Measuring Entity Semantic Relatedness using Wikipedia

Liliana Medina[1], Ana L. N. Fred[2], Rui Rodrigues[3] and Joaquim Filipe[3]
[1]*INSTICC, Setúbal, Portugal*
[2]*Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, Portugal*
[3]*Escola Superior de Tecnologia, Instituto Politécnico de Setúbal, Setúbal, Portugal*

Keywords: Semantic Relatedness, Wikipedia, Ontological Entities.

Abstract: In this paper we propose a semantic relatedness measure between scientific concepts, using Wikipedia as an hierarchical taxonomy. The devised measure examines the length of Wikipedia category path between two concepts, assigning a weight to each category that corresponds to its depth in the hierarchy. This procedure was extended to measure the relatedness between two distinct concept sets (herein referred to as *entities*), where the amount of *shared nodes* in the paths computed for all possible concept sets is also integrated in a global relatedness measure index.

1 INTRODUCTION

The *semantic relatedness* of two concepts indicates how far apart these concepts are when represented in a conceptual network or taxonomy, by using in the computation of this distance value all the semantic relationships that exist between them(Ponzetto and Strube, 2007). These semantic relationships may be hierarchical (IS-A type, hypernymy, hyponomy), of equivalence (synonyms) or associative (cause/effect) (Jiang and Conrath, 1997).

Semantic relatedness computation techniques may be grouped into two categories:

1. Distributive measures, that rely on non-structured data, such as large corpora. The underlying hypothesis is that similar words appear in similar contexts, thus they must have similar meanings. These approaches capture relationships between words.
2. Measures based on structured databases, such as taxonomies or ontologies, where relationships between semantic concepts are captured.

This paper focuses on the second category of relatedness measures, using Wikipedia as a taxonomy.

Our proposed measure takes into account both the distance between concepts and the relationship of the concepts with others in the taxonomy. It is then generalized to measure the relatedness between *concept sets* or *entities*. As defined in (Rodrguez and Egenhofer, 2003), the term entity refers to groups of concepts or objects in the real world that are somehow semantically related.

Our goal is then to present a cost function that allows us to make a decision on similarity between two or more entities, based on their representations as concept sets.

As an ontological basis, we will use the english version of Wikipedia[1] to help establish semantic relationships between concepts and entities. In the recent years, Wikipedia has been explored as a potential knowledge base for a number of information retrieval tasks, such as text categorization, named entity recognition and semantic relatedness computation(Zesch et al., 2008).

The remaining sections of this paper are organized as follows: in Section 2 we describe related work in this area; Section 3 presents the proposed similarity measure, Section 4 deals with the results obtained after applying this measure to a set of entities. Finally, in Section 5 we draw the main conclusions and proposals of future work.

2 RELATED WORK

Given two words or expressions represented in a taxonomy, the computation of the semantic relatedness between these two objects may be transformed into the evaluation of their conceptual distance in the con-

[1]http://en.wikipedia.org

Medina L., L. N. Fred A., Rodrigues R. and Filipe J..
Measuring Entity Semantic Relatedness using Wikipedia.
DOI: 10.5220/0004180204310437
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (SSTM-2012), pages 431-437
ISBN: 978-989-8565-29-7

ceptual space generated by a taxonomy (Jiang and Conrath, 1997), being that each object is represented by a node in the resulting graph.

Semantic relatedness measures in hierarchical taxonomies can be categorized into three types (Slimani et al., 2006):

1. **Information Content or Node-based:** evaluation of the information content of a concept represented by a node such as described in (Resnik, 1999). The semantic relatedness between two concepts reflects the amount of shared information between them, generally in the form of their least common subsumer (LCS).
2. **Path or Edge-based:** evaluation of the distance that separates concepts by measuring the length of the edge-path between them (Wu and Palmer, 1994) (Rada et al., 1989). A weight is assigned to each edge, being that the weight computation must reflect some of the graph properties (network density, node depth, link strength, etc.) (Jiang and Conrath, 1997)
3. **Hybrid:** a combination of the former two (Jiang and Conrath, 1997) (Leacock and Chodorow, 1998).

Lexical databases, such as WordNet, have been explored as knowledge bases to measure the semantic similarity between words or expressions. However, WordNet provides generic definitions and a somewhat rigid categorization that does not reflect the intuitive semantic meaning that a human might assign to a concept.

In this paper we use the english version of the Wikipedia[2], a web-based encyclopedia which has approximately 4 million articles edited and reviewed by volunteers. The contributors are asked to assign these articles to one or more categories: Wikipedia may be thus viewed as either a folksonomy (Nastase and Strube, 2008) or a Collective Knowledge Base (Zesch et al., 2008), where human knowledge and human intuition on semantic relationships emerges in the form of a category network. It is then natural that this web-resource has been increasingly explored as a conceptual feature space, such that articles and categories are represented as nodes in the Wikipedia graph.

Techniques such Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) represent texts in the high-dimensional concept space of Wikipedia as weighted vectors. A textual fragment is thus considered as a weighted mixture of a predetermined set of "natural" concepts. Wikipedia Link-based Measure (WLM), first described in (Milne and Witten, 2008), uses its hyperlink structure, rather than the category hierarchy or textual content to compute semantic relatedness. In (Gouws et al., 2010) semantic relatedness is computed by spreading activation energy over the aforementioned hyperlink structure.

Our proposed measure takes into account not only the connections between nodes, but also the nature of nodes themselves by means of their location and connectivity degree in the overall category network. The measure takes into account only the categories and subcategories which encompass a given pair of concepts, discarding the actual textual content of the concepts article pages in the Wikipedia. We also generalize this approach to measure the semantic relatedness between sets of concepts.

Measurement of semantic similarity between concept sets can provide particular value for tasks concerning the semantics of entities (Liu and Birnbaum, 2007). An entity may represent, for instance, (1) an author, by means of his/her research interests, (2) a publication, such as a scientific journal, by means of its main topics, (3) a conference, by means of its submission topics. In Information Retrieval, the similarity between documents is generally estimated by means of their Vector Space Models. Each feature vector represents the bag-of-words of the respective document, assigning a weight to each feature/term that reflects its importance in the overall context of either the document or the document set. The definition of entity can also be extended to represent a document, where instead of a weighted feature vector, we have a set of terms that can be related to other entities (which may also be documents or other types of entities) by means of a semantic relatedness measure between entities, such as the one presented in this paper.

3 PROPOSED SIMILARITY RELATEDNESS MEASURE

The implementation of the proposed measure is based on the assumption that each pair of concepts is connected by a category path, such as the one depicted in Figure 1 for the pair of concepts "Feature Learning" and "Boosting". In this Figure we may observe that the least common subsumer (LCS) of both concepts is "Machine Learning".

The proposed relatedness measure is computed from the following sequence of steps

Distance between Concepts - Weighted Edges Sum. Let c_1 and c_2 be two concepts represented in the Wikipedia categories network. Find the shortest

[2] http://en.wikipedia.org

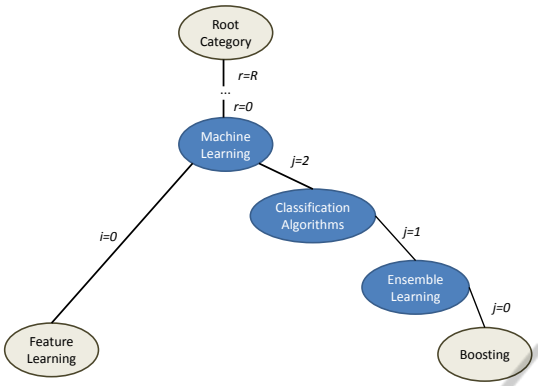

Figure 1: Category path between concepts *Feature Learning* and *Boosting*.

category path between the concepts. Compute the edge-based semantic relatedness between c_1 and the LCS node, which is the sum of the weights of the edges that link c_1 to the LCS node. Repeat this procedure to find the edge-based semantic relatedness between c_2 and the LCS node.

The overall edge-based relatedness measure between the two concepts is given by

$$d(c_1,c_2) = \frac{\sum_{i=0}^{I} w_i^1 + \sum_{i=0}^{J} w_i^2}{\sum_{i=0}^{R} w_i^1 + \sum_{i=0}^{R} w_i^2} \quad (1)$$

where w_i^1 is the weight of the edge with index i in the category path between c_1 and the LCS category, w_i^2 is the weight of the edge with index i in the category path between c_2 and the LCS category, I is the depth of the last edge of the path that connects c_1 to the LCS and J is the depth of the last edge of the path that connects c_2 to the LCS category, with R denoting the index of the last edge in the path between the root node and a concept node, with the restriction that this path must include the LCS.

The exponential weight of the $i-th$ edge is given by

$$w_i = \beta^{\alpha i} \quad (2)$$

where α and β are predefined parameters.

Edge-based Similarity between Entities. Given two entities E_1 and E_2 represented by discrete sets of concepts $C_1 = \{c_1^1,...,c_n^1\}$ and $C_2 = \{c_1^2,...,c_m^2\}$, respectively, we define the edge-based distance between sets $D(E_1,E_2)$ is

$$D(E_1,E_2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d(c_i,c_j)}{n \times m} \quad (3)$$

Finally, we have the following similarity measure between entities

$$S(E_1,E_2) = 1 - D(E_1,E_2) \quad (4)$$

Weighted Shared Nearest Neighbor Nodes Similarity between Entities (SNN). Let c_1 and c_2 be two concepts such that $c_1 \in E_1$ and $c_2 \in E_2$. Compute all the shortest paths between c_1 and the concepts belonging to E_2 and follow the same procedure for all possible pairs of c_2 with concepts that belong to E_1.

We compute the shortest paths between c_1 and all the categories belonging to E_2, and define C^1 to be the set of categories found in these paths. Conversely, we determine paths from c_2 to the entity E_1 and define C^2 in a similar way.

We refer to C^1 and C^2 as nearest-neighbor sets of c_1 and c_2, respectively. The shared nearest-neighbors of c_1 and c_2 correspond to their intersection, $C^1 \cap C^2$. We define then the following weight, proportional to the amount of shared neighbors

$$SNN(c_1,c_2) = 2\frac{|C^1 \cap C^2|}{|C^1| + |C^2|} \quad (5)$$

which is generalized to measure the weight of shared categories between two entities

$$SNN(E_1,E_2) = 2\sum_{i \in E1}^{n}\sum_{j \in E2}^{m}\frac{|C_i \cap C_j|}{|C_i| + |C_j|} \quad (6)$$

The location of a node in the category network may influence its relevance in the overall computation of the relatedness measure between the entities. A node located deeper in the hierarchy is more specific, and therefore more relevant to the characterization of the semantic proximity of two concepts. If a category path contains only a few categories, located at deeper levels of the hierarchy, then the concepts encompassed by these categories are closer together than concepts encompassed by categories further up in the hierarchy.

By assigning weights to the nodes, Equation 6 becomes

$$W_{SNN}(E_1,E_2) = 2\frac{\sum_{l \in C^i \cap C^j} w_n(l)}{\sum_{l \in C^i} w_n(l) + \sum_{l \in C^j} w_n(lcs)} \quad (7)$$

where w_{nl} is the depth of the l node (equal to the number of edges between the current node and the root node).

Proposed Relatedness Measure. A proposed measure results from the combination of the former weighted similarity measures.

$$M(E_1,E_2) = \frac{\alpha_1 S(E_1,E_2) + \alpha_2 W_{SNN}(E_1,E_2)}{\alpha_1 + \alpha_2} \quad (8)$$

where α_1 and α_2 are predefined parameters.

This measure may be further enhanced by the **Weighted Shared Least Common Subsumer Nodes** (S_{LCS}). This weighting reflects the assumption that, if two entities share a large amount of least common subsumers, and if these nodes are located further down in the category hierarchy, than the entities must be strongly related. On the other hand, if the common subsumers are few, or located in the upper levels of the hierarchy than the semantic relatedness between the entities must be weak.

The following weight is assigned to the shared LCS nodes of two entities,

$$S_{LCS}(E_1,E_2) = \frac{\sum_{lcs\in\{C^1\cap C^2\}} w_n(lcs)}{\sum_{lcs\in C^1} w_n(lcs) + \sum_{l\in C^2} w_n(lcs)} \quad (9)$$

where the weight of the node, $w_{node}(l)$ corresponds to its depth in the hierarchy.

Hence, Equation 8 takes the following form

$$M(E_1,E_2) = \frac{\alpha_1 S(E_1,E_2) + \alpha_2 S_{LCS}(E_1,E_2) + \alpha_3 SNN(E_1,E_2)}{\alpha_1+\alpha_2+\alpha_3} \quad (10)$$

where α_1, α_2 and α_3 are predefined parameters.

3.1 Computation of Shortest Paths in the Wikipedia Graph

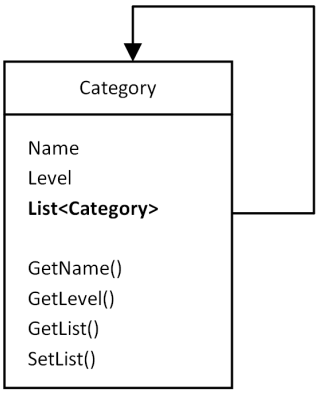

Figure 3: An object of the type CATEGORY.

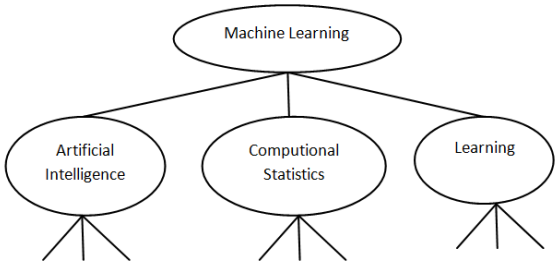

Figure 4: Instatiation of the concept "Machine Learning".

Several versions of the Wikipedia maybe be accessed at http://dumps.wikimedia.org/backup-index.html. For the results presented in this paper, we used a recent english version. To store all the Wikipedia pages and links we used the MySQL structure provided by the Java Wikipedia Library API (available at http://www.ukp.tu-darmstadt.de/software/jwpl/), further described in (Zesch et al., 2008). This API also helps us to determine if a page is of the "Disambiguation pages" type. We did not, however, used the API to build category paths, having specifically devised procedures for this task.

Each Wikipedia object of the type CATEGORY is assigned a level within the current search and a list of its nearest neighbors, when examined for shortest path computation. A representation of these CATEGORY objects is depicted in Figure 3 By regarding this shortest-path search as a tree-search, each instance of the object category will be a leaf of the tree.

```
1  Category nextLevel(Category c)
2      Begin
3      ForEach Category in c.List
4          Begin
5              If (IteratedCategory.List = null)
                   ->leaf node
6                  Begin
7                  WikiList=wikipedia.
                       GetAboveLevelCategories
                       ();
8                  c.List=newList;
9              End
10             Else
11                 Begin
12                 NextLevel(IteratedCategory);
                       ->recursive method
13                 End
14             End
15         End
```

Listing 1: Procedure to examine the upper level of a node.

After the instantiation of concept *Machine Learning* (see Figure 4), all of its parent categories ("Artificial Intelligence", "Learning" and "Computational Statistics") will have a level attribute of two. The instatiation of each of these categories will return their corresponding list of parent categories and a level attribute of 3 and so on. The pseudo code in Listing 1 illustrates this procedure.

The GETABOVELEVELCATEGORIES() method searches the parent categories of the current category, C. For each computation of a category path between two concepts, two trees are built, one for each concept. The level attribute will grow until the algorithm finds a common ancestor (the LCS).

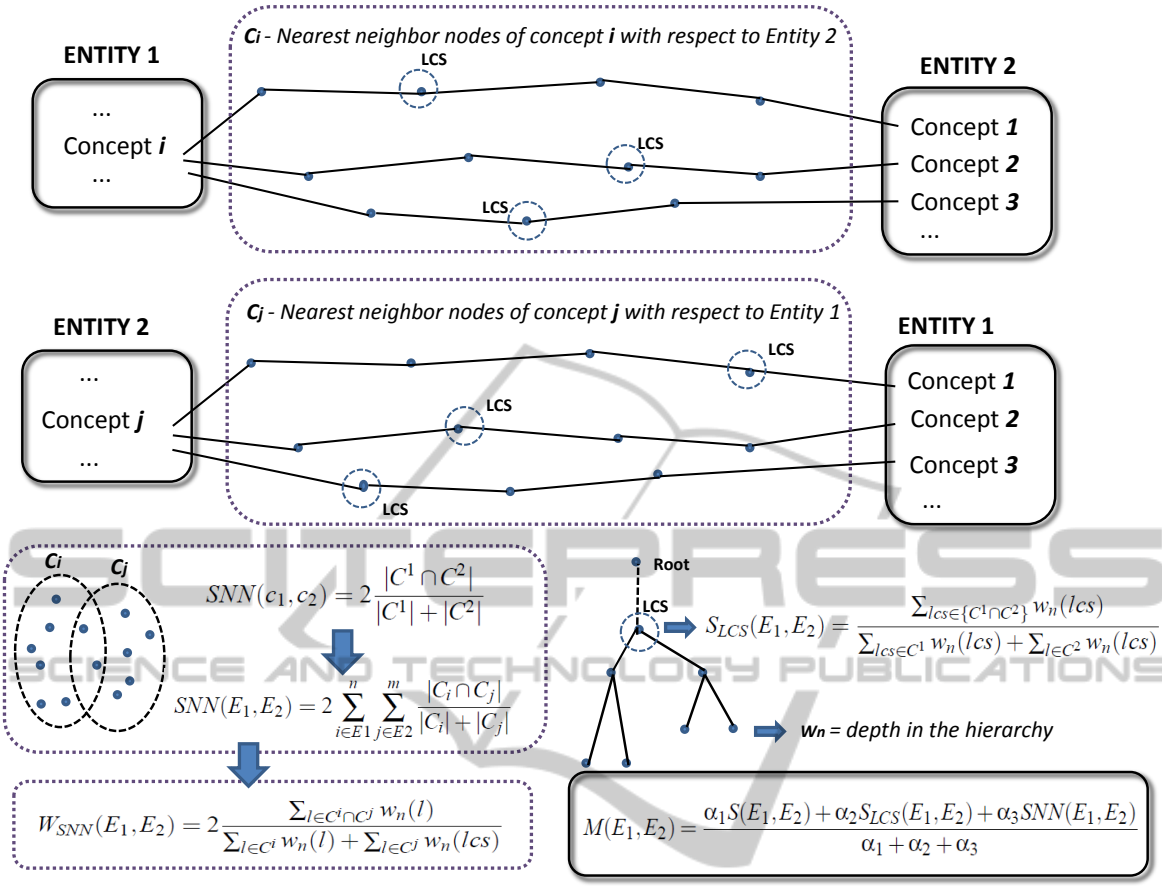

Figure 2: Illustration of the various components of the proposed semantic relatedness measure.

```
1  Void Main()
2      Begin
3       List c1 = wikipedia.
            GetAboveLevelCategories("concept1")
            ;
4         List c2 = wikipedia.
              GetAboveLevelCategories("concept2
              ");
5
6       While(ExistMatch(c1,c2) )
7           Begin
8            nextLevel(c1);
9            nextLevel(c2);
10          End
11     End
```

Listing 2: Procedure to find a path by means of a least common subsumer search.

The pseudo code in Listing 2 illustrates this procedure for example concepts ”concept1” and ”concept2”.

The method EXISTMATCH examines all the nodes of the two trees: if a match is found, then this node (which is the Least Common Subsummer) is returned. The shortest category path between the concepts is then found by examining the categories between the concepts and the LCS. To reduce the computational costs, each search is stored in a cache table.

These procedures were implemented with Java. Java is not the most adequate technology for this type of tree search, since it lacks tail call elimination for security reasons, as further detailed in the Bugs section of Oracles website[3] but it was sufficiently effective to accomplish our goals.

4 RESULTS AND DISCUSSION

For testing the proposed measure, we have chosen 6 entities: three of them represent well known conferences (CVPR[4], KDD[5] and RECOMB[6]). These conferences were chosen because each corresponds to a

[3] See http://bugs.sun.com/view_bug.do?bug_id=4726340
[4] http://www.cvpr2012.org/
[5] http://kdd2012.sigkdd.org/
[6] http://bioinfo.au.tsinghua.edu.cn/recomb2013/

Table 1: Entities represented as sets of scientific topics. The first three entities represent actual conferences (CVPR, KDD and RECOMB respectively). The other three entities represent authors.

E1	E2	E3	E4	E5	E6
Computer Vision	Knowledge Discovery	Molecular Biology	Computer Vision	Information Extraction	Genetics
Object Recognition	Data Mining	Gene Expression	Robotics	Machine Learning	Human Genome
Structure from Motion	Web Mining	Computational Biology	Object Recognition	Natural Language Processing	Gene Expression
Image Segmentation	Recommender Systems	Genomics	Structure from Motion	Information Retrieval	Systems Biology
Image Processing	Cluster Analysis	Population Genetics	Human Computer Interaction	Data Mining	Clinical Medicine
Object categorization	Text Mining		Virtual Reality	Graphical Model	Bioinformatics
Optical Flow	Data Analytics		Facial Expression	Social Network	
Pattern Recognition	Structure Mining			Reinforcement Learning	
				Web Mining	

Table 2: Proposed similarity measure results. The chosen parameter values are $\alpha = 1$ and $\beta = 3$.

Pair	(E_1,E_4)	(E_1,E_5)	(E_1,E_6)	(E_2,E_4)	(E_2,E_5)	(E_2,E_6)	(E_3,E_4)	(E_3,E_5)	(E_3,E_6)
S Eq.4	0.948	0.931	0.912	0.869	0.954	0.826	0.786	0.862	0.989
W_{SNN} Eq. 7	0.721	0.655	0.522	0.672	0.599	0.542	0.538	0.432	0.681
S_{LCS} Eq. 9	0.356	0.251	0.209	0.265	0.331	0.261	0.194	0.206	0.331
M Eq. 10	0.675	0.612	0.547	0.602	0.628	0.543	0.506	0.500	0.667

distinct scientific research area: CVPR to Computer Vision and Pattern Recognition; KDD to Data Mining and Knowledge Discovery; RECOMB to Computational Molecular Biology. The other three entities represent well known authors. Each of these authors was chosen based on the strong correspondence of their research interests with one of the three conferences:

- **Author E_4:** research interests in the area of computer vision and object recognition (from images), which matches CVPR represented by E_1.
- **Author E_5:** research interests in the area of data mining and machine learning, which are more related to the scientific areas of KDD, represented by E_2, but can also be related to CVPR by means of the topic "Pattern Recognition".
- **Author E_6:** research interests in the area of genetics and bioinformatics, which is more closely related to RECOMB (represented by E_3) than the other two conferences, although bioinformatics also overlaps the areas of KDD.

The topic set that represents each entity was derived from the conference or authors official website.

Some scientific concepts do not exist in the Wikipedia version or lead to disambiguation pages. A solution for the first case was to replace the concept with a similar one that does exist in the Wikipedia. For instance, the concept "Image Segmentation" of E_1 had to be replaced with the page "Segmentation (image processing)". A possible solution for the second case would be to access the disambiguation page, retrieve the Wikipedia links there listed and choose from these the most appropriate one for our search by examining their nearest neighbor categories. This alternative will be explored in future work.

The chosen parameter values are $\alpha = 1$, $\beta = 3$, $\alpha_1 = 1$, $\alpha_2 = 1$ and $\alpha_3 = 1$. The depth of each node in the Wikipedia hierarchy is determined with respect to the category entitled "Main topic classifications"[7] which encompasses Wikipedia's major categories.

The concept sets that represent the entities are depicted in Table 1. The values obtained with the proposed measure are in Table 2, were the different components are depicted in separate rows. The overall relatedness measure values correspond to the last row.

From these values we observe a high value of similarity relatedness for the following entity pairs: (E_4,E_1), (E_5,E_2), and (E_6,E_3). This is expected due to the semantic overlapping of these entities. It was also expected that the similarity values for (E_1,E_5) would be much lower than the value found for (E_1,E_4). The underlying cause may be the "Pattern Recognition" concept of E_1 which is strongly correlated with the concepts of E_5. Other possible cause may be the S component of the proposed measure (see Equation 4): it relies strongly on the computation of the distance from the nodes in the category paths to the root node which was chosen to be "Main topic classifications". We observe that in many cases this distance is very high, which originates high similarity values that are not quite differentiated from entity to entity. This also has some impact in the other components of the measure. A solution for this

[7] http://en.wikipedia.org/wiki/Category:Main_topic_classifications

would be to choose as root node a category lower in the Wikipedia hierarchy than "Main Topic Classifications", possibly a node that still encompasses the overall topics of the entity, but not as generic as the one chosen here. The values of the parameters are being fine-tuned in ongoing work in order to further improve the proposed measure.

5 CONCLUSIONS

In this paper we presented a new semantic relatedness measure between entities, using Wikipedia as a hierarchy of scientific categories. The devised measure examines the Wikipedia category paths between all the possible concept pairs of two distinct entities, assigning weights according to the category's relevance in the resulting path set and in the Wikipedia graph. We examined the proposed measure values for selected entities, observing that these match the intuitive human assessment of their semantic similarity. We conclude then that this is a valid approach to automatically assess the proximity of scientific researchers and other scientific entities such as conferences and journals.

Future Work. Ongoing work includes comparison of the results obtained with manual annotations done by volunteers, using a website specifically deployed for this task (www.insticc.org/SemanticDistance.aspx). Further work includes continuing exploration of the measure for other entity pairs, comparison of our measure with other state-of-the-art metrics, devising tasks of semantic disambiguation of Wikipedia articles and clustering of concept sets such that an entity may be represented by several subsets of scientific topics, each subset representing a particular area.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of the Instituto de Telecomunicações (IT-IST) and Escola Superior de Tecnologia, Instituto Politécnico de Setúbal (EST-IPS).

REFERENCES

Gabrilovich, E. and Markovitch, S. (2007). 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. Morgan Kaufmann Publishers Inc.

Gouws, S., Rooyen, G., and Engelbrecht, H. (2010). Measuring conceptual similarity by spreading activation over wikipedia's hyperlink structure. In *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*.

Jiang, J. J. and Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*.

Leacock, C. and Chodorow, M. (1998). *Combining Local Context and WordNet Similarity for Word Sense Identification*, chapter 11, pages 265–283. The MIT Press.

Liu, J. and Birnbaum, L. (2007). Measuring semantic similarity between named entities by searching the web directory. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 461–465.

Milne, D. and Witten, I. H. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *In Proceedings of AAAI 2008*.

Nastase, V. and Strube, M. (2008). Decoding wikipedia categories for knowledge acquisition. In *AAAI*, pages 1219–1224.

Ponzetto, S. P. and Strube, M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Int. Res.*, 30:181–212.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130.

Rodrguez, M. A. and Egenhofer, M. J. (2003). Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15:442–456.

Slimani, T., Yaghlane, B. B., and Mellouli, K. (2006). A New Similarity Measure based on Edge Counting. In *Proceedings of world academy of science, engineering and technology*, volume 17.

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138. Association for Computational Linguistics.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.