Combining a co-occurrence-based and a semantic measure for entity linking

Bernardo Pereira Nunes[1,2], Stefan Dietze[2], Marco Antonio Casanova[1], Ricardo Kawase[2], Besnik Fetahu[2], Wolfgang Nejdl[2]

[1] Department of Informatics - PUC-Rio - Rio de Janeiro, RJ, Brazil
{bnunes, casanova}@inf.puc-rio.br

[2] L3S Research Center - Leibniz University Hannover - Hannover, Germany
{nunes, dietze, kawase, fetahu, nejdl}@l3s.de

Abstract. One key feature of the Semantic Web lies in the ability to link related Web resources. However, while relations within particular datasets are often well-defined, links between disparate datasets and corpora of Web resources are rare. The increasingly widespread use of cross-domain reference datasets, such as Freebase and DBpedia for annotating and enriching datasets as well as documents, opens up opportunities to exploit their inherent semantic relationships to align disparate Web resources. In this paper, we present a combined approach to uncover relationships between disparate entities which exploits (a) graph analysis of reference datasets together with (b) entity co-occurrence on the Web with the help of search engines. In (a), we introduce a novel approach adopted and applied from social network theory to measure the connectivity between given entities in reference datasets. The connectivity measures are used to identify connected Web resources. Finally, we present a thorough evaluation of our approach using a publicly available dataset and introduce a comparison with established measures in the field.

Keywords: Semantic connectivity, co-occurrence-based measure, linked data, data integration, link detection, semantic associations

1 Introduction

The emergence of the Linked Data approach has led to the availability of a wide variety of structured datasets on the Web[1] which are exposed according to Linked Data principles [3]. However, while the central goal of the Linked Data effort is to create a well-interlinked graph of Web data, links are still comparatively sparse, often focusing on a few highly referenced datasets such as DBpedia[2], YAGO [28] and Freebase[3], while the majority of data exists in a rather isolated fashion. This is of particular concern for datasets which describe the same or potentially *related* resources or real-world *entities*.

[1] http://lod-cloud.net/state
[2] http://dbpedia.org
[3] http://www.freebase.com

For instance, within the academic field, a wealth of potentially connected entities are described in bibliographic datasets and domain-specific vocabularies, while no explicit relationships are defined between equivalent, similar or connected resources [8].

Furthermore, knowledge extraction and Named Entity Recognition (NER) tools and environments such as GATE [5], DBpedia Spotlight[4], Alchemy[5], AIDA[6] or Apache Stanbol[7] are increasingly applied to automatically generate structured data (entities) from unstructured resources such as Web sites, documents or social media. For example, such automatically generated data may provide some initial classification and structure, such as the association of terms with entity types defined in a structured RDF schema (as in [22]). However, entities extracted via Natural Language Processing (NLP) techniques usually are noisy, ambiguous and lack sufficient semantics. Hence, identifying links between related entities within a particular dataset, as well as with pre-existing knowledge, serves three main purposes (a) enrichment, (b) disambiguation and (c) data consolidation. Often, dataset providers aim at *enriching* a particular dataset by adding links (*enrichments*) to comprehensive reference datasets. Current interlinking techniques usually resort to mapping entities which refer to the same resource or real-world entity, e.g., by creating `owl:sameAs` references between an extracted entity representing the city "Berlin" with the corresponding Freebase and Geonames[8] entries.

However, additional value lies in the detection of related entities within and across datasets, e.g., by creating `skos:related` or `so:related` references between entities that are to some degree connected [10, 14]. In particular, the widespread adoption of reference datasets opens opportunities to analyse such reference graphs to detect the *connectivity*, i.e., the *semantic association* [2, 26] between a given set of entities. However, uncovering these connections would require the assessment of very large data graphs in order to (a) identify the paths between given entities and (b) measure their meaning with respect to a definition of semantic connectivity.

In this paper, we present a general-purpose approach that combines a co-occurrence-based and a semantic measure to uncover relationships between entities within reference datasets in disparate datasets. Our novel semantic connectivity score is based on the Katz index [16], a score for measuring relatedness of actors in a social network, which has been adopted and expanded to take into account the semantics of data graphs, while the co-occurrence-based method relies on Web search results retrieved from search engines. Finally, we evaluate the approach using the publicly available USAToday corpus and compare our entity connectivity results with related measures.

The remainder of this paper is structured as follows. Section 2 discusses previous related work in the field. Section 3 presents the use case scenario that motivated our approach. Section 4 presents our entity connectivity approach. Section 5 and Section 6 show the evaluation strategies and their results. Finally, Section 7 summarizes our contributions and discusses future work.

[4] `http://dbpedia.org/spotlight`
[5] `http://www.alchemyapi.com`
[6] `http://adaptivedisclosure.org/aida/`
[7] `http://incubator.apache.org/stanbol`
[8] `http://www.geonames.org`

2 Related Work

Lehmann et al. [17] introduces RelFinder, which shows semantic associations between multiple entities from a RDF dataset, based on a breadth-first search algorithm, that is responsible for finding all related entities in the tripleset. Contrasting with RelFinder, Seo et al. [25] proposed the OntoRelFinder that uses a RDF Schema for finding semantic associations between two entities through its class links. Scarlet [23, 24] is another approach that relies on different schemas to identify relationships between entities.

Han et al. [15] proposes a slightly different approach. Instead of finding connections between two given entities, they expect to find the entities that are most connected, with respect to a given relationship and entity. This approach is interesting since it throws another perspective on the problem that we consider. However, they look for connected entities by means of a known relationship, while we aspire to uncover such connections between known entities.

Anyanwu et al. [1] present the SemRank, a customizable query framework that allows different setups for ranking methods, resulting in different perspectives for the same query. Thus, given two entities, depending on the setup the search results vary from more traditional (e.g. common connections or closest paths between entities) to less traditional (e.g. longer paths). In our approach, we consider both short and long paths to determine connectivity between two entities and Web resources.

Work from Leskovec et al. [18] presents a technique suggesting positive and negative relationships between people in a social network. This notion is also addressed in our method, but we take into account the path length. The longer is the path, the smaller is its contribution to the score.

The problem of discovering relationships between entities was also addressed by Damljanovic et al. [6] in Open Innovation scenarios, where companies outsource tasks on a network of collaborators. Their approach exploits the links between entities extracted from both the user profiles and the task descriptions in order to match experts and tasks. For this task, they use reference datasets and distinguish between entities as hierarchical and transversal. Following her approach, we distinguish between both relations types, although we focus on transversal relations.

Related work in the field of recommender systems includes the work by Passant [20], which presents a linked data semantic distance measure (LDSD) for music recommendation, by taking mainly into account incoming and outgoing links as well as indirect links between resources (i.e., songs and singers) to determine a recommendation score, used for recommending both direct and lateral music. In later work [19], he introduces a filtering step, by removing properties between resources that are not meaningful in the music context. Work on movie recommendation by Souvik et al. [7] considers an approach based on object features in order to improve movie recommendation, by using several similarity functions that deal with nominal, boolean and numeric features. Furthermore, they also use a linear regression method to assign weights for each feature type. Although this method presents good results, they do not consider semantic connections to uncover latent features.

Fang et al. [9] introduces the REX system, which computes a ranked list of entity pairs to describe entity relationships. The graph structure is decomposed for an entity pair resulting in unique graph patterns and ranks, where these patterns are matched ac-

cording to a measure of interestingness, based on the traditional random walk algorithm and the patterns found between an entity pair.

Sieminski [27] presents a method to measure the semantic similarity between texts on the Web, which consists of a modified *tf-idf* model and semantic analysis that makes use of WordNet structure. However, unlike his work, we explore the connections given by transversal properties in order to uncover latent connections between texts, rather than to explore similarity between them.

From the approaches outlined, we combine different techniques to uncover connections between disparate entities, which allows us to exploit the relationships between entities to identify connected Web resources.

3 Motivation

In this section we describe an example originating from actual Web information integration problems to illustrate the motivation of our work on discovering *latent semantic relationships* through its *semantic relations*.

The example below shows two descriptions of documents extracted from the USAToday corpus. Note that, the underlined terms refer to the recognised entities in each document derived from an entity recognition and enrichment process.

(i) The Charlotte Bobcats could go from the NBA's worst team to its best bargain.

(ii) The New York Knicks got the big-game performances they desperately needed from Carmelo Anthony and Amar'e Stoudemire to beat the Miami Heat.

Although both documents are clearly related to Basketball/Sports topics, linguistic and statistical approach would struggle to point out that both documents are connected. First, both textual descriptions are rather short and lack sufficient contextual information what makes it harder for purely linguistic or statistical approaches to detect their connectivity. Second, in this particular case, there are no significant common words between the documents. Usually, statistical and linguistic approaches are particularly suitable for cases where large amounts of textual content is available to detect the relationships between Web resources. In particular, some common terminology is required for detecting similarities between Web resources.

On the other hand, these challenges can be partially overcome by taking advantage of structured background knowledge to disambiguate and enrich the unstructured textual information. The example shows two documents, each associated with a particular entity, where the term *Charlotte Bobcats* was enriched with the entity `http://dbpedia.org/resource/Charlotte_Bobcats` in the document (i) and the term *Carmelo Anthony* was enriched with the entity `http://dbpedia.org/resource/Carmelo_Anthony` in the document (ii). Thus, analysing the DBpedia graph uncovers a connection between *Charlotte Bobcats* and *Carmelo Anthony* (being a basketball team and player, respectively) and hence allows us to establish a connection between the entities and their connected Web resources. Specifically, both entities are connected through the path: *Charlotte Bobcats* $\leftrightarrow$ *Eastern Conference (NBA)* $\leftrightarrow$ *New York Knicks* $\leftrightarrow$ *Carmelo Anthony*, where the intermediary entities uncover a connection between *Charlotte Bobcats* and *Carmelo Anthony*.

4 Approach

In this section, we introduce two novel measures for entity interlinking, a semantic graph-based connectivity score and one which utilises co-occurrence on the Web. Both detect complementary relationships between entities as results show in Section 6.

4.1 Semantic Connectivity Scores (SCS)

In this section, we define a semantic connectivity score between entities, based on a reference graph that describes entities and their relations. Similar to Damljanovic et al. [6], we distinguish between *hierarchical* and *transversal* relations in a given graph. Typical hierarchical properties in RDF graphs are, for instance, `rdfs:subclassOf`, `dcterms:subject` and `skos:broader`, and usually serve as an indicator for similarity between entities. In contrast, transversal properties do not indicate any classification or categorisation of entities, but describe non-hierarchical relations between entities which indicate a form of connectivity independent of their similarity.

To illustrate the semantic connectivity, we refer to the pair of entities "Jean Claude Trichet" and "European Central Bank", which have no equivalence or taxonomic relation, but have a high connectivity according to transversal properties. For example, the "European Central Bank" is linked to the entity "President of the European Central Bank" through the RDF property `http://dbpedia.org/property/leaderTitle` that, for its part, links to "Jean Claude Trichet" through the RDF property `http://dbpedia.org/property/title`.

Let R be a reference triple set and G be the associated undirected graph, in the sense that the nodes of G correspond to the individuals occurring in R and the edges of G correspond to the properties between individuals defined in R. From this point on, we will refer to the individuals occurring in R as *entities*.

We define the *semantic connectivity score* (SCS) between a pair of entities (e_1, e_2) in G as follows:

$$SCS(e_1, e_2) = \sum_{l=1}^{\tau} \beta^l \cdot |paths^{<l>}_{(e_1,e_2)}| \tag{1}$$

where $|paths^{<l>}_{(e_1,e_2)}|$ is the number of transversal paths between the entities e_1 and e_2 of length l, τ is the maximum length of paths considered (in our case $\tau = 4$, as explained in more details below), and $0 < \beta \leq 1$ is a positive damping factor. The damping factor β^l is responsible for exponentially penalizing longer paths. The smaller this factor, the smaller the contribution of longer paths to the final score. Obviously, if the damping factor is 1, all paths will have the same weight independently of length. In previous experiments, we observed that $\beta = 0.5$ presented better results in terms of precision [21].

The semantic connectivity score between entities is a variation of the Katz index [16] introduced to estimate the relatedness of actors in a social network. We introduced a number of derivations to improve its applicability to large graphs and to reflect the added semantics provided by labelled edges in RDF graphs, as opposed to the limited semantics of edges in a social network. A detailed discussion of the advantages and limitations of our approach is provided in Section 7.

As one main adaptation of Katz, we exploit the semantics of edges in a given data graph by excluding hierarchical properties from our connectivity score computation. As defined earlier, connectivity is indicated by transversal properties. Currently, no further distinction between property types has been introduced into our formula, though we explicitly envisage such an adaptation. However, given the vast amount of property types in datasets such as DBpedia, a distinction at the general and domain-independent level is computationally too expensive and therefore does not scale. Instead, we particularly suggest the adaptation of our formula to specific domains or entity types, which allows the consideration of more fine-grained semantics provided by distinct property types.

In addition, we opted for an undirected graph model in order to reduce computational complexity, since a property is often found in its inverse form (e.g. fatherOf/sonOf) [13]. While most current entity interlinking techniques apply their approaches to a restricted set of entity types to allow some sort of tailoring and, as consequence, more precise results, our experiments in Section 5 show that even our fairly generic score produces useful and promising results, which can be improved by means of domain-specific adaptations.

As the semantic score is based on the number of paths and distances (length of a path) between entities, SCS considers only paths with a maximum length ($\tau = 4$), as also adopted in [9]. This maximum length was identified by investigating the semantic score behaviour for edge distances ranging from 1 to 6, as detailed below.

In our experiments, we randomly selected 200 entity pairs and computed the semantic connectivity score (SCS) (see Eq. 1) for the aforementioned path length range (see Figure 1a). As expected, the average number of paths grows exponentially with the distance (i.e. the path length), see Figure 1a.

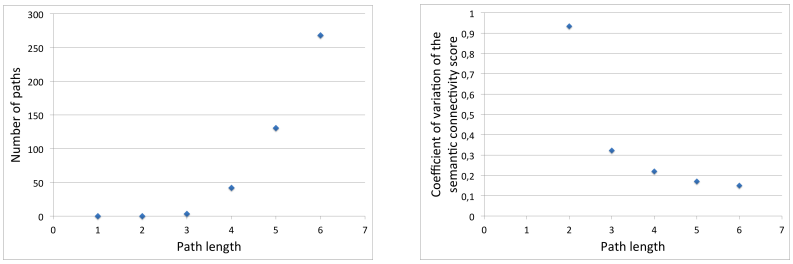
Thus, as in the small world assumption [29], beyond a certain path length, every node pair is likely to be connected. However, as opposed to the small world assumption that people are interlinked through a maximum distance of 6 connections, we found that for interlinking entities this number is lower, approximately by two degrees. This decision is backed up according to several experiments, detailed below.

After computing all entity pairs for different path lengths, we evaluated the coefficient of variation of the semantic score, $C_v = \sigma/\mu$, where, for a given length, σ is the standard deviation of the number of paths and μ is the mean number of paths. This coefficient is used to measure the spread of the semantic score distribution, taking into account an upper bound path length (see Figure 1b).

From the behaviour of the curve in Figure 1b, it is apparent that the contribution of paths with distances greater than 4 edges is low. Also as expected, the average running time to compute the path grows exponentially with the distance. Hence, including longer path lengths increases significantly the computational costs, while producing only minimal gains in performance. Thus, we obtain the best balance between performance and informational gain to the semantic score. That is, we minimise the path length considered, while maximise the contribution in the overall score.

4.2 Co-occurrence-based Measure (CBM)

We introduce in this section a co-occurrence-based measure between entities that relies on an approximation of the number of existing Web pages that contain their labels. For



(a) (b)

Fig. 1: Maximum path length analysis. Figure (a) shows the number of paths with respect to length and (b) shows the gain of information when considering different path lengths.

example, we estimate the CBM score of a pair of entities by submitting queries (such as "Jean Claude Trichet" + "European Central Bank") to a search engine and retrieving the total number of search results that contain the entity labels in their text body. Thus, we define the CBM score of a pair of entities e_1 and e_2 as follows:

$$CBM(e_1, e_2) = \begin{cases} 0, & \text{if } count(e_1) = 0 \text{ or } count(e_2) = 0 \\ 1, & \text{if } count(e_1) = count(e_2) = count(e_1, e_2) = 1 \\ \dfrac{Log(count(e_1, e_2))}{Log(count(e_1))} \cdot \dfrac{Log(count(e_1, e_2))}{Log(count(e_2))}, & \text{otherwise} \end{cases} \tag{2}$$

where $count(e_i)$ is the number of Web pages that contain an occurrence of the label of entity e_i, and $count(e_1, e_2)$ is the number of Web pages that contain occurrences of the labels of both entities. Note that $count(e_1, e_2)$ is a non-negative integer always less than or equal to $count(e_i)$, for $i = 1, 2$. Hence, the final score is already normalised to $0 \leq CBM(e_1, e_2) \leq 1$.

There are other similar approaches to quantify the relation between entities, such as Pointwise Mutual Information (PMI)[4] and Normalised Google Distance (NGD)[12]. However, they take into account the joint distribution and the probability of their individual distributions, which requires to know a priori the total number of Web pages searched by a search engine.

To illustrate the co-occurrence-based score (CBM), consider the values $count(e_1) = count(e_2) = count(e_1, e_2)$, meaning that all occurrences of e_1 and e_2 appear together. In this case, the resulting co-occurrence-based score is 1, disregarding the number of search results.

For example, having $count(e_1) = count(e_2) = count(e_1, e_2) = 10$ or $count(e_3) = count(e_4) = count(e_3, e_4) = 1000$, would result in the same score. Evidently, if we would consider the probabilities, as in PMI or NGD, the latter case would get a higher score. Nevertheless, since we are not interested in disjoint comparisons, e.g., $CBM(e_1, e_2)$ against $CBM(e_3, e_4)$, we do not need to estimate the total number of pages, neither include it in the formula.

4.3 Towards a Combined Measure

As shown in previous work [21], although there is an overlap between the semantic and co-occurrence based approaches, some relationships cannot be uncovered by co-occurrence methods or by semantic methods alone. Thus, given that the results from SCS and CBM are seen as complementary, one conclusion is to combine them, which provides the advantage of scalability at discovering entity connections, where CBM would be used as a default approach, and SCS could be employed as an extensive search for finding latent connections in the resulting set of entity pairs deemed unconnected according to CBM, see Eq. 3.

$$\alpha_{CBM+SCS}(e_i, e_j) = \begin{cases} CBM(e_i, e_j), & \text{if } CBM(e_i, e_j) > 0 \\ SCS(e_i, e_j), & \text{otherwise} \end{cases} \tag{3}$$

where e_i and e_j are entities and $i \neq j$.

5 Evaluation Method

5.1 Dataset

The dataset for assessing entity connectivity consists of a set of 40,000 document pairs randomly selected from the USAToday news Website[9], where each document contains a title and a summary as textual content. The summary of each document has on average 200 characters. The corpus was annotated using DBpedia Spotlight[10] which resulted in approximately 80,000 entity pairs.

5.2 Gold Standard

Given the lack of benchmarks for validating latent relationships between entities, we created a gold standard using CrowdFlower[11], a crowdsourcing platform. To ensure a sufficient quality of the results, we required each user to pass through a set of tests where correct answers were known already, what allowed us to filter out poor assessors. In this way, we were able to avoid relevance judgements from untrusted workers. Moreover, as our corpus is focused on American news, we restrict the assessment only to workers located in the United States.

Thus, in order to construct the gold standard, we randomly selected 1000 entity pairs and 600 document pairs to be evaluated. The evaluation process consisted of a questionnaire in a 5-point Likert scale model where participants are asked to rate their agreement of the suggested semantic connection between a given entity pair. Additionally, we inspected participants' expectations regarding declared connected entities. In this case, presenting two entities deemed to be connected, we asked participants if such connections were expected (from *extremely unexpected* to *extremely expected* in the Likert scale).

[9] `http://www.usatoday.com`
[10] `http://spotlight.dbpedia.org/`
[11] `https://www.crowdflower.com/`

The collected judgements provided a gold standard for the analysis of our techniques. Note that in the case of this work, additional challenges are posed with respect to the gold standard, because our semantic connectivity score is aimed at detecting possibly unexpected relationships which are not always obvious to the user. To this end, a gold standard created by humans provides an indication of the performance of our approach with respect to precision and recall, but it may lack appreciation of some of our found relationships (see Section 6.2 for a detailed discussion).

5.3 Evaluation Methods

We also present a comparison of our approach against competing methods which measure connectivity via co-occurrence-based metrics to detect entity connectivity. In this evaluation we compared the performance of CBM against SCS and a third method (Explicit Semantic Analysis (ESA)) that is based on statistical and semantic methods.

Specifically, ESA [11] measures the relatedness between Wikipedia concepts by using a vector space model representation, where each vector entry is assigned using the *tf-idf* weight between the entities and its occurrence in the corresponding Wikipedia article. The final score is given by the cosine similarity between the weighted vectors. Note that ESA can be applied to measure any kind of corpora, not just Wikipedia concepts.

5.4 Evaluation Metrics

We measure the performance of the entity connectivity using the standard metrics of precision (P), recall (R) and $F1$ measure. Note that in these metrics, as relevant entity pairs, we consider those that were marked in the gold standard (gs) as connected according to the 5-point Likert Scale (*Strongly Agree* & *Agree*).

(P) is defined as the ratio of the set of retrieved entity pairs that have relevant uncovered connections over the set of entity pairs that have connections, see Eq. (4).

$$P = \frac{|\mu^{\tau}_{retrieved} \cap \mu_{relevant}|}{|\mu^{\tau}_{retrieved}|} \tag{4}$$

where $\mu_{relevant}$ is the set of retrieved entity pairs that are relevant and $\mu^{\tau}_{retrieved}$ is the set of retrieved connections that has a semantic connectivity score greater than a given threshold (τ). The threshold used in our experiments is shown in Section 6).

The recall measure is the ratio of the set of the retrieved entity pairs (R) that have relevant uncovered connections over all relevant connected entity pairs according to the gold standard, see Eq. (5).

$$R = \frac{|\mu^{\tau}_{retrieved} \cap \mu_{relevant}|}{|\mu_{relevant_{(gs)}}|} \tag{5}$$

where $\mu_{relevant_{(gs)}}$ is the set of all relevant entity pairs.

Finally, $F1$ measure shows the balance between precision and recall, and is computed as $F1 = 2 \cdot \frac{P \cdot R}{P+R}$.

6 Results

For each method described in the Sections 4 and 5, we present the results on their ability to discover latent connections over the entities. Furthermore, we also present an in depth-analysis of their shortcomings and advantages for discovering connections between entities.

6.1 Entity Connectivity Results

Table 1 shows the results obtained by the questionnaire and used as gold standard for the entity connectivity. The results are presented in a 5-point Likert scale of agreement ranging from *Strongly Agree* to *Strongly disagree*.

Table 1: Number of entity-pairs in each category (5-point Likert scale) in gold standard.

Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
63	178	127	227	217

In Figure 2, we report the performance for the co-occurrence-based score (CBM), Explicit Semantic Analysis (ESA) and our proposed adaptation of the Katz score (SCS). We considered as relevant all the entity pairs which had relevance judgements as *Strongly Agree* and *Agree*, and scores greater than a threshold. Since our task is to uncover latent relationships between entities rather than ranking them, we set the threshold to 0 (i.e. we include all results), but for some tasks we might want to raise this, e.g. for ranking or recommending.

According to Figure 2, SCS performs better in terms of precision whereas CBM achieves highest recall value. SCS and CBM present only minimal differences with respect to precision and recall, while ESA has the lowest values for all metrics.

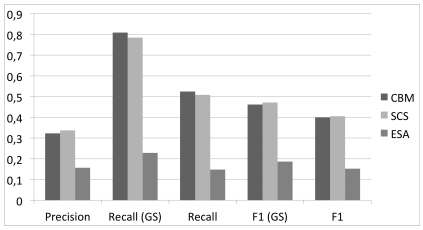


Fig. 2: P/R/F1 measure according to the gold standard (GS) amongst methods.

In addition to performance, we are also interested in the agreement between the methods. Identifying missed and detected relationships amongst all measures provides an indicator of their complementarity. In Table 2 we present a pairwise comparison of methods where we show the ratio of connections that are found by one method and missed by another. It is notable that CBM and SCS capture most of the connections, even though CBM misses 3.1% and 11.2%, and SCS misses 9.5% and 12.3% for *Strongly Agree* and *Agree* respectively.

Table 2: Ratio of connections detected by each method, according to the gold standard.

	CBM (*not* in SCS)	CBM (*not* in ESA)	SCS (*not* in CBM)	SCS (*not* in ESA)	ESA (*not* in CBM)	ESA (*not* in SCS)
Strongly Agree	**9.5%**	**76%**	**3.1%**	**71%**	**7.9%**	**9.5%**
Agree	**12.3%**	**63.4%**	**11.2%**	**60.1%**	**8.9%**	**6.7%**
Undecided	9.4%	60.6%	6.3%	59.8%	5.5%	7.9%
Disagree	15.0%	63.0%	7.1%	53.3%	7.1%	5.3%
Strongly Disagree	18.4%	63.1%	51.6%	4.6%	4.6%	6.9%

Besides the missed connections, we also take into account the expectedness of a connection between entity pairs. The expectedness shows how well established the connection is: an unexpected connection would be a relevant inferred indirect link between the entities. Thus, unexpectedness can be interpreted as a creation of novel links between entities. We see that SCS uncovers 25% of the unexpected connections, while CBM uncovers 16%. For this task, ESA was not able to uncover any new connections.

6.2 Results Analysis

In this section, we provide a detailed analysis of the results. The analysis is guided by the initial aims of our work on discovering latent connections between entities within a data graph (at varying path lengths), rather than competing with well established methods such as co-occurrence-based approaches widely deployed by search engines. To this end, the results of the listed approaches are complementary, where each of the approaches is able to establish unique entity connections.

In Figure 3, we show the agreement of entity pair ranking retrieved by SCS compared with CBM. The entity pair ranking follows an expected decline, where most connections are found at high ranks, whereas only a few are found at very low ranks.

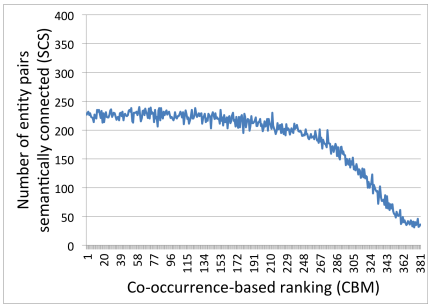


Fig. 3: The x-axis represents the ranking position x of entity pairs according to the CBM rankings. The y-axis represent the number of entity pairs ranked at xth position that have a semantic connection according to our connectivity threshold.

Table 3: Kendall tau and Jaccard-index between SCS and CBM entity rankings.

	k@2		**k@5**		**k@10**	
Dataset	Kendall tau	Jaccard-index	Kendall tau	Jaccard-index	Kendall tau	Jaccard-index
USAToday	0.40	0.09	0.47	0.19	0.52	0.21

As we can see in Figure 3, for the topmost rank of co-occurrence-based entity pairs, 225 of them have a semantic connection. Ideally, since these pairs are ranked in the top position, we expected to find a semantic connection between all of them. Arguably, the dependency rank-position to semantic connection should follow the trend where the lower the rank position, the higher the number of semantic connected entity pairs. In this sense, we can estimate which items have some missing relations. This is the first step in the task of actually discovering the missing relations. By observing the missing semantic ranked pairs on the *x*-axis, we can identify which entities miss some connection induced by the co-occurrence-based score (the problem introduced on Section 3). It is worth noting that, after the 260th rank position in the *x*-axis, the behaviour of the curve is in line with our expectations, i.e., the lower the correlation induced by the co-occurrence-based score, the lower that induced by the semantic connectivity score.

To show the complementarity between CBM and SCS, we used the Kendall tau rank correlation coefficient to assess the agreement of the entity ranks induced by the semantic connectivity score based on the DBpedia graph against the entity ranks induced by CBM. Table 3 shows the results.

As we can see from Table 3, the overlap between the rankings is not high. However, as our previous evaluation with the gold standard shows, this indicates that the scores induce different relationships between entities. The CBM score induces a relationship that reflects the overall co-occurrence of entities in the Web, whereas the semantic connectivity score mirrors the DBpedia graph.

Thus, as shown in Table 4, the CBM+SCS is the best performing approach compared to the other methods for the task of entity connectivity. Moreover, when comparing the F1 results from the CBM+SCS and SCS, we achieve significantly different results for p-value = 0.04 with 95% confidence.

Table 4: P/R/F1 measures according to gold-standard and amongst methods.

	CBM	SCS	ESA	CBM+SCS
Precision	0.32	0.34	0.16	0.34
Recall (GS)	0.81	0.78	0.23	0.90
Recall	0.52	0.51	0.15	0.58
F1 (GS)	0.46	0.47	0.19	0.50
F1	0.40	0.41	0.15	0.43

We would also like to point out the challenges posed by our approach on creating a gold standard. As mentioned previously, while our work aims at detecting semantic entity connections beyond traditional co-occurrences, this results in connections which might be to some extent unexpected yet correct, according to background knowledge

(such as DBpedia in our case). Hence, using a manually created gold standard, though being the only viable option, necessarily impacts the precision values for our work in a negative way, as correct connections might have been missed by the evaluators. This has been partially confirmed by the large number of detected co-occurrences which were marked as *undecided* by the users, where manual inspection of samples in fact confirmed a positive connection. This confirms that in a number of cases, connections were not necessarily incorrect but simply unknown to the users. Thus, we believe that a more thorough evaluation providing the evaluators with information on how a connection emerged, by showing all properties and entities that are part of a path greater than one, would give us more reliable judgements.

An example found in our evaluation is between the politicians "Barack Obama" and "Olympia Snowe", where the first is the current US president and the latter is one of the current senior US senators. Although the evaluators did not identify a connection between them, our semantic connectivity approach found several paths with length 2 or more. Additionally, they are related via several topics in real life, which confirms the validity of the paths found by our approach. For instance, this information could be exploited by news Websites for improving the user experience on finding related topics or news.

7 Discussion and Outlook

We have presented a general-purpose approach to discover relationships between entities, utilising structured background knowledge from reference graphs as well as co-occurrence of entities on the Web. To compute entity connectivity, we first introduced a semantic-based entity connectivity approach (SCS), which adapts a measure from social network theory (Katz) to data graphs, in particular Linked Data. We were able to uncover 14.3% entity connections not found by the state of the art method described here as CBM. While using a combination of CBM+SCS, we achieved a F1 measure of 43% for entity connectivity.

Our experiments show that SCS enables the detection of entity relationships that a priori linguistic and co-occurrence approaches would not reveal. Contrary to the latter, SCS relies on semantic relations between entities as represented in structured background knowledge, captured in reference datasets.

While both approaches (CBM and SCS) produce fairly good indicators for entity and document connectivity, an evaluation based on Kendall's tau rank correlation showed that the approaches differ in the relationships they uncover [21]. A comparison of agreement and disagreement between different methods revealed that both approaches are complementary and produce particularly good results in combination with each other. The semantic approach is able to find connections between entities that do not necessarily co-occur in documents (found on the Web), while the CBM tends to emphasise entity connections between entities that are not necessarily strongly connected in reference datasets. Thus, a combination of our semantic approach and traditional co-occurrence-based measures provide promising results for detecting related entities.

Despite the encouraging results, one of the key limitations of our Katz-based measure is the limited consideration of edge semantics in its current form. At the moment,

property types are distinguished only at a very abstract level, while valuable semantics about the meaning of each edge (i.e., each property) is left unconsidered during the connectivity computation. We are currently investigating approaches to take better advantage of the semantics of properties in data graphs.

Another issue faced during the experimental work is related to the high computational demands when applying our approach to large-scale data, which restricted our experiments to a limited dataset. In particular, the combination of traditional measures with our approach could help in improving performance, for instance, by computing our semantic connectivity only between entity pairs deemed unconnected by traditional measures. In addition, reducing the gathering of paths to a limited set of nodes ("hub nodes") might help in further improving scalability.

References

1. K. Anyanwu, A. Maduko, and A. Sheth. Semrank: ranking complex relationship search results on the semantic web. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 117–127, New York, NY, USA, 2005. ACM.
2. K. Anyanwu and A. Sheth. p-queries: enabling querying for semantic associations on the semantic web. In *Proceedings of the 12th international conference on World Wide Web*, pages 690 – 699, Budapest, Hungary, 2003. ACM Press New York, NY, USA.
3. C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.
4. K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, Mar. 1990.
5. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, pages 168–175, Philadelphia, July 2002.
6. D. Damljanovic, M. Stankovic, and P. Laublet. Linked data-based concept recommendation: Comparison of different methods in open innovation scenario. In E. Simperl, P. Cimiano, A. Polleres, Ó. Corcho, and V. Presutti, editors, *ESWC*, volume 7295 of *LNCS*, pages 24–38. Springer, 2012.
7. S. Debnath, N. Ganguly, and P. Mitra. Feature weighting in content based recommendation system using social network analysis. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1041–1042, New York, NY, USA, 2008. ACM.
8. S. Dietze, H. Q. Yu, D. Giordano, E. Kaldoudi, N. Dovrolis, and D. Taibi. Linked education: interlinking educational resources and the web of data. In S. Ossowski and P. Lecca, editors, *SAC*, pages 366–371. ACM, 2012.
9. L. Fang, A. D. Sarma, C. Yu, and P. Bohannon. Rex: explaining relationships between entity pairs. *Proc. VLDB Endow.*, 5(3):241–252, Nov. 2011.
10. A. Ferrara, A. Nikolov, and F. Scharffe. Data linking for the semantic web. *Int. J. Semantic Web Inf. Syst.*, 7(3):46–76, 2011.
11. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.
12. R. Gligorov, W. ten Kate, Z. Aleksovski, and F. van Harmelen. Using google distance to weight approximate ontology matches. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 767–776, New York, NY, USA, 2007. ACM.

13. A. Graves, S. Adali, and J. Hendler. A method to rank nodes in an rdf graph. In C. Bizer and A. Joshi, editors, *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008), Karlsruhe, Germany, October 28, 2008*, volume 401 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.
14. H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson. When owl: sameas isn't the same: An analysis of identity in linked data. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *International Semantic Web Conference (1)*, volume 6496 of *LNCS*, pages 305–320. Springer, 2010.
15. Y.-J. Han, S.-B. Park, S.-J. Lee, S. Y. Park, and K. Y. Kim. Ranking entities similar to an entity for a given relationship. In *Proceedings of the 11th Pacific Rim international conf. on Trends in AI*, PRICAI'10, pages 409–420, Berlin, Heidelberg, 2010. Springer-Verlag.
16. L. Katz and L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, Mar. 1953.
17. J. Lehmann, J. Schüppel, and S. Auer. Discovering unknown connections - the dbpedia relationship finder. In S. Auer, C. Bizer, C. Müller, and A. V. Zhdanova, editors, *CSSW*, volume 113 of *LNI*, pages 99–110. GI, 2007.
18. J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 641–650, New York, NY, USA, 2010. ACM.
19. A. Passant. dbrec - music recommendations using dbpedia. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *International Semantic Web Conference (2)*, volume 6497 of *LNCS*, pages 209–224. Springer, 2010.
20. A. Passant. Measuring semantic distance on linking data and using it for resources recommendations. In *AAAI Spring Symposium: Linked Data Meets AI*. AAAI, 2010.
21. B. Pereira Nunes, R. Kawase, S. Dietze, D. Taibi, M. A. Casanova, and W. Nejdl. Can entities be friends? In G. Rizzo, P. Mendes, E. Charton, S. Hellmann, and A. Kalyanpur, editors, *Proceedings of the WoLE Workshop in conjuction with the 11th International Semantic Web Conference*, volume 906 of *CEUR-WS.org*, pages 45–57, Nov. 2012.
22. T. Risse, S. Dietze, W. Peters, K. Doka, Y. Stavrakas, and P. Senellart. Exploiting the social and semantic web for guided web archiving. In *Proceedings of the Second international conference on Theory and Practice of Digital Libraries*, TPDL'12, pages 426–432, Berlin, Heidelberg, 2012. Springer-Verlag.
23. M. Sabou, M. d'Aquin, and E. Motta. Exploring the semantic web as background knowledge for ontology matching. *J. Data Semantics*, 11:156–190, 2008.
24. M. Sabou, M. d'Aquin, and E. Motta. Relation discovery from the semantic web. In C. Bizer and A. Joshi, editors, *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008), Karlsruhe, Germany, October 28, 2008*, volume 401 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.
25. D. Seo, H. Koo, S. Lee, P. Kim, H. Jung, and W.-K. Sung. Efficient finding relationship between individuals in a mass ontology database. In T.-H. Kim, H. Adeli, J. Ma, W.-C. Fang, B. H. Kang, B. Park, F. E. Sandnes, and K. C. Lee, editors, *FGIT-UNESST*, volume 264 of *CCIS*, pages 281–286. Springer, 2011.
26. A. P. Sheth and C. Ramakrishnan. Relationship web: Blazing semantic trails between web resources. *IEEE Internet Computing*, 11(4):77–81, 2007.
27. A. Sieminski. Fast algorithm for assessing semantic similarity of texts. *IJIIDS*, 6(5):495–512, 2012.
28. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.
29. D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.