# Wikipedia Category Graph and New Intrinsic Information Content Metric for Word Semantic Relatedness Measuring

Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha,
Mohamed Tmar, and Abdelmajid Ben Hamadou

MIRACL Laboratory, Sfax University, Tunisia
mohamedali.hadjtaieb@gmail.com, mohamed.benaouicha@irit.fr,
{mohamed.tmar,abdelmajid.benhamadou}@isimsf.rnu.tn

**Abstract.** Computing semantic relatedness is a key component of information retrieval tasks and natural processing language applications. Wikipedia provides a knowledge base for computing word relatedness with more coverage than WordNet. In this paper we use a new intrinsic information content (IC) metric with Wikipedia category graph (WCG) to measure the semantic relatedness between words. Indeed, we have developed a performed algorithm to extract the categories assigned to a given word from the WCG. Moreover, this extraction strategy is coupled with a new intrinsic information content metric based on the subgraph composed of hypernyms of a given concept. Also, we have developed a process to quantify the information content subgraph. When tested on common benchmark of similarity ratings the proposed approach shows a good correlation value compared to other computational models.

**Keywords:** Wikipedia, Information content, Semantic relatedness, Wikipedia category graph.

## 1 Introduction

Semantic relatedness (SR) measures how much two words or concepts are related using all types of relations between them [1]. Budantsky & Hirst, in their paper [2] show that semantic relatedness is more general than semantic similarity. In fact, two concepts or two words can be related but are not necessary similar like *cars* and *gasoline* cited in [3]. SR is used as a necessary pre-processing step to many Natural Language Processing (NLP) tasks, such as Word Sense Disambiguation (WSD) [4,5]. Moreover, SR constitutes one of the major stakes in the Information Retrieval (IR) [6-8] especially in some tasks like semantic indexing [9]. A powerful semantic relatedness measure influences on Semantic Information Retrieval (SIR) system. It exists in some information retrieval systems that support retrieval by *Semantic Similarity Retrieval Model* (*SSRM*) [38].

These methods are based on knowledge sources to assess the semantic relatedness degree between words and concepts. The knowledge sources can be unstructured

Y. Xiang et al. (Eds.): IDCS 2012 and ICDKE 2012, LNCS 7646, pp. 415–427, 2012.

documents or (semi-)structured resources such as Wikipedia, WordNet, and domain specific ontologies (e.g., the Gene Ontology). Research in [1,12,13] has shown that the accuracy of an SR method depends on the choice of the knowledge sources, without concluding an importance order between these knowledge sources.

Methods for computing SR use the semantic network features. Therefore, these methods can be classified according to path based, Information Content (IC) based, and statistical methods. Path based methods (Rada [21]; Wu and Palmer [22]; Hirst and St-Onge [18]; Leacock and Chodorow [4]) measure SR between concepts as a function of their distance in a semantic network, usually calculated using the path connecting the concepts following certain semantic links (typically is-a or known as hypernym/hyponym). IC based methods use the amount of information shared between the concerned concepts. It is usually determined by a higher level concept that subsumes both concepts in a taxonomic structure. (Resnik [26], Jiang and Conrath [23], Lin [24], Lord [14], Seco [27], Sebti [28], Pirro [25]). Statistical methods measure relatedness between words or concepts based on their distribution of contextual evidence. This can be formalized as cooccurrence statistics collected from unstructured documents or distributional concept or word vectors with features extracted from either unstructured documents [33-34] or (semi-)structured knowledge resources [10,13, 35-37].

Computing SR requires a knowledge resource about concepts or words. (Semi-)Structured knowledge sources organize explicitly the semantic knowledge about concepts and words and interlink them with semantic relations. There are some studies of SR, that include lexical resources such as WordNet Fellbaum [17], Roget's thesaurus [39], Wiktionary, and (semi-)structured encyclopedic resources such as Wikipedia. Earlier studies used WordNet such as (Hirst and St-Onge [18]; Jiang and Conrath [23]; Lin [24]; Leacock and Chodorow [4]; Resnik [26]; Seco et al. [27]; Wu and Palmer [22]). Wordnet is still a preferred knowledge source in recent works [35]. However, its effectiveness has a problem concerning its lack of coverage of specialized lexicons and domain specific concepts [1,13]). Wikipedia and Wiktionary are collaboratively maintained knowledge sources by volunteers and therefore may overcome this limitation. Wikipedia in particular, is found to have a reasonable coverage of many domains [20,36,38]). Recently, It has become increasingly popular in SR studies. Wikipedia grows exponentially and has probably become the largest collection of freely available knowledge.

Recent Wikipedia-based lexical semantic relatedness approaches have been found to outperform measures based on the WordNet graph. Recent work has shown that Wikipedia can be used as the basis of successful measures of semantic relatedness between words or text passages. Such methods stand out WikiRelate! [1], Explicit Semantic Analysis (ESA) [32], Wikipedia Link Vector Model (WLVM) [31] and WikiWalk [30].

This work exploits our novel IC metric based on hypernyms subgraph with Wikipedia categories graph as semantic taxonomic resource for measuring semantic relatedness between words. The rest of the paper is organized as follows. Section 2 focuses on the Wikipedia category system structure as semantic taxonomy. Section 3 describes the extraction categories strategy to determine the semantic relatedness between words using our novel IC metric with the Wikipedia category graph as semantic resource. Section 4 includes the detailed elaboration of our proposed metric. Section 5 includes the experiments and the evaluation to show the effectiveness of our proposed method. Concluding remarks and some future directions of our work are described in Section 6.

## 2 Wikipedia's Category System

In Wikipedia, at the bottom of each page all assigned categories are listed with links to the category page where an automatic index of all pages tagged with this category is shown. The number of categories has evolved from 165744 categories on 25 September 2006 to 337705 categories on March 12, 2008. For the current study, the English Wikipedia dump dated 12 March 2008 is downloaded[1] to form the category graph used as semantic resource [10].

Most of the categories are connected to a selected main category "*Contents*" that is superordinated to all other categories. Wikipedia categories and their relations do not have explicit semantics like known ontologies in computer science. The Wikipedia categorization system does not form a taxonomy like the WordNet "is a" taxonomy with a fully subsumption hierarchy, but only a thematically organized thesaurus. As an example *Computer systems* is categorized in the upper category *Technology systems* (is a) and *Computer hardware* (has part).

So the categories form a directed acyclic graph which can be taken to represent a conceptual network with unspecified semantic relations [10]. To use WCG as semantic resource for semantic relatedness computing we realize a cleanup for this network. Indeed, we start with the full categorization network consisting of 337705 nodes. We first clean the network from meta-categories used for management, e.g the categories under *Wikipedia administration*. We remove instead all those nodes whose labels contain any of the following strings: *Wikipedia*, *wikiproject*, *lists*, *mediawiki*, *template*, *user*, *portal*, *categories*, *articles*, *pages* and *stub*. Secondly, we browse the result graph to remove orphan nodes and we keep only the category *Contents* as root. The maximum depth of the graph passed from 291 before the cleanup to 221 for the result graph.

Moreover, in WCG a path linking two categories does not represent necessarily a semantic meaning. For example there exists a path connecting categories *video games* and *water pollution*. This property is a hard problem that decreases the results in the semantic relatedness task through the WCG. In the WCG each category is taken as a concept.

Figure 1 describes the steps followed in our system to exploit the WCG as semantic network for computing semantic relatedness between words. Indeed, firstly we must derive the WCG from a Wikipedia dump. After, using the two concerned words and the categories extraction strategy we can locate the sets of categories $C_1$ and $C_2$ assigned respectively to $w_1$ and $w_2$. The next step concerns the information content

[1] http://download.wikimedia.org/enwiki/20080312/enwiki-20080312-pages-articles.xml.bz2

computing of each extracted category using its subgraph formed by its hypernyms as indicated for the category *Insects* in figure 2. In the next paragraph, we will detail the extraction strategy.

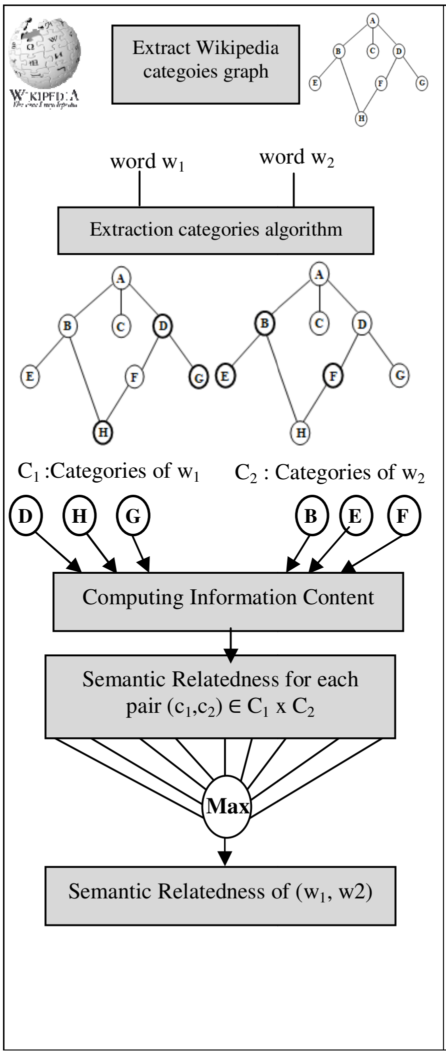
Extract Wikipedia categoies graph
word $w_1$
word $w_2$
Extraction categories algorithm
$C_1$ :Categories of $w_1$
$C_2$ : Categories of $w_2$
A
B
C
D
E
F
G
H
Computing Information Content
Semantic Relatedness for each pair $(c_1,c_2) \in C_1 \times C_2$
Max
Semantic Relatedness of ($w_1$, w2)

**Fig. 1.** Information flow for measuring semantic relatedness using the Wikipedia category graph

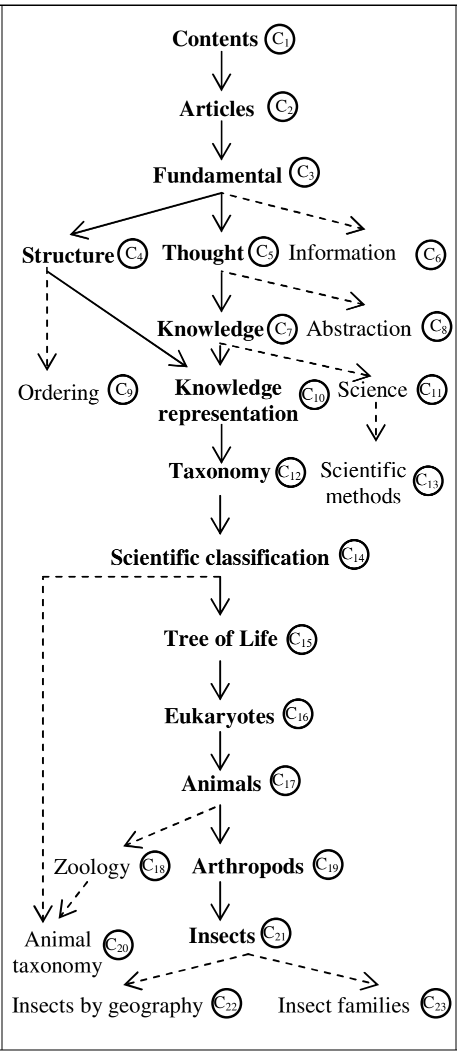

**Fig. 2.** An excerpt from Wikipedia category graph containing 23 categories. The solid arrows represent the information content subgraph of the category "*Insects*".

## 3 Categories Extraction Strategy

In WordNet, polysemous words may have more than one corresponding node in “is a” taxonomy, the resulting semantic relatedness of two words $w_1$ and $w_2$ can be calculated as:

$$SemRel(w_1, w_2) = \max_{(c_1,c_2) \in Syn(w_1) \times Syn(w_2)} SemRel(c_1, c_2) \quad (1)$$

With *Syn(w*$_i$*)* is the set of $w_i$ synsets. Thus, the relatedness of two words is equal to the most related synsets pair.

In our study, we use the WCG to compute the semantic relatedness between two words $w_1$ and $w_2$. So, similar to synsets for Wordnet, we must extract the set of categories $C_i$ that represents the word $w_i$. We define $C_1$ and $C_2$ as categories sets assigned respectively to words $w_1$ and $w_2$. Then, we compute the SR for each category pair $(c_k, c_j) \in C_1 x C_2$ using our IC computing method. Therefore, we define a categories extraction strategy. In equation 1 the set of synsets are replaced by $C_1$ and $C_2$:

$$SemRel(w_1, w_2) = \max_{(c_1,c_2) \in C_1 \times C_2} SemRel(c_1, c_2) \quad (2)$$

The categories extraction process starts by fetching the Wikipedia article allocated to the word $w_i$. If it is a redirected page, we collect all redirected pages of the concerned word. Then, we traverse all returned pages to extract their categories. On the other hand, if the page is not a redirected page, we extract its categories. The polysemous problem in Wikipedia is marked by the category “*Disambiguation*” that indicates the disambiguation pages which play a sense inventory role for a word $w_i$. when we find this category during the extraction process, we pass to a special treatment.

Resolving ambiguous page queries is an important task in our categories extraction strategy due to queries in Wikipedia that return a disambiguation page. This treatment contains principally the three following steps:

- **Step 1:** We start extracting all *outLinks* of the page containing the category “disambiguation”. These links of the disambiguation page are ordered in the decreasing order according to the outLinks number of their target articles. In fact, we suppose that an article which contains more outLinks is the most semantically rich. We can compare it to a concept which inherits its features from several other concepts. Finally, we use the first two pages links that match with the patterns *$w_1$ ($w_2$)* or *$w_2$ ($w_1$)*. If no pages are returned, we take the page links which contain $w_2$ and $w_1$ but not between parenthesis.
- **Step 2:** If the first step does not return an appropriate page, we move to this step. Indeed, using the same ordered set of outLinks extracted from the disambiguation page; we try to find links containing disambiguation tag in parenthesis. It is enough to treat the two first turned pages to not overload the set of categories which influence negatively on the semantic relatedness computing.

- **Step 3:** If no pages are found in step2, we extract the categories from the corresponding articles of the two first links existing in the ordered set that contains the concerned word $w_i$.

Finally, if none of the categories are found, we take the categories of the article assigned to the first link existing in the ordered set. Once the whole of the categories is ready we calculate the IC of each pair $(c_1,c_2) \in C_1 \times C_2$. The IC calculation uses a novel metric that is detailed in the next paragraph.

# 4 The Information Content Metric

The IC concept was introduced for the first time by Resnik [26] following the standard argumentation of information theory. In Some previous works, was obtained the IC through statistical analysis of corpora, from where probabilities of concepts occurring are inferred. Other authors feel that the taxonomic semantic network can also be used as a statistical resource with no need for external ones. Their methods of obtaining IC values rest on the assumption that the taxonomic structure is organized in a meaningful and principled way. Following the standard argumentation of information theory [43], the IC of a concept c can be quantified as negative log likelihood: -log p(c). Notice that quantifying information content in this way makes intuitive sense in this setting: as probability increases, informativeness decreases, so the more abstract a concept is, the lower its information content is.

## 4.1 The Previous IC Methods

In this paragraph we present some previous IC methods like:

- The conventional way of measuring the IC of word senses is to combine the knowledge of their hierarchical structure from the semantic network with the statistics of their actual usage in a large corpus.

  The IC value is then calculated by negative log likelihood formula as follow:

$$IC(c) = -\log\big(p(c)\big) \tag{3}$$

  Where $c$ is a concept and $p$ is the probability of encountering $c$ in a given corpus. If sense-tagged text is available, frequency counts of concepts can be attained directly, since each concept will be associated with a unique sense. If sense tagged text is not available, it will be necessary to adopt an alternative counting scheme. Resnik [21] suggests counting the number of occurrences of a word type in a corpus, and then dividing that count by the number of different concepts/senses associated with that word. This value is then assigned to each concept. Resnik showed that semantic similarity depends on the amount of information that two concepts have in common, this shared information is given by the *Most Specific Common Abstraction* (MSCA) that subsumes both concepts.

Other works use only the taxonomy structure without making recourse to the external corpora.

- (Nuno Seco et al., 2004), in [27] they present a completely IC intrinsic measurement which is connected only to the hierarchical structure of WordNet. This minimizes the complexity treatment of a corpus in order to extract the probabilities from the concepts. The computation equation is the following:

$$IC(c) = 1 - \frac{Log(hypo(c) + 1)}{Log\,(\text{max})} \qquad (4)$$

  Where *hypo(c)* is a function which returns the hyponyms number of a given concept and *max* is a constant which represents the maximum number of concepts.
- (Sebti et al., 2008) in [28] they present a new approach for measuring semantic similarity between words via concepts. Their proposed measure is a hybrid system based on using a new IC metric with the WordNet hierarchical structure. This method is based on the number of direct hyponyms of each concept pertaining to the initial way of the root until the target concept.

### 4.2 Our IC Metric

In this study, we present a novel IC metric that is completely derived from our works on WordNet [29, 40]. Indeed, in the "is a" hierarchical structure, a subordinate concept inherits the basic features from the superordinate concept and adds its own specific features to form its meaning. Thus, a concept is an accumulation of the propagated information from a distant ancestor to another less distant by adding specificities to each descendant. Therefore, a concept depends strongly on its direct parents and ancestors. Direct and indirect hypernym relations of a concept *c* from a subgraph which will take part in its IC quantification. The contribution of a concept pertaining to the hypernyms subgraph of a target concept depends on its depth. Figure 2 presents an example of hypernyms subgraph for the concept "*Insects*". Indeed, this subgraph is formed by the category set {$C_1$,$C_2$,$C_3$,$C_4$,$C_5$,$C_7$,$C_{10}$,$C_{12}$,$C_{14}$,$C_{15}$,$C_{16}$,$C_{17}$,$C_{19}$,$C_{21}$} (subgraph formed by solid arrows in figure 2).

In order to quantify the hypernyms subgraph of a given concept (for example IC(*Insects*)= 314.2565 ) we will present these notations:

**Hyper(c):** the set of direct hypernyms of the concept *c*. For example, *Hyper($C_{10}$)={$C_4$,$C_7$}* .

**Hypo(c):** the number of concepts subsumed by the concept *c*. For example, *Hypo($C_7$)=277508 and Hypo($C_{21}$)=552* (in WCG).

**SubGraph(c):** set of all concepts pertaining to hypernyms subgraph modeling the IC of the concept *c*. For example subgraph(*$C_{21}$*)= {$C_1$,$C_2$,$C_3$,$C_4$,$C_5$,$C_7$,$C_{10}$, $C_{12}$,$C_{14}$,$C_{15}$,$C_{16}$,$C_{17}$,$C_{19}$,$C_{21}$}.

**Depth(c):** the maximal depth of the concept *c* in the WCG taxonomy. For example, Depth($C_{21}$)=*12.*

**AverageDepth (c):** the average depth of the hypernyms subgraph of the concept *c*. It is computed as follows:

$$AverageDepth(c) = \frac{1}{|\mathrm{SubGraph(c)}|} \times \sum_{\mathrm{c' \in SubGraph(c)}} \mathrm{Depth(c')} \qquad (5)$$

Where *c'* is a concept.

To compute the *IC (Categ)* where *Categ* is a category. ∀ *c* ∈ *SubGraph(Categ)*, we calculate a score:

$$Score(c) = \left( \sum_{c' \in Hyper(c)} \frac{Depth(c')}{Hypo(c')} \right) \times Hypo(c) \qquad (6)$$

Indeed, for each category *c'* ∈ $Hyper(c)$ we calculate a term as follows: $\frac{Hypo(c)}{Hypo(c')} \times Depth(c')$ (term). *Score(c)* represents the contribution of each concept pertaining to *subgraph(Categ)* on *IC(Categ)*. Finally, we calculate the total IC of the concept *Categ* as follows:

$$IC(Categ) = \left( \sum_{c \in SubGraph(Con)} Score(c) \right) \times \mathrm{AverageDepth}(Categ) \qquad (7)$$

As presented in the quantification process, the hyponyme number and the maximal depth of each concept are the important taxonomy properties used in our metric. Indeed, they indicate semantic richness for each concept pertaining to the semantic network. The average depth of the *subgraph(c)* gives information about the vertical distribution of *subgraph(c)*. When this value is large, it indicates an important features enrichment. This method is adapted to an “is a” taxonomy formed basically by hyponym/hypernym relations. But, the WCG cannot be taken as an “is a” taxonomy because it is not based only on hypernym/hyponym relations. In the experiment part we discuss the solution that we found to use this IC metric with the WCG as semantic knowledge.

## 5 Experiments and Evaluation

The Wikipedia based measure used in this article is implemented based on JWPL[2], a high-performance Java-based Wikipedia API. JWPL operates on an optimized database that is created from the database dumps available from the Wikimedia foundation. This allows for fast access to Wikipedia articles, categories, links, redirects, etc. A more detailed description of JWPL can be found in [19].

[2] http://www.ukp.tu-darmstadt.de/software

Following the literature on semantic similarity, we evaluate the performance by taking the Pearson product moment correlation coefficient *r* between the similarity scores and the corresponding human judgments. We perform an evaluation by computing semantic relatedness on three commonly used datasets, namely Miller & Charles' list of 30 noun pairs (M&C) [41] and the 65 word synonymy list for Rubenstein & Goodenough (R&G) [42] and Finkelstein (F) [6] dataset composed of 353 noun pairs. We use the Lin formula [24] coupled with our IC metric as follow:

$$Sim(c_1, c_2) = \frac{2\times IC(\mathrm{MSCA})}{IC(c_1)+\mathrm{IC}(\mathrm{c}_2)} \tag{8}$$

Then, we use the maximum function to extract the semantic similarity degree between these words:

$$Sim(w_1, w_2) = \max_{\mathrm{c}_1\neq \mathrm{c}_2,(\mathrm{c}_1,\mathrm{c}_2)\in \mathrm{C1}\times\mathrm{C2}} \mathrm{Sim}(\mathrm{c}_1, \mathrm{c}_2) \tag{9}$$

**Table 1.** Results on correlation with human judgments of relatedness measures

| M&C | R&G | F |
|---|---|---|
| 0.40 | 0.34 | 0.38 |

Table 1 reports the scores obtained by computing semantic relatedness using formula (9). Considering that the experiment on benchmark datasets revealed a performance far lower than expected, we sought the problem in our semantic network (WCG). Indeed, many unrelated pairs received a score higher than expected, because of coarse grained over connected categories in the WCG. In fact, the WCG does not contain only is-a relations. So, a path which links two categories does not mean a strong semantically relation, for example we can find a path connecting the two categories *video games* and *food*. Therefore we chose to limit our search to a chosen depth when browsing the WCG. After an empirical study, we found that limiting the search improves the results. Indeed, the best maximum depth is 4. So, the subgraph formed by hypernyms is limited to the depth 4 and the function *hypo* detailed in our IC metric is limited to the same value. The analysis of results found with the value 4 shows that the problem of higher scores for unrelated pairs is resolved but another problem is appeared concerning the median values. Therefore, we chose to correct these values by computing semantic relatedness using the depth 5. In fact, when a computing value with depth 4 is inferior to a threshold (0.1 in our case), it will be replaced by the novel value found with the depth 5 if it exceeds 0.1. An excellent amelioration was noticed as indicated in table 2:

**Table 2.** Results on correlation with human judgments of relatedness measures after fixing depth limited search

| M&C | R&G | F |
|---|---|---|
| 0.73 | 0.65 | 0.61 |

In next experiments, we choose to compare the IC methods based on the taxonomy structure with our measure. Therefore, we use the IC-based formulas (Res: Resnik 1995 [26], Lin 1998 [24], and JC: Jiang and Conrath 1997 [23]).

**Table 3.** The coefficients correlation between human similarity judgments of the Finkelstein dataset and the suggested similarity measures

| **WCG** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Seco, 2004** | | | **Sebti, 2008** | | | **Our metric** | | |
| **Res** | **Lin** | **JC** | **Res** | **Lin** | **JC** | **Res** | **Lin** | **JC** |
| 0.45 | 0.51 | 0.21 | 0.39 | 0.42 | 0.46 | 0.41 | **0.61** | 0.35 |
| **WordNet** | | | | | | | | |
| **Seco, 2004** | | | **Sebti, 2008** | | | **Our metric** | | |
| **Res** | **Lin** | **JC** | **Res** | **Lin** | **JC** | **Res** | **Lin** | **JC** |
| 0.32 | 0.38 | 0.35 | 0.30 | 0.34 | 0.33 | 0.45 | 0.58 | 0.41 |

Table 3 shows that our IC metric used with Lin formula and WCG as semantic resource outperforms all other measures using IC methods Seco [27] and Sebti [28]. Indeed, the result 0.61 exceeds the best result found with WCG for other methods (0.51, Seco and Lin formula). It allows us to have a profit of 0.10.

In order to show the effectiveness of our categories extraction strategy, we compared it to the strategy followed by Ponzetto [11]. In fact, the results presented in their paper [11] are 0.47 (M&C), 0.52 (R&G) and 0.49 (F) which proves the good performance of our extraction strategy. Moreover, the comparison with the WordNet shows an important advance for the WCG semantic network.

To improve the correlation value 0.61 found with Finkelstein dataset we chose to add a third set $C_{1,2}$ to sets $C_1$ and $C_2$ that contain the extracted categories from respectively word $w_1$ and $w_2$. The set $C_{1,2}$ contains the categories of the page (If it exists) which its title matches with the patterns "$w_1$ $w_2$" , "$w_2$ $w_1$", "*redirects*($w_1$) $w_2$", "$w_2$

**Table 4.** Performance of semantic relatedness measures

| Measure | Correlation with manual judgments | Reference |
|---|---|---|
| WordNet | 0.33-0.35 | [2] |
| Roget's thesaurus | 0.55 | [39] |
| LSA[3] | 0.56 | [44] |
| WikiRelate ! | 0.19-0.48 | [1] |
| ESA | 0.72 | [32] |
| WLVM | 0.69 | [31] |
| Our method | **0.69** | |

[3] LSA accessible plateform via `http://lsa.colorado.edu/`

*redirects($w_1$)*", "*redirects($w_2$) $w_1$*" and "*$w_1$ redirects($w_2$)*". In Wikipedia a redirect link may provide information about synonyms, spelling variations, related terms and abbreviations. Afterwards, we apply the formula 9 on the sets ($C_1$,$C_{1,2}$) and ($C_2$,$C_{1,2}$) and take the maximum value as the semantic relatedness between $w_1$ and $w_2$. This idea leads to a remarkable improvement showed in the table 4.

The correlation 0.69 shown in table 4 can be compared directly to the results of other measures exploiting the same knowledge base (Wikipedia) and other semantic resources (WordNet and Roget's thesaurus). It is only slightly lower than the most accurate method ESA.

## 6 Conclusion and Future Work

This paper has demonstrated that performing our novel IC metric and a strategy for categories extraction over Wikipedia categories graph is a feasible and potentially fruitful means of computing semantic relatedness for words. Our Wikipedia-based relatedness measure proved to be competitive with other studies centered on Wikipedia features.

In future work, we will concentrate to make a thorough analysis of the WCG as an ontological source and compare it to the articles graph based on the hyperlink structure of Wikipedia. Moreover, we think about using the Wikipedia page content coupled with the category graph to study their contribution on semantic relatedness task.

## References

1. Strube, M., Ponzetto, S.: WikiRelate! Computing semantic relatedness using Wikipedia. In: Proceedings of the 21st National Conference on Artificial Intelligence, AAAI (2006)
2. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of semantic distance. Computational Linguistics 32(1), 13–47 (2006)
3. Resnik, P.: 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 (1999)
4. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (ed.) WordNet. An Electronic Lexical Database, ch. 11, pp. 265–283 (1998)
5. Han, X., Zhao, J.: Structural semantic relatedness: a knowledge-based method to named entity disambiguation. In: The 48th Annual Meeting of the Association for Computational Linguistics (2010)
6. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: the concept revisited. ACM Transactions on Information Systems 20(1), 116–131 (2002)
7. Gurevych, I., Müller, C., Zesch, T.: What to be? – electronic career guidance based on semantic relatedness. In: Proceedings of ACL, pp. 1032–1039. Association for Computational Linguistics, Prague (2007)

8. Baziz, M., Boughanem, M., Aussenac-Gilles, N.: Evaluating a Conceptual Indexing Method by Utilizing WordNet. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 238–246. Springer, Heidelberg (2006)
9. Zargayouna, H.: Contexte et sémantique pour une indexation de documents semi-structurés. In: ACM COnférence en Recherche Information et Applications, CORIA 2004 (2004)
10. Zesch, T., Gurevych, I.: Analysis of the Wikipedia Category Graph for NLP Applications. In: Proceedings of the TextGraphs-2 Workshop, NAACL-HLT (2007)
11. Ponzetto, S.P., Strube, M.: Knowledge Derived From Wikipedia For Computing Semantic Relatedness. Journal of Artificial Intelligence Research 30, 181–212 (2007)
12. Zesch, T., Gurevych, I.: Wisdom of crowds versus wisdom of linguists: measuring the semantic relatedness of words. Journal of Natural Language Engineering 16, 25–59 (2010)
13. Zhang, Z., Gentile, A., Xia, L., Iria, J., Chapman, S.: A random graph walk based approach to compute semantic relatedness using knowledge from Wikipedia. In: Proceedings of LREC 2010 (2010)
14. Lord, P., Stevens, R., Brass, A., Goble, C.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19(10), 1275–1283 (2003)
15. Patwardhan, S., Pedersen, T.: Using WordNet based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL 2006 Workshop Making Sense of Sense, Italy (2006)
16. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic Wikipedia. In: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland (2006)
17. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
18. Hirst, G., St-Onge, D.: Lexical chains as representation of context for the detection and correction malapropisms. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database and Some of Its Applications, Cambridge, pp. 305–332 (1998)
19. Zesch, T., Müller, C., Gurevych, I.: Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: Proceedings of the Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco (2008)
20. Zesch, T., Gurevych, I., Mühlhäuser, M.: Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In: Biannual Conference of the Society for Computational Linguistics and Language Technology, pp. 213–221 (2007)
21. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19(1), 17–30 (1989)
22. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 133–138 (1994)
23. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the International Conference on Research in Computational Linguistics, pp. 19–33 (1997)
24. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 296–304 (1998)
25. Pirro, G.: A semantic similarity metric combining features and intrinsic information content. Data and Knowledge Engineering 68(11), 1289–1308 (2009)
26. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of IJCAI 1995, pp. 448–453 (1995)

27. Seco, N., Hayes, T.: An intrinsic information content metric for semantic similarity in WordNet. In: Proceedings of the 16th European Conference on Artificial Intelligence (2004)
28. Sebti, A., Barfrouch, A.A.: A new word sense similarity measure in WordNet. In: Proceedings of the International Multiconference on Computer Science and Information Technologie, Poland (2008)
29. Hadj Taieb, M., Ben Aouicha, M., Tmar, M., Ben Hamadou, A.: New Information Content Metric and Nominalization Relation_for a new WordNet-based method to measure the semantic relatedness. In: 10th IEEE International Conference on Cybernetic Intelligent Systems, University of East London (2011)
30. Yeh, E., Ramage, D., Manning, C., Agirre, E., Soroa, A.: WikiWalk: Random walks on Wikipedia for Semantic Relatedness. In: ACL Workshop"TextGraphs-4: Graph-based Methods for Natural Language Processing (2009)
31. Milne, D.: Computing Semantic Relatedness using Wikipedia Link Structure. In: Proc. of NZ CSRSC 2007 (2007)
32. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, India (January 2007)
33. Harrington, B.: A semantic network approach to measuring relatedness. In: Proceedings of COLING 2010 (2010)
34. Wojtinnek, P., Pulman, S.: Semantic relatedness from automatically generated semantic networks. In: Proceedings of the Ninth International Conference on Computational Semantics, IWCS 2011 (2011)
35. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., Soroa, A.: A Study on Similarity and Relatedness Using Distributional and WordNet based Approaches. In: Proceedings of NAACL 2009 (2009)
36. Halavais, A.: An Analysis of Topical Coverage of Wikipedia. Journal of Computer-Mediated Communication 13(2) (2008)
37. Gouws, S., Rooyen, G., Engelbrecht, H.: 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 (2010)
38. Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E., Milios, E.: Semantic similarity methods in WordNet and their application to information retrieval on the web. In: 7th ACM Intern. Workshop on Web Information and Data Management (WIDM 2005), Bremen, Germany (2005)
39. Jarmasz, M.: Roget's thesaurus as a lexical resource for natural language processsing. Master's thesis, University of Ottawa (2003)
40. Hadj Taieb, M., Ben Aouicha, M., Tmar, M., Ben Hamadou, A.: New WordNet-based semantic relatedness measurement using new information content metric and k-means clustering algorithm. In: Global WordNet Conference, Matsue, Japan (2012)
41. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language and Cognitive Processes 6(1), 1–28 (1991)
42. Rubenstein, H., Goodenough, J.: Contextual correlates of synonymy. Communications of the ACM 8(10), 627–633
43. Ross, S.: A first Course in Probability. Macmillan (1976)
44. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. JASIS 41(6) (1990)