based IR framework which relies on a multidimensional representation of documents, where each document defines a subspace built by segmenting the document into fragments and by associating each fragment to a weighted set of information need states based on document terms. The query vector is built by getting the context window of the term for each incidence in the document collection, building a distributional representation for that specific collection. Compared to the approach of Piwowarski et al., the T-Space model also uses a multidimensional representation of information. However, the structure of the final semantic space and the distributional approach used are fundamentally different. In addition, the approach introduced by Piwowarski et al. does not define a compositionality model.

The proposed approach converges three trends, where the representational power provided by multidimensional vector spaces implementing semantic compositionality meets the flexibility of distributional semantics. In addition, the application of the T-Space model over RDF data provides a bidirectional bootstrap benefit, where the additional structure typing provided by RDF data can improve the semantics of the space for different applications (e.g. unstructured text indexing).

In the space of structure indexes for RDF, Semplore [16] 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 (individuals, classes, properties and objects) 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. Also in the structure

index space, Dong & Halevy [17] proposes 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 [16][17] 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 the previous Treo query approach [8], this work generalizes the basic elements present in [8], to build a multidimensional semantic space. The multidimensional semantic space generalization allows the conceptual alignment between the proposed approach and existing works and trends in semantic spaces and multidimensional vector spaces. The generalization also includes a change from the previous semantic relatedness approach, which was based on a link-based relatedness measure (Wikipedia Link Measure [18]), to a distributional approach based on Explicit Semantic Analysis (ESA). An additional refinement includes the entity indexing strategy which moved from a uniform entity indexing to an entity index which differentiates instances (TF/IDF) and classes (ESA). Differently from the previous approach which had to compute the semantic relatedness for each relation during query time, this work proposes the introduction of a relation subspace which is materialized into a relation index associated with each entity, bringing a principled solution to reduce the original associated query execution time.

VII. Conclusion & Future Work

This work proposes a semantic space focused on addressing a fundamental challenge for RDF data queries, where the data model heterogeneity of the Web demands a query approach focused on abstracting users from an a priori understanding of the data model behind datasets. Key elements in the construction of the approach are the application of *distributional semantics*, which in this work is defined by Explicit Semantic Analysis (ESA), a *compositional semantic model* based on the structure of RDF and on the use of dependency analysis and a *hybrid search model* where *entity-centric search* is complemented by *spreading activation search*. The final multidimensional semantic space allows data model independent and expressive natural language queries over the RDF data, achieving *mean reciprocal rank* = 0.516, *avg. precision* = 0.482 and *avg. recall* = 0.491, evaluated using 50 natural language queries over DBPedia. By introducing an additional semantically-rich structure provided by RDF data into the construction of a semantic space, the proposed model also points into a promising direction for investigation, where RDF data embedded in a flexible semantic space can be used to bootstrap the creation of more complex semantic spaces.

Future work will include the implementation of optimizations in the index construction process and evaluation of the index construction time, the elimination of limitations which are considered addressable in the short term and the implementation of a question answering (QA) system over RDF data using the proposed index. The implementation of QA features will allow the comparative evaluation against existing QA systems [3].

Acknowledgment

This work 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, http://www.w3.org/DesignIssues/LinkedData.html, 2006.
[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann, DBpedia - A crystallization point for the Web of Data, Web Semantics: Science, Services and Agents on the World Wide Web 7, 2009.
[3] 1st Workshop on Question Answering over Linked Data (QALD-1), http://www.sc.cit-ec.uni-bielefeld.de/qald-1, 2011.
[4] Evaluation Dataset, http://treo.deri.ie/semanticspace/icsc2011.htm, 2011.
[5] E. Gabrilovich and S. Markovitch, Computing semantic relatedness using Wikipedia-based explicit semantic analysis, International Joint Conference On Artificial Intelligence, 2007.
[6] G. Furnas, T. Landauer, L. Gomez, and S. Dumais, The vocabulary problem in human-system communication, Communications of the ACM, vol. 30, no. 11, pp. 964-971, 1987.
[7] 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, vol. 20, no. 1, 2008.
[8] A. Freitas, J.G. Oliveira, S. O'Riain, E. Curry, and J.C Pereira 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, 2011.
[9] P. D. Turney and P. Pantel, From Frequency to Meaning: Vector Space Models of Semantics, Journal of Artificial Intelligence Research 37, pp. 141-188, 2010.
[10] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, International Joint Conference On Artificial Intelligence, 1995.
[11] S. Pado and M. Lapata: Dependency-Based Construction of Semantic Space Models, Computational Linguistics, 33:2, 161-199, 2007.
[12] J. Lambek, The mathematics of sentence structure, The American Mathematical Monthly, 65(3), pp. 154-170, 1958.
[13] S. Clark and S. Pulman, Combining symbolic and distributional models of meaning, In Proceedings of the AAAI Spring Symposium on Quantum Interaction, Stanford, CA, pp. 52-55, 2007.
[14] C. J. van Rijsbergen, The Geometry of Information Retrieval, Cambridge University Press New York, NY, USA, 2004.
[15] B. Piwowarski, M. Lalmas, I. Frommholz and C. J. van Rijsbergen, Exploring a Multidimensional Representation of Documents and Queries, In Proceedings of RIAO 2010, Adaptivity, Personalization and Fusion of Heterogeneous Information, 2010.
[16] H. Wang, Q. Liu, T. Penin, L. Fu, L. Zhang, T. Tran, Y. Yu and Y. Pan, Semplore: A scalable IR approach to search the Web of Data, Web Semantics: Science, Services and Agents on the World Wide Web, vol. 7, no. 3, pp. 177-188, 2009.
[17] X. Dong and A. Halevy, Indexing dataspaces, In Proceedings Proceedings of the 2007 ACM SIGMOD international conference on Management of Data, 2007.
[18] D. Milne and 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), Chicago, IL, 2008.