Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 22 (2013) 231 – 240

Procedia Computer Science

17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013

Interlinking documents based on semantic graphs

Bernardo Pereira Nunes[a,b,*], Ricardo Kawase[b], Besnik Fetahu[b], Stefan Dietze[b], Marco A. Casanova[a], Diana Maynard[c]

[a]*Department of Informatics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro/RJ – Brazil, CEP 22451-900*
[b]*L3S Research Center, Leibniz University Hannover, Appelstr. 9a, 30167 Hannover, Germany*
[c]*Department of Computer Science, University of Sheffield, Sheffield, UK*

Abstract

Connectivity and relatedness of Web resources are two concepts that define to what extent different parts are connected or related to one another. Measuring connectivity and relatedness between Web resources is a growing field of research, often the starting point of recommender systems. Although relatedness is liable to subjective interpretations, connectivity is not. Given the Semantic Web's ability of linking Web resources, connectivity can be measured by exploiting the links between entities. Further, these connections can be exploited to uncover relationships between Web resources. In this paper, we apply and expand a relationship assessment methodology from social network theory to measure the connectivity between documents. The connectivity measures are used to identify connected and related Web resources. Our approach is able to expose relations that traditional text-based approaches fail to identify. We validate and assess our proposed approaches through an evaluation on a real world dataset, where results show that the proposed techniques outperform state of the art approaches.

Selection and peer-review under responsibility of KES International

Keywords: Document connectivity, semantic connections, semantic graphs, document recommendation

1. Introduction

User-generated content is characterized by a high degree of diversity and heavily varying quality. Given the ever increasing pace at which this form of Web content is evolving, adequate preservation and detection of correlations has become a cultural necessity. Extraction of entities from Web content, in particular social media, is a crucial challenge in order to enable the interlinking of related Web content, semantic search and navigation within Web archives, and to assess the relevance of a given set of Web objects for a particular query or crawl. As part of earlier work, we have developed a processing chain dealing with entity extraction and enrichment, consisting of a set of dedicated components which handle named entity recognition (NER) and consolidation (enrichment, clustering, disambiguation) as part of one coherent workflow (see [3] for more details).

Traditional approaches to finding related Web resources (e.g. documents) are often addressed using a combination of Information Retrieval (IR) and Natural Language Processing (NLP) techniques. These techniques

*Corresponding author.
E-mail address: bnunes@inf.puc-rio.br.

Selection and peer-review under responsibility of KES International
doi:10.1016/j.procs.2013.09.099

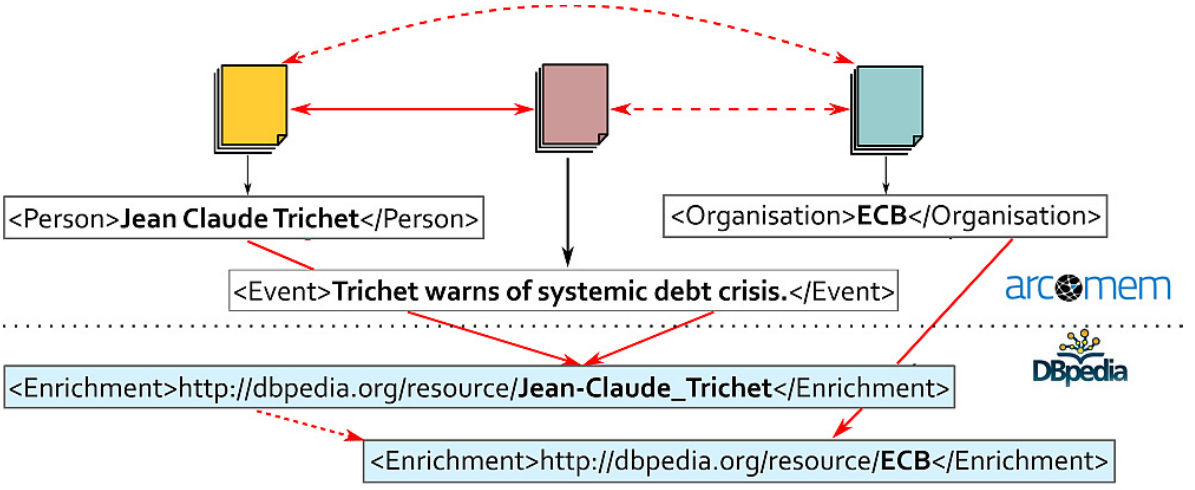

Fig. 1: Example: connections between Web documents, extracted entities and DBpedia enrichments within ARCOMEM dataset.

compute the similarities between a set of terms from specific resources based on their overlap, or through latent semantic analysis [4] measuring relatedness of individual terms and resources. Nonetheless, most of these techniques require large corpora and a partially common vocabulary/terminology between the resources. Thus, in such cases, they fail to detect latent semantic relationships between resources.

On the other hand, semantic approaches exploit knowledge defined in a data graph to compute notions of similarity and connectivity [18]. Our approach explicitly targets *connectivity* as a measure of the *relationship* between two Web resources, as opposed to their *similarity*.

An example is derived from datasets specific to the ARCOMEM project[1], which primarily consist of extracted information about events and entities (see [3]). ARCOMEM follows a use case-driven approach based on scenarios aimed at creating focused Web archives, particularly of social media, by adopting novel entity extraction and interlinking mechanisms. These archives deploy a document repository of crawled Web content and a structured RDF knowledge base containing metadata about entities and events detected in the archived content.

For instance, Figure 1 shows three sets of Web resources (depicted at the top), each associated with one particular entity/event, where the entity ("Jean Claude Trichet") and event ("Trichet warns of systemic debt crisis") are both enriched with the same DBpedia[2] entity (`http://dbpedia.org/resource/Jean-Claude-Trichet`). This allows us to cluster the respective entity and event, and their connected Web resources, as an example of direct connection (solid red line in the diagram). However, the third set of Web resources is connected with a third entity ("ECB") which refers to the *European Central Bank*, enriched with the corresponding DBpedia resource (`http://dbpedia.org/resource/ECB`). While NLP and standard IR approaches would fail to detect a connection between them, analysing the DBpedia graph uncovers a close connection between ECB and *Jean Claude Trichet* (being a former ECB president), and hence allows us to establish a connection (dashed line) between all involved entities/events and their connected Web resources. Analysis of the reference data graph thereby allows us to identify implicit connections between entities and documents.

In this paper, we present a general-purpose approach to detect and measure semantic connectivity between entities within reference datasets as a means to compute connectivity between Web resources (documents) in disparate datasets and document corpora. Our semantic connectivity score is based on the Katz index [13], a score for measuring relatedness of actors in social networks, which has been adopted and expanded to take into account the semantics of data graphs.

[1] `http://www.arcomem.eu`
[2] `http://www.dbpedia.org/`

In previous works [16, 15], we have introduced the semantic connectivity score between entities and an in-depth analysis of how semantic graphs can be exploited to uncover latent connections between entities. In this paper, we extend the previous approach based on entity connectivity to find latent connections across documents.

The remainder of the paper is structured as follows. Section 2 presents an overview of related research. Section 3 introduces the semantic connectivity score between documents. Sections 4 and 5 show the evaluation method and the outcomes of our method. Finally, Section 6 presents the conclusion and future work.

2. Related Work

Kaldoudi et al. [12] discusses how to apply the overall approach of actor/network theory to data graphs. Graph summarization is an 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 summarisation techniques can create compact representations of the original graph, by adopting a set of criteria for creation, correction and deletion of edges and grouping of nodes. Thus, a prediction function ranks the edges with the most potential, and then suggests possible links between two given genes.

Potamias et al. [17] presents another approach based on Dijkstra's shortest path along with random walks in probabilistic graphs to define distance functions that identify the k closest nodes from a given source node.

Lehmann et al. [11] introduces RelFinder, which shows semantic associations between two different entities from RDF datasets, based on a breadth-first search algorithm responsible for finding all related entities in the tripleset. In this work, we use the RelFinder approach to exploit the connectivity between entities.

In the field of Social Networks, Hasan and Zake [10] present a survey of link prediction techniques, where they classify the approaches into the following categories: *feature based link prediction*, *bayesian probabilistic models*, *probabilistic relational models* and *linear algebraic methods*. According to this classification, our approach can be classified as a *feature based link prediction* method. Work from Leskovec et al. [14] 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 as mentioned previously.

Finding semantic relationships between two given entities is also discussed in the context of ontology matching [9, 20, 21]. In our case, hub ontologies could also be used to infer missing relationships into another ontology.

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 connections between Web resources.

3. Document Connectivity

In this section we present the main steps of the process chain of our approach. The whole process is composed of four steps, described as follows:

S1. *Entity Extraction* – pre-processing of documents for finding and extracting term references and named entities;
S2. *Entity Enrichment* – matching of references in external knowledge bases such as DBpedia and Freebase[3];
S3. *Entity Connectivity* – uncovering of latent relationships between entities and induction of connections amongst entities.
S4. *Document Connectivity* – uncovering latent relationships between documents through entity connections and inducing connections amongst documents.

Steps 1-3 have been introduced in our previous work [16, 15] and therefore, in this section, we focus on Step 4, the contribution of this work, in which we discover latent connections between documents. However, Step 3, defined in our previous works, is of paramount importance in order to fully understand Step 4.

[3] http://www.freebase.com

3.1. A novel approach to document connectivity

In this section, we define a document connectivity score which relies on connections between entities based on reference graphs. Before introducing the document connectivity approach, we recall how Step 3 (described in [15]) uncovers latent connections between entities which our approach builds upon.

3.1.1. Entity Connectivity

As the main goal of this work is to uncover latent information between documents, we first exploit the content of the documents to find connections between terms and entities that occur in the documents that would in turn induce connections between the documents themselves. For this, we first process the documents to find and extract term references and named entities, and then enrich these mentions using reference datasets (e.g. DBpedia).

Assuming that this process is already solved by previous approaches, we stick to the problem of finding latent connections between entities. We briefly introduce the semantic connectivity score (SCS_e) responsible for discovering latent connections between entity pairs. SCS_e is based on the Katz index [13] which is used to estimate the relatedness of actors in social networks. To adapt the Katz index for finding latent connections between entity pairs in large graphs, we have applied three main adaptions described as follows:

1. *Maximum path length*: Traversing large graphs is computationally expensive and the computation of all paths between entity pairs is computationally intractable. Thus, to make our approach feasible, we restrict the computation of paths between entity pairs with a maximum path length of four intermediate edges (links) in-between.

   Note that the maximum path length exploited was previously determined after comprehensive tests presented in [15], and also adopted in [5].

2. *Undirected graphs*: Reference graphs like DBpedia and Freebase have object properties that are often found in their inverse form. For instance, as described in [8], the property *fatherOf* is the inverse property of *sonOf*. Thus, we explore connectivity between entity pairs without taking into account the edge direction. Hence, the semantic connectivity scores between entities are the same for both directions.

3. *Transversal paths*: As described in [2], we distinguish relation types found in reference graphs as *hierarchical* and *transversal*. Concisely, *hierarchical* relations indicate similarity through typical hierarchical relations between entity pairs. Examples of hierarchical relations are: `rdfs:subclassOf`, `dcterms:subject` and `skos:broader`. Unlike *hierarchical* relations, *transversal* relations indicate entity connectivity independent of their similarity, i.e. non-hierarchical relations. Thus, to compute the semantic connectivity score between entity pairs we consider only transversal relations. An example of transversal relation is given by the entity pairs *"Jean Claude Trichet"* and *"European Central Bank"* introduced in Section 1, where the *"European Central Bank"* is linked to the entity *"President of the European Central Bank"* through the transversal RDF property `http://dbpedia.org/property/leaderTitle` that, for its part, links to *"Jean Claude Trichet"* through another transversal RDF property `http://dbpedia.org/property/title`.

Having introduced and defined the scenario in which we compute the *semantic connectivity score* (SCS_e) between an entity pair (e_1, e_2), we now present the Equation:

$$SCS_e(e_1, e_2) = 1 - \frac{1}{1 + (\sum_{l=1}^{\tau} \beta^l \cdot |paths^{<l>}_{(e_1,e_2)}|)} \quad (1)$$

where $|paths^{<l>}_{(e_1,e_2)}|$ is the number of *transversal* paths of length l between entities e_1 and e_2, τ is the maximum length of paths considered (in our case $\tau = 4$), 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 is to the final score. Obviously, if the damping factor is 1, all paths will have the same weight independent of the length. In previous experiments, we observed that $\beta = 0.5$ achieved better results in terms of precision [16]. Equation 1 is normalised to range between $[0, 1)$.

Returning to the example presented in Section 1, we compute the semantic connectivity score for the entities *"Jean Claude Trichet"* (JCT) and *"European Central Bank"* (ECB), using DBpedia as the reference tripleset. Omitting the details, let us assume that we obtained 8 paths of length 2, and 14 paths of length 3, resulting in the following score:

$$SCS_e(JCT, ECB) = 1 - \frac{1}{1 + (0.5^2 \cdot 8 + 0.5^3 \cdot 14)} = 1 - \frac{1}{1 + (2 + 1.75)} = 0.79 \quad (2)$$

Note that even for a small number (i.e., 8) of short paths (of length 2), the contribution to the overall score is larger than for longer paths (of length 3). Evidently, the score obtained by a longer path can overcome a shorter path depending on the number of paths found and the damping factor assigned.

3.1.2. Document Connectivity

Based on the semantic connectivity score between entity pairs (SCS_e), we then define the *semantic connectivity score* (SCS_w) between two Web resources W_1 and W_2 as follows:

$$SCS_w(W_1, W_2) = \begin{cases} 0, \text{ iff } |E_1| = 0 \text{ or } |E_2| = 0 \\ \left(\sum_{\substack{e_1 \in E_1 \\ e_2 \in E_2 \\ e_1 \neq e_2}} SCS_e(e_1, e_2) + |E_1 \cap E_2| \right) \cdot \frac{1}{|E_1| \cdot |E_2|}, \text{ otherwise} \end{cases} \quad (3)$$

where E_i is the set of entities found in W_i, for $i = 1, 2$. Note that the score is normalised between [0,1]. The score $SCS_w(W_1, W_2)$ is 0 when no connection between entity pairs across documents exists or iff $|E_1| = 0$ or $|E_2| = 0$. Otherwise, the score is represented by the sum of semantic connectivity scores between entities, normalised over the total number of entity pair comparisons.

To illustrate the semantic connectivity score between document pairs, we present two descriptions of documents extracted from the USAToday[4] corpus. We observe that the underlined terms are entities previously recognised through the entity recognition and enrichment process (S.1 and S.2).

(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.

Thus, for each entity in document (i) and document (ii), we compute the semantic connectivity score (SCS_e) between entities. Table 1 summarises the scores between entity pairs between documents (i) and (ii).

Table 1: Semantic connectivity scores between entity pairs in document (i) and (ii).

Entities from document (i)	Entities from document (ii)	SCS_e
Charlotte Bobcats	*New York Knicks*	0.87
Charlotte Bobcats	*Carmelo Anthony*	0.63
Charlotte Bobcats	*Amar'e Stoudemire*	0.60
Charlotte Bobcats	*Miami Heat*	0.89
NBA	*New York Knicks*	0.85
NBA	*Carmelo Anthony*	0.60
NBA	*Amar'e Stoudemire*	0.63
NBA	*Miami Heat*	0.87

[4] http://www.usatoday.com

Thus, the final score between the documents (i) and (ii) is:

$$SCS_w(W_1, W_2) = \frac{(0.87 + 0.63 + 0.60 + 0.89) + (0.85 + 0.60 + 0.63 + 0.87)}{2 \cdot 4} = \frac{5.96}{8} = 0.74 \quad (4)$$

4. Evaluation Method

In this section, we describe in detail the evaluation methodology and experiment setup used to validate our hypothesis of uncovering latent relationships between Web resources (*entities* and *documents*) using the semantic connectivity score SCS_w.

4.1. Dataset

The dataset used to evaluate our approach consists of a subset of randomly selected news articles (documents) from the USAToday news Website. In total, we consider document connectivity for 40,000 document pairs. Each document contains a title and a summary, where the latter is 200 characters long on average. We performed the *entity extraction* step using DBpedia Spotlight[5]. The resulting set of annotations consists of approximately 80,000 entity pairs.

4.2. Gold standard

In order to validate the results of our evaluation, the first step is to obtain a ground truth of relationships between documents. Given the lack of such benchmarks, we conducted a user evaluation to collect user judgements with the aim of creating a gold standard. The user evaluation was set up in CrowdFlower[6], a crowdsourcing platform. In order to construct the gold standard, we randomly selected 600 document pairs to be evaluated. The evaluation process consisted of a questionnaire on a 5-point Likert scale model where participants were asked to rate their agreement of the suggested semantic connection between a given document pair.

Additionally, we inspected participants' expectations regarding declared connected document pairs. In this case, presenting two documents deemed to be connected, we asked participants if such connections were expected (from *extremely unexpected* to *extremely expected*, also on a 5-point Likert scale). The judgements collected provide us with a gold standard for the analysis of our techniques. Note that in 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.

4.3. Evaluation Methods

To emphasise the benefits of measuring connectivity between documents using our approach, we compared it against competing methods which measure connectivity via co-occurrence-based metrics to detect entity and document connectivity. In the first evaluation, we compared the performance of SCS_w against two methods outlined below.

Co-occurrence-based method (CBM) is a co-occurrence-based score between entities that relies on an approximation of the number of existing Web pages that contain these entities. For example, Nunes et al. [16] estimates the co-occurrence score of entity pairs by issuing 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. We interpret a large number of pages as an indicator of high connectivity, and a small number of pages as an indicator of low connectivity between the queried terms (which represent entities in our case). Besides CBM, there are other similar approaches to quantify the relatedness between entities, such as Pointwise

[5] http://spotlight.dbpedia.org/
[6] https://www.crowdflower.com/

Mutual Information (PMI)[1] and Normalised Google Distance (NGD)[7]. However, they take into account the joint distribution and the probability of their individual distributions, which requires knowing a priori the total number of Web pages searched by a search engine. Thus, in this case, the document connectivity score is given by a small adjustment in Equation (3) where, instead of SCS_w, we use CBM.

Explicit Semantic Analysis (ESA) proposed by Gabrilovich and Markovitch [6] measures the relatedness between Wikipedia[7] 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.

In order to evaluate the document connectivity, we compared our method with the traditional statistical *tf-idf* method, in addition to ESA and CBM. As mentioned, the latter method was slightly modified to measure the connectivity between documents, where in Equation (3) we replaced the semantic score with the co-occurrence-based score.

4.4. Evaluation Metrics

For measuring the performance of the doccument connectivity approaches, we used standard evaluation metrics like precision (P), recall (R) and $F1$ measure. Note that in these metrics, as relevant pairs, we consider those marked in the gold standard (gs) as connected according to the 5-point Likert Scale (*Strongly Agree* & *Agree*).

For the document connectivity, the precision measure (P_w) is the ratio of the set of all retrieved document pairs deemed as relevant over the set of document pairs that are connected. Thus, the relevant documents are those that were marked as *Strongly Agree* & *Agree*, while the set of document pairs that are connected consists of those that have a semantic connectivity score greater than a given threshold (see Equation (5)).

$$P_w = \begin{cases} 0, & \text{iff } |\Phi^{\tau}_{retrieved}| = 0 \\ \dfrac{|\Phi^{\tau}_{retrieved} \cap \Phi_{relevant}|}{|\Phi^{\tau}_{retrieved}|}, & \text{otherwise} \end{cases} \tag{5}$$

where $\Phi_{relevant}$ is the set of retrieved document pairs that are relevant and $\Phi^{\tau}_{retrieved}$ is the set of all connected document pairs greater than a given threshold (τ).

The recall (R_w) is the ratio of the set of retrieved documents that are relevant over the set of all relevant document pairs according to the gold standard (see Equation (6)).

$$R_w = \begin{cases} 0, & \text{iff } |\Phi_{relevant_{(gs)}}| = 0 \\ \dfrac{|\Phi^{\tau}_{retrieved} \cap \Phi_{relevant}|}{|\Phi_{relevant_{(gs)}}|}, & \text{otherwise} \end{cases} \tag{6}$$

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

Finally, $F1$ measure shows the balance between precision and recall, and is computed as in Equation (7).

$$F1_w = \begin{cases} 0, \text{ iff } (P_w + R_w) = 0 \\ 2 \cdot \dfrac{P_w \cdot R_w}{P_w + R_w}, \text{ otherwise} \end{cases} \tag{7}$$

5. Results

In this section, we report evaluation results for the document connectivity approaches. For each method, we present the results for their ability to discover latent connections between pairs of resources. Furthermore, we also present an in-depth analysis of their shortcomings and advantages for discovering connections between documents.

[7] http://www.wikipedia.org

5.1. Document connectivity results

Table 2 shows the results according to the gold standard presented in the Likert scale, where users evaluated if a given entity pair could be connected in a document. Compared with the gold standard, 368 entity pairs out of 812 could have some connection.

From the set of entities that could co-occur in a document, 51% of those entities were also connected based on our gold standard, while 34% were *Undecided*. Analysis of the results for the *Undecided* category will be provided in Section 5.2, since these results are of particular interest in establishing latent relationships between Web resources.

Table 2: Total number of results for the GS in Likert-scale.

Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
96	272	139	165	140

The performance of each method is shown in Table 3. As in the task of entity connectivity, SCS_w performs slightly better than CBM in terms of precision, while CBM is better in terms of recall. $F1$ measure is similar, with 60.0% and 59.6% for SCS_w and CBM, respectively. In both cases, ESA has the lowest performance.

The positive correlation of entity connectedness and their co-occurrence in the same document was 79.6%, 78.0% and 23.5% for SCS_w, CBM and ESA respectively, considering only the *Strongly Agree* and *Agree* relevance judgement results.

As already indicated in the introduction in Section 1, our proposed semantic approach can be exploited to measure document connectivity by taking into account the connectedness of entities that describe a document and their semantic connections. Indeed, as shown by the positive correlation of entity connectivity and entity co-occurrence in a document, we claim that our approach can be used as method for inferring document "relatedness" where other statistical models would fail.

To validate the usefulness of our approach, we compared the results against the well established document relatedness measure *tf-idf*. Our approach was able to find 500 unique connections between documents, whereas *tf-idf* found only 25. As described in Section 4.1, our corpus is composed of small descriptions of the news articles, which severely limits the ability of *tf-idf* to identify connections between them.

We also conducted an experiment to evaluate the uncovered connections by both methods. We found that 16% of the connections found by our approach were relevant, compared with 12% using *tf-idf*. We took into consideration that the recall achieved by *tf-idf* is only 3.6%, whereas for SCS_w, it is close to 86%.

5.2. Analysis of the Results

Table 3 shows the results for the task of document connectivity. The mixed approach $CBM{+}SCS_w$ performs best on finding the co-occurrence of entity pairs in a document. It is worth noting as well that the co-occurrence of entity pairs for documents can be retrieved with high recall (90%) when using the proposed combination of $CBM{+}SCS_w$.

A positive correlation of entity connectivity and co-occurrence in a document is of high importance for our proposed approach, allowing to establish newly constructed knowledge that can be represented as an aggregate of the entity connections.

We would also like to point out the challenges posed by our approach to creating a gold standard. As previously mentioned, while our work aims at detecting semantic connectivity of entities beyond traditional co-occurrence based approaches, this results in connections that 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, might impact 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 meaningful connections between entity pairs. This confirms that in a number of cases connections

Table 3: Precision, recall and F1 measure amongst methods.

	CBM	SCS_w	ESA	CBM+SCS_w
Precision	0.47	0.49	0.21	0.51
Recall (GS)	0.80	0.77	0.25	0.89
Recall	0.49	0.48	0.15	0.54
F1 (GS)	0.59	0.60	0.23	0.64
F1	0.48	0.48	0.18	0.52

were not necessarily incorrect, but simply presented information that was unknown to the users. Thus, we believe that a more thorough evaluation providing the evaluators with information on how a connection emerged, where we show all properties and entities that are part of a path greater than one, would give more reliable judgements.

6. Conclusion and future work

We have presented a general-purpose approach to discover and quantify document connectivity. To compute document connectivity, we first introduced a semantic-based entity connectivity approach (SCS_e) which adapts a measure from social network theory (Katz) to data graphs, in particular Linked Data, and extended it to interlink documents (SCS_w). SCS_w was able to uncover 16% of unique inferred document connections based on entity co-occurrence, not found by the state of the art method CBM. Additionally, while using a combination of CBM+SCS_w we achieved an $F1$ measure of 52%.

Our experiments show that SCS_w enables the detection and establishment of document connectivity that a priori linguistic and co-occurrence approaches would not reveal. Contrary to the latter, our approach relies on semantic relations between entities as represented in structured background knowledge, available via reference datasets. A combination of our semantic approach and traditional co-occurrence-based measures provided very promising results for detecting connected documents. While both approaches (CBM and SCS_w) produce fairly good indicators for document connectivity, an evaluation based on Kendall's τ rank correlation showed that the approaches differ in the relationships they uncover [16]. A comparison of agreement and disagreement between different methods revealed that both approaches are complementary and produce particularly good results when combined: 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 emphasize entity connections between entities that are not necessarily strongly connected in reference datasets.

As for future work, we aim to apply weights to different edge/property types according to their inherent semantics in order to provide a more refined score and to investigate means to combine our complementary approaches.

7. Acknowledgements

This work has been partially supported by CAPES (Process n^o 9404-11-2), the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No 270239 (ARCOMEM) and No 317620 (LinkedUp) as well as CNP under grants 301497/2006-0 and 475717/2011-2 and by FAPERJ under grants E-26/103.070/ 2011.

References

[1] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, Mar. 1990.
[2] D. Damljanovic, M. Stankovic, and P. Laublet. Linked data-based concept recommendation: Comparison of different methods in open innovation scenario. In *ESWC*, pages 24–38, 2012.
[3] S. Dietze, D. Maynard, E. Demidova, T. Risse, W. Peters, K. Doka, and Y. Stavrakas. Entity extraction and consolidation for social web content preservation. In A. Mitschick, F. Loizides, L. Predoiu, A. Nürnberger, and S. Ross, editors, *SDA*, volume 912 of *CEUR Workshop Proceedings*, pages 18–29. CEUR-WS.org, 2012.
[4] S. T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004.

[5] 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.

[6] 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.

[7] 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.

[8] A. Graves, S. Adali, and J. Hendler. A method to rank nodes in an rdf graph. In C. Bizer and A. Joshi, editors, *International Semantic Web Conference (Posters & Demos)*, volume 401 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[9] A. Groß, M. Hartung, T. Kirsten, and E. Rahm. Mapping Composition for Matching Large Life Science Ontologies. In *Proceedings of the 2nd International Conference on Biomedical Ontology*, ICBO 2011, 2011.

[10] M. A. Hasan and M. J. Zaki. A survey of link prediction in social networks. In C. C. Aggarwal, editor, *Social Network Data Analytics*, pages 243–275. Springer, 2011.

[11] P. Heim, S. Hellmann, J. Lehmann, S. Lohmann, and T. Stegemann. RelFinder: Revealing relationships in RDF knowledge bases. In *Proceedings of the 3rd International Conference on Semantic and Media Technologies (SAMT)*, volume 5887 of *Lecture Notes in Computer Science*, pages 182–187. Springer, 2009.

[12] E. Kaldoudi, N. Dovrolis, and S. Dietze. Information organization on the internet based on heterogeneous social networks. In *Proceedings of the 29th ACM international conference on Design of communication*, SIGDOC '11, pages 107–114, New York, NY, USA, 2011. ACM.

[13] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953.

[14] 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.

[15] B. Pereira Nunes, S. Dietze, M. A. Casanova, R. Kawase, B. Fetahu, and W. Nejdl. Combining a co-occurrence-based and a semantic measure for entity linking. In *ESWC*, 2013 (to appear).

[16] 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 Web of Linked Entities Workshop in conjuction with the 11th International Semantic Web Conference*, volume 906 of *CEUR-WS.org*, pages 45–57, Nov. 2012.

[17] M. Potamias, F. Bonchi, A. Gionis, and G. Kollios. k-nearest neighbors in uncertain graphs. *Proc. VLDB Endow.*, 3(1-2):997–1008, Sept. 2010.

[18] A. Sheth, B. Aleman-Meza, F. S. Arpinar, A. Sheth, C. Ramakrishnan, C. Bertram, Y. Warke, K. Anyanwu, B. Aleman-meza, I. B. Arpinar, , K. Kochut, C. Halaschek, C. Ramakrishnan, Y. Warke, D. Avant, F. S. Arpinar, K. Anyanwu, and K. Kochut. Semantic association identification and knowledge discovery for national security applications. *Journal of Database Management*, 16:33–53, 2005.

[19] A. Thor, P. Anderson, L. Raschid, S. Navlakha, B. Saha, S. Khuller, and X.-N. Zhang. Link prediction for annotation graphs using graph summarization. In *10th International Conference on The Semantic Web, Vol. Part I*, ISWC'11, pages 714–729, Berlin, Heidelberg, 2011.

[20] V. M. P. Vidal, J. A. F. de Macedo, J. C. Pinheiro, M. A. Casanova, and F. Porto. Query processing in a mediator based framework for linked data integration. *IJBDCN*, 7(2):29–47, 2011.

[21] L. Xu and D. W. Embley. Discovering direct and indirect matches for schema elements. In *DASFAA*, pages 39–46. IEEE Computer Society, 2003.