Can Entities be Friends?

Bernardo Pereira Nunes[1,2], Ricardo Kawase[1], Stefan Dietze[1], Davide Taibi[3], Marco Antonio Casanova[2], Wolfgang Nejdl[1]

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

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

[3] Italian National Research Council - Institute for Educational Technology - Palermo - Italy
davide.taibi@itd.cnr.it

Abstract. The richness of the (Semantic) Web lies in its ability to link related resources as well as data across the Web. 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 document corpora, opens up opportunities to exploit their inherent semantics to uncover semantic relationships between disparate resources. In this paper, we present an approach to uncover relationships between disparate entities by analyzing the graphs of used reference datasets. We adapt a relationship assessment methodology from social network theory to measure the connectivity between entities in reference datasets and exploit these measures to identify correlated Web resources. Finally, we present an evaluation of our approach using the publicly available datasets Bibsonomy and USAToday.

Keywords: Linked data, data integration, link detection, semantic associations

1 Introduction

The emergence of the Linked Data principles [2] has led to the availability of a wide variety of structured datasets[1] on the Web. 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, YAGO [18] and Freebase, 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*. For instance, within the academic field, a wealth of potentially related entities are described in bibliographic datasets and domain-specific vocabularies, while no explicit relationships are defined between equivalent, similar or related resources [5].

[1] `http://lod-cloud.net/state`

Furthermore, knowledge extraction, Named Entity Recognition (NER) tools and environments such as GATE [4], DBpedia Spotlight[2], Alchemy[3], AIDA[4] or Apache Stanbol[5] are increasingly applied to automatically generate structured data (entities) from unstructured resources such as Web sites, documents or social media. However, while such automatically generated data usually provides an initial classification and structure, for instance, the association of terms with entity types defined in a structured RDF schema (as in [14]), entities extracted via Natural Language Processing (NLP) techniques are usually noisy, ambiguous and lack sufficient semantics. Hence, identifying links between entities within such 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 such comprehensive reference datasets. Current interlinking techniques usually resort to map 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[6] entries.

However, additional value lies in the identification of related entities within and across datasets, e.g., by creating `skos:related` or `so:related` references between entities that are to some degree related [7]. In particular, the widespread adoption of reference datasets such as DBpedia or Freebase opens opportunities to discover related entities by analyzing the graph of used joint reference datasets to measure the relatedness, i.e., the semantic association [1, 17] between a given set of enrichments and, thus, entities. However, uncovering this relation would require the assessment of such reference graphs in order to (a) identify the paths between these given enrichments and (b) measure their meaning with respect to some definition of semantic relatedness.

In this paper, we describe an approach to identify relationships between disparate entities by analyzing the graphs of reference datasets using an algorithm adopted from social network theory and extended to the needs of our overall vision. The main goal is to detect and quantify the relatedness between given sets of disparate entities and thus, Web resources. We provide a general-purpose approach, which exploits the number of paths and the distance (length of a path) between given entities to compute a relatedness score between (a) extracted entities and (b) associated Web resources such as documents.

The remainder of this paper is structured as follows. Section 2 formally describes the problem addressed. Section 3 introduces our method. Section 4 and Section 5 show the evaluation strategies and their results, respectively. Section 6 reviews the literature. Finally, Section 7 summarizes our contributions and discusses future work.

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

2 Problem Definition

In this work, we aim at finding and measuring the connectivity, i.e. semantic association, between disparate entities and use it as a measure to compute the relatedness of documents which refer to such entities. Exploiting implicit semantic relationships between entities, beyond mere linguistic similarity between different Web resources, allows to uncover different kind of semantic relationships between Web resources.

According to Sheth et al. [16], a semantic association between two resources exists if they have semantic connectivity or semantic similarity. In this work we focus on the semantic association given by semantic connectivity.

For instance, let $G = (E, P)$ be a graph (e.g. RDF dataset), where E and P denote a finite set of entities and properties, respectively. A property $p_i \in P$ is represented by a finite set of entities $\{e_i, e_j\}$, where $e_i, e_j \in E$. Thus, given two entities e_1 and e_n, they have *semantic connectivity* [16] iff exists at least one path $\rho_{(e_1,e_n)}^{<max(l)>} = \{\{e_1, e_2\}, \{e_2, e_3\}, ..., \{e_{n-1}, e_n\}\}$ that links each other with a maximum l properties between them. Contrasting with [16], we constrained the paths to a maximum length $max(l)$, since reference datasets (e.g. DBpedia and Freebase) are densely connected and, hence, the probability that any two entities be connected through longer paths tends to be high.

For performance reasons (see Section 3), we assume undirected graphs. Therefore, the paths $\rho_{(e_1,e_n)}^{<max(l)>} = \{\{e_1, e_2\}, \{e_2, e_3\}, ..., \{e_{n-1}, e_n\}\}$ and $\rho_{(e_n,e_1)}^{<max(l)>} = \{\{e_n, e_{n-1}\}, ..., \{e_3, e_2\}, \{e_2, e_1\}\}$ are considered to be equal, that is, $\rho_{(e_1,e_n)}^{<max(l)>} = \rho_{(e_n,e_1)}^{<max(l)>}$.

Thus, the semantic connectivity between two given entities e_i and e_j can be measured by a score $\lambda(\delta_{(e_i,e_j)}^{<max(l)>})$, where $\delta_{(e_i,e_j)}^{<max(l)>}$ is a set of paths $\rho_{(e_i,e_j)}^{<max(l)>}$. We say that there is a semantic association between e_i and e_j iff $\lambda(\delta_{(e_i,e_j)}^{<max(l)>}) > 0$, and that there is no semantic association between e_i and e_j iff $\lambda(\delta_{(e_i,e_j)}^{<max(l)>}) = 0$.

Section 3 provides the details about the measure chosen to compute the score between two entities. This measure is applied to detect connectivity between entities and connectivity between Web resources (e.g. documents).

3 Approach

In this section, we present a method for computing the semantic connectivity between entities as well as corresponding Web documents. The process is divided into the following steps: (a) entity recognition and enrichment; (b) discovery of semantic associations between entities; (c) computation of semantic connectivity scores that express the relatedness between the entities.

3.1 Entity Recognition and Enrichment

The entity recognition and enrichment process extract rich, structured data about entities, such as locations, organizations or persons from unstructured Web resources. One fundamental goal is to, not only recognize named entities but, to enrich these with references to established reference datasets such as DBpedia or Freebase as means to disambiguate and expand entity descriptions.

Our approach currently applies two different methodologies: (1) gradual named entity recognition (NER) followed by subsequent enrichment, (2) integrated NER and enrichment. The first approach is currently exploited by the previously introduced ARCOMEM project and deploys GATE[7] components as a NER tool together with self-developed enrichment techniques using typed queries on DBpedia and Freebase [14]. While GATE extracts isolated typed entities, for instance an entity of type location with the label "Athens", enrichment is used to expand each entity with additional knowledge and to provide means for disambiguation.

The second approach exploits combined NER and disambiguation techniques which directly extract DBpedia and Freebase entities out of unstructured resources. As part of the current experiments proposed in this paper, we use a local deployment of the DBpedia Spotlight Web Service. While both approaches show particular advantages and disadvantages, a thorough evaluation with respect to precision and recall of retrieved entities is currently ongoing. However, since the focus of this paper is on the next two steps, our experiments use an evaluated set of extracted and enriched DBpedia entities.

3.2 Discovery of Semantic Associations between Entities

The second step of our approach aims at retrieving all paths with up to a maximum length between two given entities in the DBpedia graph. As this is a computationally expensive task, we adopted a pre-processing strategy, also used in [10], which computes the maximal connected subgraphs through a breadth-first search algorithm.

Instead of starting to find all the paths between two nodes, the algorithm verifies if both nodes belong to the same subgraph in the triple set. If the two nodes do not belong to the same subgraph, then a priori we know that no path with up to a predetermined maximum length exists between them. Otherwise, the process of finding all paths between two given nodes is initiated.

The maximum length of a path will be discussed in the next section. However, it is obvious that calculating long paths is expensive.

3.3 Semantic Connectivity and Document Relatedness Score

In order to compute the connectivity between two given enriched entities, we applied the Katz index proposed in [9] to calculate the relatedness of actors in a social network. This index takes into account the set of all paths between two nodes. The index also uses a damping factor β^l that is responsible for exponentially penalizing longer paths. The equation to compute the Katz index is as follows:

$$Katz(a,b) = \sum_{l=1}^{\tau} \beta^l \cdot |paths_{(a,b)}^{<l>}| \tag{1}$$

where $|paths_{(a,b)}^{<l>}|$ is the number of paths between a and b of length l and $0 < \beta \leq 1$ is a positive damping factor. The smaller this factor is, the smaller is the contribution of longer paths to the Katz index. Obviously, if the damping factor is 1, all paths will have

[7] http://gate.ac.uk

the same weight independently of the path length. In this work, we used $\beta = 0.5$ as our damping factor, since this value presented better results.

After computing the semantic connectivity score for a set of entities in a enriched dataset, a ranking of the most related entities is generated for each entity.

A major problem with the Katz index is that it is computationally expensive, since finding all paths between two nodes in not practical for large graphs. Thus, to overcome this problem, we set a threshold ($\tau = 4$) to the maximum path length between nodes, a decision that is backed up by the small world [21] phenomenon, which indicates that a pair of nodes is separated by a small number of connections. Thus, to compute all paths above this threshold would mean to add a constant factor for all indices.

One of the main applications of measuring entity connectivity is to discover *document relatedness*. In order to achieve such goal, we combine the results of the Katz index formula with entity co-occurrence scores. Thus, documents that contain the same entities receive an extra similarity bonus that would not be granted by the Katz index. The semantic relatedness score between documents is computed by the Eq. 2.

$$DRS(A, B) = \sum_{i \in A, j \in B, i \neq j} Katz(i, j) + \frac{|entity(A) \cap entity(B)|}{2} \tag{2}$$

where $entity(A)$ and $entity(B)$ denote the set of entities occurring in documents A and B, respectively.

4 Evaluation

In this section, we present the evaluation process to validate our approach. Our evaluation aims to assess the following criteria:

Computed connectivity between entities. Given the lack of benchmarks for validating entity connectivity, we rely on the wisdom of crowds to verify the relation between entities found by our semantic approach. Our assumption is that from the associations between terms (entity labels) suggested by Web users over time, a valid measure for connectivity emerges. In summary, two terms (that name entities) that co-occur to a high degree on the Web are considered related. With this "crowd-sourced" strategy, we exploit the wisdom of crowds to detect the co-occurrence of entities on the Web (See Section 4.2 for details). To assess the agreement of both approaches, we use a variation of the Kendall's Tau method [3].

Validity of computed document relationships. This evaluation is fundamental to prove the importance of considering the semantic associations between entities. Furthermore, although our motivation examples show a very strict scenario, where linguistic techniques would fail, our evaluation intends to show that this strategy also can be useful to improve linguistic approaches in common datasets.

4.1 Dataset

In this section, we describe the characteristics of the two distinct datasets used for the evaluation process. The first dataset consists of 200 randomly selected articles from the

USAToday[8] news Web site. Each article contains a title and a summary of the whole textual content. The second dataset consists of randomly selected documents from Bibsonomy[9], a repository of research publications, annotated based on a folksonomy. To sample the data, we randomly selected 5 tags and gathered the Bibsonomy entries for each of these tags. In total we ended up with 213 documents (titles and abstracts).

The entity recognition and enrichment process (Section 3.1) extracted 399 unique entities from the USAToday corpus while the Bibsonomy corpus was annotated with 1118 unique entities. That resulted in rather large number of entity pairs, each requiring to compute an individual relatedness score. For example, in the case of the USAToday dataset, we obtained approximately 80,000 pairs of entities and over 600,000 pairs of entities for the Bibsonomy dataset. As reported in Section 3.3, each comparison of entity pairs returns several distance values that are used to compute the semantic relatedness score between documents in our two corpora.

4.2 Crowd-sourced Connectivity Score

To compare and assess our retrieved connectivity scores, we introduced a crowd-sourced relationship detection approach. For this purpose, the Bing search engine[10] was used to identify entity correlations based on term co-occurrence on the Web.

In order to estimate a co-occurrence score, a query is submitted to the Bing search engine, retrieving the total number of search results that contain the labels of the queried pair of entities in their text body. Note that Bing and other search engines return an approximation of the number of existing Web pages that contain the queried terms. In addition, since search terms are *untyped* strings as opposed to entities, we are aware that this approach might carry ambiguous and misleading results. However, we assume that a large number of pages indicates high connectivity and a small number of pages indicates low connectivity between the queried terms. Thus, given two entities a and b, the final score is estimated by the Eq. 3.

$$CrowdScore(a,b) = \frac{Log(count(ab))}{Log(count(a))} \cdot \frac{Log(count(ab))}{Log(count(b))} \tag{3}$$

where *count(a)* is the number of Web pages that contain entity a, *count(b)* is the number of Web pages that contain entity b and *count(ab)* is the number of Web pages that contain both entities a and b. It is important to note that *count(ab)* is always less than or equal to *count(a)* and less than or equal to *count(b)*. Hence, the final score is already normalized to $0 \leq CrowdScore(a,b) \leq 1$. This score will be used as our benchmark. Thus, we rely on the wisdom of crowds to validate our approach.

4.3 Entity Connectivity Evaluation & Document Relatedness Evaluation

In the first step, we aim to evaluate the entity rankings given by both methods - semantic and crowd-sourced - we used a variation of Kendall's tau method, which is used for measuring the similarity of the top k items in two ranked lists.

[8] http://www.usatoday.com
[9] http://www.bibsonomy.org
[10] http://www.bing.com/developers/

Table 1. Kendall tau and Precision between the semantic and the crowd-sourced entity rankings.

	k@2		k@5		k@10		k@20	
Dataset	Kendall tau	Precision	Kendall tau	Precision	Kendall tau	Precision	Kendall tau	Precision
USAToday	0.01	0.09	0.13	0.19	0.20	0.21	0.22	0.23
Bibsonomy	0.0010	0.0016	0.0069	0.0081	0.0102	0.0109	0.0204	0.0210

The second experiment evaluates the ability to find evidences of document relatedness. In this step, we compute for each pair of documents a semantic-based relatedness score and a crowd-sourced score. To compute the semantic relatedness score between documents, we used Eq. 2 proposed in Section 3.3. Similarly to Eq. 2, we defined a crowd-sourced score that uses the CrowdScore defined in Eq. 3 as follows:

$$DRC(A,B) = \sum_{i \in A, j \in B, i \neq j} CrowdScore(i,j) + \frac{|entity(A) \cap entity(B)|}{2} \quad (4)$$

where A and B are documents; i and j are entities contained in each document respectively. As we explained in Section 3.3, a different score is given for pairs of entities where $i = j$.

Based on Eqs. 2 and 4, a list of the most related documents for each given document is generated. In order to assess the precision of the generated ranked lists in the USAToday dataset, we performed a manual evaluation to validate the top 1 related document for each of the 200 existing documents using both methods (semantically and crowd-sourced generated). The results show the precision of the top one related item.

For the Bibsonomy dataset we used the tags of each document as a ground truth for document relatedness [12]. In this evaluation, two documents are considered related if they share a set of tags. For this evaluation, we assessed precision at different levels. As mentioned earlier, comparison with linguistic clustering techniques is not suited for the purpose of our evaluation, since it would only detect correlation of terms, while our approach also considers semantic relationships between terms (as part of extracted entity labels).

5 Results

Regarding the agreement of the semantically generated entity ranking against the crowd-sourced ranking generated by the given co-occurrence of the terms in Web pages, as explained in Section 4.3, we performed a variation of the Kendall tau rank correlation coefficient, together with precision measures at different levels (see results in Table 1).

The main reason for these rather low values is that the information captured by both relatedness strategies expresses different relationships. While the crowd-sourced one gives us the overall human perception of the relatedness between different entities, the semantic strategy provide us with actual underlying connections between the entities.

Regarding to the document relatedness evaluation, Table 2 shows the results regarding the manually assessed recommendations for both strategies, verifying the validity of the semantic entity-document score. On the USAToday dataset the success of both strategies perform quite good (over 76% for the semantic-based relatedness and over

Table 2. Precision @k for the documents relatedness recommendations in Bibsonomy dataset (left table). Precision @1 manually evaluated for the documents relatedness recommendations in USAToday dataset (right table).

Bibsonomy				
Precision	@1	@2	@5	@10
Semantic-based	0.78	0.86	0.82	0.76
Crowd-sourced	0.70	0.70	0.74	0.70

USAToday	
Precision	@1
Semantic-based	0.76
Crowd-sourced	0.65

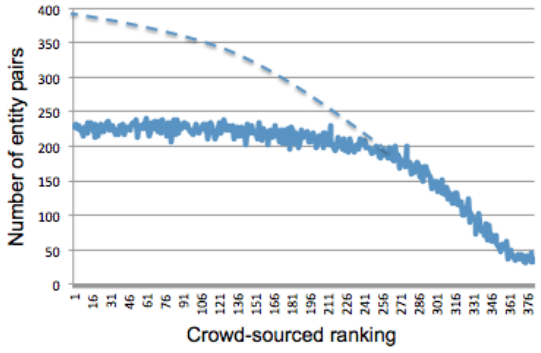

Fig. 1. The X-axis represents the ranking position x of entity pairs according to our crowd-sourced connectivity rankings. The Y-axis represent the number of entity pairs ranked at xth position that have a semantic relation according to our connectivity threshold.

65% for the crowd-sourced), giving us the proof of concept that, both strategies can be used for suggesting related items, even though the top related items are not the same.

For the Bibsonomy dataset we performed an automatic evaluation to access the precision of the document placement regarding the tag assignments. As explained in Section 4.3, we assumed the tag assignments as the ground truth for document relatedness [12]. Table 2 exposes the results for precision of the recommended documents considering the top k results. Both methods reached over 70% of performance that demonstrates their significant potential.

After computed all semantic and crowd-sourced scores between the pairs of entities, we obtained for each entity two ordered lists ranking all entities (m) according to each score. In Fig. 1, we represent data generated based on the USAToday dataset. The X-axis represents particular sets of entity-pairs ordered according to their connectivity ranking achieved based on the crowd-sourced activity ranking. The (x) value denotes all entity pairs (mx) which are ranked at the xth position in each particular entity ranking list. The Y-axis represent the number of entity pairs (nx) that have a semantic relation according to our semantic connectivity scores (solid line, $(\lambda(\delta_{(e_i,e_j)}^{<max(l)>}) > 0)$) within the particular set mx at xth ranking position.

Ideally, we expect that for every entity pair ranked at the top position (left on X-axis), would exist some semantic relation. The plot shows that for the top 1 crowd-sourced pairs, we found around 225 pairs that have such relation.

In this sense, the dotted line represents the ideal result. From these results, we can deduct that the pairs that are in between the area below the dotted line and above the solid line are most probably missing some semantic relation. Identifying the correct items that have some missing relations is the first step for the task of actually discovering which ones are the exactly missing relations. Complementary, by observing the missing semantic ranked pairs on the *X*-axis, we can identify which entities miss some relation given by the crowd-sourced. It is worth noting that since the 260th rank position in the *X*-axis, the behavior of the curve are in line with our expectations, i.e., the lower the correlation between crowd-sourced, the lower is the semantic connectivity.

As for a qualitative analysis of the document relatedness evaluation, we picked up a document (i) from the USAToday corpus and its most related document (ii) according to our semantic-based approach. The underlined terms refer to the recognized entities in each document derived from the entity recognition and enrichment process (see Section 3.1).

(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 related to basketball, a linguistic approach would fail to point out both documents as related. First, both documents have too short descriptions, which make it harder for a linguistic approach to detect their similarity. Second, in this particular case, there are no significant common words between the documents. However, by applying our semantic-based approach, it is possible to measure a score of connectivity between both documents. For example, once the term *Charlotte Bobcats* was enriched by the entity `http://dbpedia.org/resource/Charlotte_Bobcats` in the document (i) and the term *New York Knicks* was enriched by the entity `http://dbpedia.org/resource/New_York_Knicks` in the document (ii), a semantic score is assigned to each pair of entities found to generate an overall score of connectivity between both documents.

6 Related Work

The approach of applying actor/network theory to data graphs has been discussed by Kaldoudi et al. [8]. Graph summarization is a very interesting approach to exploit semantic knowledge in annotated graphs. Thor et al. [19] exploited this technique for link prediction between genes in the area of Life Sciences. Their approach relies on the fact that the graph summarization techniques create compact representations of the original graph adopting some criteria for the creation, correction and deletion of edges and for grouping nodes. Thus, a prediction function ranks the most potential edges and then suggests possible links between two given genes.

Another approach to identify potential links between nodes is presented by Potamias et al. [13], where they describe an algorithm based on Dijkstra's shortest path along with random walks in probabilistic graphs to define distance functions that identifies the k closest nodes from a given source node. Lehmann et al. [10] introduces the RelFinder

that is able to show relationships between two different objects in DBpedia. Their approach is based on the breadth-first search algorithm, which is responsible for finding all related objects in the tripleset. Then, the information gathered is stored in a relational database for further querying and visualization. In this work, we use the RelFinder approach to exploit the relationship between objects (see Section 3.2). Contrasting with RelFinder, Seo et al. [15] proposed the OntoRelFinder that uses a RDF Schema for finding the relationships between two objects through its class relationships.

An interesting work in social networks is also presented by Leskovec et al. [11]. Their technique suggests positive and negative relationships between people in a social network. The notion of negative and positive relationships is also addressed in our method, but taking into account the length of the paths, as aforementioned. Similarly, Xiang et al. [22] present a work based on the homophily principle (i.e., people tend to associate and interact with people with similar characteristics) to estimate relationship strength between people. For this, they present an unsupervised model that takes into account the shared attributes and interactions between individuals in a social network. This approach meets our assumptions that the closer two objects are, the higher is the proximity between them.

Finding semantic associations between two given objects is also discussed in the context of ontology matching [6, 20, 23]. In our case, hub ontologies could also be used to infer missing relationships into another ontology.

Contrasting with the approaches just outlined, we combine different techniques to uncover relationships between disparate entities, which allows us to exploit the relationships between entities to identify correlated Web resources.

7 Discussion and Outlook

We have presented a general-purpose approach to discover relationships between Web resources based on the relationships between extracted entities together with an evaluation and discussion of experimental results. We found that, uncovering relationships between data entities helps to detect correlations of documents that, a priori, linguistic approaches would not reveal. Linguistic methods are based on the co-occurrence of words in a set of documents, while our semantic-based approach relies on semantic relations between entities as represented in reference datasets. A hybrid approach would overcome this deficiency. However, in cases where extracted entities have to be matched, term frequency or linguistic similarity-based approaches cannot be applied. An interesting application of our work lies in document and data clustering which can be exploited, for instance, for entity based document recommenders.

During our evaluation experiments, we achieved an average of 80% of precision for the Bibsonomy dataset when suggesting the most related documents given one document (top 1, top 2, top 5, top 10), while for the USAToday dataset we achieved 0.76% of precision. We also presented a crowd-sourced strategy that takes into account the co-occurrence of entities in Web searches, thus relying on the wisdom of crowds. This approach achieved an average of 71% of precision for the Bibsonomy dataset, while the USAToday presented 65% of precision. This leads to the conclusion that both produce fairly good indicators for document relatedness.

Although both approaches have achieved good results, an evaluation based on the Kendall's tau rank correlation has shown that both differ in the relationships they uncovered. Naturally, the numbers presented by the Kendall's tau evaluation are subjected to noise caused by *misannotated* entities during the NER/enrichment process and the approximated values given by the search engine. Nevertheless, we believe that these proposed evaluations are the first step to identify missing connections in a semantically enriched dataset. Finally, we deduct that each strategy is complementary to each other. Semantically deducted relations are able to find connections between entities that do not necessarily often co-occur in contrast to the crowd-sourced analysis based on co-occurrence.

The main issues faced during the experimental work were the low performance and accuracy of the NER tools at hand, and high computational demands when applying our relatedness computation to larger amounts of data. That restricted our experiments to a limited dataset. Moreover, one of the key weaknesses of the Katz index in the context of our work is the fact that it treats all edges equally. Thus, when applying it to Linked Data graphs, valuable semantics about the meaning of each edge (i.e., property) is not considered during the relatedness computation. We are currently investigating ways to extend the Katz index by distinguishing between different property types. Hence, future work will plan to (a) apply weights to different path types between the entities according to the semantics of the properties they represent in order to provide a more refined score; and (b) investigate means to combine our two complementary relationship discovery approaches.

8 Acknowledgement

This work has been partially supported by CAPES (Process n^o 9404-11-2) and the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement No 270239 (ARCOMEM), CNPq grants 301497/2006-0 and 475717/2011-2 and FAPERJ grants E-26/103.070/2011.

References

1. Anyanwu, K., Sheth, A.: p-queries: enabling querying for semantic associations on the semantic web. In: Proceedings of the 12th international conference on World Wide Web. pp. 690 – 699. ACM Press New York, NY, USA, Budapest, Hungary (2003)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS) 5(3), 1–22 (2009)
3. Carmel, D., Cohen, D., Fagin, R., Farchi, E., Herscovici, M., Maarek, Y.S., Soffer, A.: Static index pruning for information retrieval systems. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 43–50. SIGIR '01, ACM, New York, NY, USA (2001), http://doi.acm.org/10.1145/383952.383958
4. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (2002)

5. Dietze, S., Yu, H., Giordano, D., Kaldoudi, E., Dovrolis, N., Taibi, D.: Linked education: interlinking educational resources and the web of data. In: Proceedings of the 27th ACM Symposium On Applied Computing, Special Track on Semantic Web and Applications. SAC '12, ACM, New York, NY, USA (2012)
6. Groß, A., Hartung, M., Kirsten, T., Rahm, E.: Mapping Composition for Matching Large Life Science Ontologies. In: Proceedings of the 2nd International Conference on Biomedical Ontology. ICBO 2011 (2011)
7. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl: sameas isn't the same: an analysis of identity in linked data. In: Proc. of the 9th International Semantic Web Conference,Vol. Part I. pp. 305–320. Berlin, Heidelberg (2010), http://dl.acm.org/citation.cfm?id=1940281.1940302
8. Kaldoudi, E., Dovrolis, N., Dietze, S.: Information organization on the internet based on heterogeneous social networks. In: Proceedings of the 29th ACM international conference on Design of communication. pp. 107–114. SIGDOC '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2038476.2038496
9. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), 39–43 (March 1953), http://ideas.repec.org/a/spr/psycho/v18y1953i1p39-43.html
10. Lehmann, J., Schppel, J., Auer, S.: Discovering unknown connections - the DBpedia relationship finder. In: Proceedings of 1st Conference on Social Semantic Web. Leipzig (CSSW07), 24.-28. September. Lecture Notes in Informatics (LNI), vol. P-113 of GI-Edition. Bonner Kllen Verlag (September 2007), http://www.informatik.uni-leipzig.de/~auer/publication/relfinder.pdf
11. Leskovec, J., Huttenlocher, D.P., Kleinberg, J.M.: Predicting positive and negative links in online social networks. CoRR abs/1003.2429 (2010), http://dblp.uni-trier.de/db/journals/corr/corr1003.html#abs-1003-2429
12. Peters, I., Haustein, S., Terliesner, J.: Crowdsourcing in article evaluation. In: Proceedings of the 3rd ACM International Conf. on Web Science. pp. 1–4. Koblenz, Germany (2011), http://journal.webscience.org/487/
13. Potamias, M., Bonchi, F., Gionis, A., Kollios, G.: k-nearest neighbors in uncertain graphs. PVLDB 3(1), 997–1008 (2010), http://dblp.uni-trier.de/db/journals/pvldb/pvldb3.html#PotamiasBGK10
14. Risse, T., Dietze, S., Peters, W., Doka, K., Stavrakas, Y., Senellart, P.: Exploiting the social and semantic web for guided web archiving. In: Proceedings of the International Conference on Theory and Practice of Digital Libraries 2012. TPDL '12, Springer LNCS (2012)
15. Seo, D., Koo, H., Lee, S., Kim, P., Jung, H., Sung, W.K.: Efficient finding relationship between individuals in a mass ontology database. In: Kim, T.H., Adeli, H., Ma, J., Fang, W.C., Kang, B.H., Park, B., Sandnes, F.E., Lee, K.C. (eds.) FGIT-UNESST, vol. 264. pp. 281–286. Communications in Computer and Information Science, Springer (2011), http://dblp.uni-trier.de/db/conf/fgit/unesst2011.html#SeoKLKJS11
16. Sheth, A., Aleman-Meza, B., Arpinar, I.B., Halaschek-Wiener, C., Ramakrishnan, C., Bertram, Y.W.C., Avant, D., Arpinar, F.S., Anyanwu, K., Kochut, K.: Semantic association identication and knowledge discovery for national security applications. Journal of Database Management 16(1), 3353 (2005)
17. Sheth, A.P., Ramakrishnan, C.: Relationship web: Blazing semantic trails between web resources. IEEE Internet Computing 11(4), 77–81 (2007), http://dblp.uni-trier.de/db/journals/internet/internet11.html#ShethR07
18. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A core of semantic knowledge. In: 16th international World Wide Web conference. ACM Press, New York, NY, USA (2007)
19. Thor, A., Anderson, P., Raschid, L., Navlakha, S., Saha, B., Khuller, S., Zhang, X.N.: Link prediction for annotation graphs using graph summarization. In: 10th International Confer-

ence on The Semantic Web, Vol. Part I. pp. 714–729. ISWC'11, Berlin, Heidelberg (2011), http://dl.acm.org/citation.cfm?id=2063016.2063062
20. Vidal, V.M.P., de Macedo, J.A.F., Pinheiro, J.C., Casanova, M.A., Porto, F.: Query processing in a mediator based framework for linked data integration. IJBDCN 7(2), 29–47 (2011), http://dblp.uni-trier.de/db/journals/ijbdcn/ijbdcn7.html#VidalMPCP11
21. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (Jun 1998), http://dx.doi.org/10.1038/30918
22. Xiang, R., Neville, J., Rogati, M.: Modeling relationship strength in online social networks. In: Rappa, M., Jones, P., Freire, J., Chakrabarti, S. (eds.) WWW. pp. 981–990. ACM (2010), http://dblp.uni-trier.de/db/conf/www/www2010.html#XiangNR10
23. Xu, L., Embley, D.W.: Discovering direct and indirect matches for schema elements. In: DASFAA. pp. 39–46. IEEE Computer Society (2003), http://dblp.uni-trier.de/db/conf/dasfaa/dasfaa2003.html#XuE03