This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

Computing Semantic Similarity of Concepts in Knowledge Graphs

Ganggao Zhu and Carlos A. Iglesias

Abstract—This paper presents a method for measuring the semantic similarity between concepts in Knowledge Graphs (KGs) such as WordNet and DBpedia. Previous work on semantic similarity methods have focused on either the structure of the semantic network between concepts (e.g. path length and depth), or only on the Information Content (IC) of concepts. We propose a semantic similarity method, namely wpath, to combine these two approaches, using IC to weight the shortest path length between concepts. Conventional corpus-based IC is computed from the distributions of concepts over textual corpus, which is required to prepare a domain corpus containing annotated concepts and has high computational cost. As instances are already extracted from textual corpus and annotated by concepts in KGs, graph-based IC is proposed to compute IC based on the distributions of concepts over instances. Through experiments performed on well known word similarity datasets, we show that the wpath semantic similarity method has produced statistically significant improvement over other semantic similarity methods. Moreover, in a real category classification evaluation, the wpath method has shown the best performance in terms of accuracy and F score.

Index Terms—Semantic Similarity, Semantic Relatedness, Information Content, Knowledge Graph, WordNet, DBpedia

1 Introduction

With the increasing popularity of the linked data initiative, many public Knowledge Graphs (KGs) have become available, such as Freebase [1], DBpedia [2], YAGO [3], which are novel semantic networks recording millions of concepts, entities and their relationships. Typically, nodes of KGs consist of a set of concepts $C_1, C_2, \ldots, C_n$ representing conceptual abstractions of things, and a set of instances $I_i, I_2, \ldots, I_m$ representing real world entities. Following Description Logic terminology [4], knowledge bases contain two types of axioms: a set of axioms is called a terminology box (TBox) that describes constraints on the structure of the domain, similar to the conceptual schema in database setting, and a set of axioms is called assertion box (ABox) that asserts facts about concrete situations, like data in a database setting [4]. Concepts of the KG contains axioms describing concept hierarchies and are usually refereed as ontology classes (TBox), while axioms about entity instances are usually referred as ontology instances (ABox). Fig. 1 shows a tiny example of a KG using the above notions. Concepts of TBox are constructed hierarchically and classify entity instances into different types (e.g., actor or movie) through a special semantic relation *rdf:type*[1](e.g., dbr:Star_Wars is a instance of concept movie). Concepts and hierarchical relations (e.g., is-a) compose a concept taxonomy which is a concept tree where nodes denote the concepts and edges denote the hierarchical relations. The hierarchical relations between concepts specify that a concept C_i is a kind of concept C_j (e.g., *actor* is a *person*). Apart from hierarchical relationships, concepts can have other semantic relationships among them (e.g., *actor* plays in a *movie*). Note that the tiny KG is a simplified example from DBpedia for illustration, and Table 1 shows examples of DBpedia entities and their types which are mapped to the example KG in Fig. 1.

TABLE 1
The Examples of Mapped Entities and Entity Types in DBpedia.

Entity	Type	Concept
dbr:Star_Wars	yago:Movie106613686, dbo:Film	Movie
dbr:Don_Quixote	yago:Novel106367879, dbo:Book	Novel
dbr:Tom_Cruise	yago:Actor109765278, dbo:Actor	Actor
dbr:Apple_Inc	yago:Company108058098, dbo:Company	Company

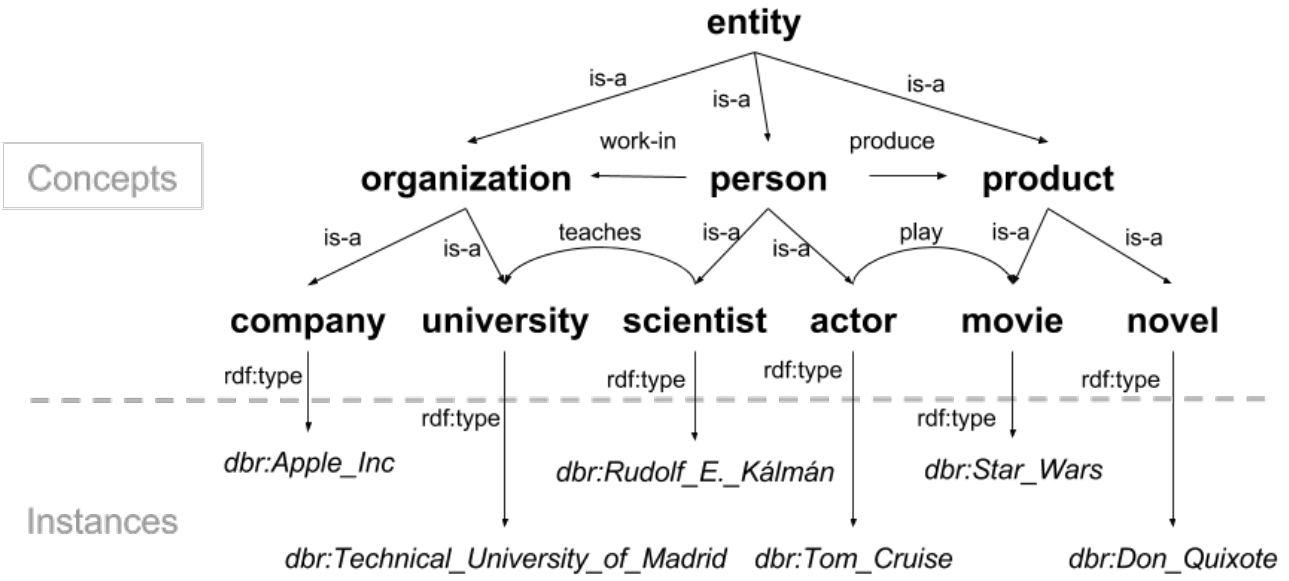
The lexical database WordNet [5] has been conceptualized as a conventional semantic network of the lexicon of English words. WordNet can be viewed as a concept taxonomy where nodes denote WordNet synsets representing a set of words that share one common sense (synonyms), and edges denote hierarchical relations of hypernym and hyponymy (the relation between a sub-concept and a super-concept) between synsets. Recent efforts have transformed WordNet to be accessed and applied as concept taxonomy in KGs by converting the conventional representation of WordNet into novel linked data representation. For example, KGs such as DBpedia, YAGO and BabelNet [6] have integrated WordNet and used it as part of concept taxonomy to categorize entity instances into different types. Such integration of conventional lexical resources and novel KGs have provided novel opportunities to facilitate many different Natural Language Processing (NLP) and Information Retrieval (IR) tasks [7], including Word Sense Disambiguation (WSD) [8], [9], Named Entity Disambiguation (NED) [10], [11], query interpretation [12], document modeling [13] and question answering [14] to name a few. Those KG-based applications rely on the knowledge of concepts, instances and their

• *G. Zhu and C.A Iglesias are with Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Avda. Complutense, 30 28040 Madrid, Spain*
E-mail: gzhu@dit.upm.es, cif@gsi.dit.upm.es

1. We abbreviate URI namespaces with common prefixes, see http://prefix.cc for details.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

Fig. 1. A Tiny Example of Knowledge Graph



relationships. In this work, we mainly exploit the concept level knowledge, while the instance level knowledge is used to support the concept knowledge. More specifically, we focus on the problem of computing the semantic similarity between concepts in KGs.

In computational linguistics, semantic similarity is a metric that represents the commonality of two concepts relying on their hierarchical relations [15], [16]. Semantic similarity is a special case of semantic relatedness which does not necessarily rely on hierarchical relations. For example, as shown in the tiny example of KG in Fig. 1, *scientist* and *actor* are semantically similar because they share the hypernym *person*. Although *actor* and *movie* are clearly related, but they are not really similar because they belong to different branches of taxonomy. Semantic relatedness usually has wider computational applications because it considers all kinds of semantic relations between concepts. Semantic similarity would be more useful when applications need to encode hierarchical relations between concepts, such as concept expansion and concept-based retrieval. In general, semantic similarity metrics can be used for weighting or ranking similar concepts based on a concept taxonomy. In such way, semantic similarity methods could be applied in KGs for concept-based entity retrieval or question answering, where those entities that contain types having similar meaning to query concepts would be retrieved. Moreover, in entity modeling, semantic similarity could be used to cluster entities based on their type concepts.

Some of the conventional semantic similarity metrics [17], [18], [19], [20] rely on measuring the semantic distance between concepts using hierarchical relations. Semantic similarity between two concepts is then proportional to the length of the path connecting the two concepts. Path based similarity metric requires the structure of semantic network to generate a similarity score that quantifies the degree of similarity between two concepts. Concepts that are physically close to each other in taxonomy are considered to be more similar than those concepts that are located far away. For instance, the concept *actor* is more similar to the concept *scientist* than the concept *movie*, because *actor* and *scientist* are located closer in concept taxonomy. Some other semantic similarity metrics [15], [21], [22] consider statistical Information Content (IC) of concepts computed from corpora in order to improve the performance of similarity metrics that are only based on the structure of concept taxonomy. IC is a measure of specificity of a concept. The higher values of IC are associated with more specific concepts (e.g., actor), while those lower values are more general (e.g., person). IC is computed based on frequency counts of concepts appearing in a textual corpus. Each occurrence of a more specific concept also implies the occurrence of the more general ancestor concepts. In order to alleviate the weaknesses of both path based metrics and IC based metrics, we propose a novel semantic similarity method, namely wpath, combining the two methods. Moreover, in order to adapt corpus-based IC methods to structured KGs, we propose a graph-based IC computation method, which can enable those semantic similarity metrics using IC to be used based on KGs without offline preparation of domain corpus. Within the graph-based IC, the wpath semantic similarity method can be used to compute semantic similarity between concepts in KGs only based on the structural knowledge of concepts and the statistical knowledge of instances in KGs.

According to the evaluation experiments in word similarity datasets, compared with the previous state of the art semantic similarity methods, the wpath method results in statistical significant improvement of correlation between computed similarity scores and human judgements. The proposed graph-based IC has shown to be effective as the corpus-based IC so that it could be used as the substitution of the corpus-based IC in KGs. Furthermore, in order to evaluate the performance of semantic similarity methods in real application datasets, we have applied semantic similarity metrics to the aspect category classification task [23], [24]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

of the restaurant domain. The evaluation results of semantic similarity based category classification have shown that the wpath semantic similarity method has the best accuracy and F-measure score.

In conclusion, this paper considers the problem of measuring semantic similarity between concepts in KGs. The main contributions of this work may be summarized as below.

- We propose a method for measuring the semantic similarity between concepts in KGs.
- We propose a method to compute IC based on the specificity of concepts in KGs.
- We evaluate the proposed methods in gold standard word similarity datasets.
- We evaluate the semantic similarity methods in aspect category classification.

The rest of this paper is organized as follows. In Section 2, we review the conventional semantic similarity methods and formalise the semantic information used in those methods. Section 3 presents the wpath semantic similarity method and the graph-based IC computation method. Section 4 reports the evaluation experiments and summarises the evaluation results. Finally, we draw conclusions and outline aspects of future work in Section 5.

2 Semantic Similarity

There is a relatively large number of semantic similarity metrics which were previously proposed in the literatures. Among them, there are mainly two types of approaches in measuring semantic similarity, namely corpus-based approaches and knowledge-based approaches [25]. Corpus-based semantic similarity metrics are based on models of distributional similarity learned from large text collections relying on word distributions. Two words will have a high distributional similarity if their surrounding contexts are similar. Only the occurrences of words are counted in corpus without identifying the specific meaning of words and detecting the semantic relations between words. Since corpus-based approaches consider all kinds of lexical relations between words, they mainly measure semantic relatedness between words. On the other hand, knowledge-based semantic similarity methods are used to measure the semantic similarity between concepts based on semantic networks of concepts. This section reviews briefly corpus-based approaches (Section 2.1) and knowledge-based semantic similarity metrics that have been observed good performance in NLP or IR applications (Section 2.2).

2.1 Corpus-based Approaches

Corpus-based approaches measure the semantic similarity between concepts based on the information gained from large corpora such as Wikipedia. Following this idea, some works exploit concept associations such as Pointwise Mutual Information [26] or Normalised Google Distance [27], while some other works use distributional semantics techniques to represent the concept meanings in high-dimensional vectors such as Latent Semantic Analysis [28] and Explicit Semantic Analysis [29]. Recent works based on distributed semantics techniques consider advanced computational models such as Word2Vec [30] and GLOVE [31], representing the words or concepts with low-dimensional vectors.

The co-occurrence information of words with the same surrounding context would make a wide variety of words to be considered as related. Since corpus-based approaches mainly rely on contextual information of words, they usually measure the general semantic relatedness between words rather than the specific semantic similarity that depends on hierarchical relations [16]. Furthermore, corpus-based semantic similarity methods represent concepts as words without clarifying their different meanings (word senses). Compared to knowledge-based approaches relying on KGs, corpus-based approaches normally have better coverage of vocabulary because their computational models can be effectively applied to various and updated corpora. Since they are modeled based on words and textual corpora rather than concept taxonomies, we briefly touch on the corpus-based methods and present a detailed review of the main knowledge-based methods in the following section.

2.2 Knowledge-based Approaches

Knowledge-based approaches measure the semantic similarity of concepts in KGs. We first give a formal definition of KG.

Definition 1. A KG is defined as a directed labeled graph, $G = (V, E, \tau)$, where V is a set of nodes, E is a set of edges connecting those nodes; and τ is a function $V \times V \rightarrow E$ that defines all triples in G.

Given a KG, knowledge-based approaches measure the semantic similarity between concepts $c_1, c_2 \in V$, formally $sim(c_1, c_2)$, using semantic information contained in KG. The most intuitive semantic information is the semantic distance between concepts, which is usually represented by the path connecting two concepts in KG. Intuitively, the shorter the path from one concept to another, the more similar they are.

Definition 2. A path $P(c_i, c_j)$ between $c_i, c_j \in V$ through G is a sequence of nodes and edges $P(c_i, c_j) = \{c_i, e_i, \ldots, v_k, e_k, v_{k+1}, e_{k+1}, \ldots, c_j\}$ connecting the concepts c_i and c_j with cardinality or size n. For every two consecutive nodes $v_k, v_{k+1} \in V$ in $P(c_i, c_j)$, there exists an edge $e_k \in E$.

Note that though KG is modeled as a directed graph we do not consider the direction of edges because semantic relations can be considered to have semantically sound inverse relation [11]. Let $Paths(c_i, c_j) = \{P_1, P_2, \ldots P_n\}$ be the set of paths connecting the concepts c_i and c_j with cardinality or size N. Let $|P_i|$ denote the length of a path $P_i \in Paths(c_i, c_j)$, then $length(c_i, c_j) = \min\limits_{1 \leq i \leq N}(|P_i|)$ denotes the shortest path length between two concepts. The **path** [17] method uses the shortest path length between concepts to represent their semantic distance and the distance can be transformed into similarity as expressed in Eq.(1):

$$sim_{path}(c_i, c_j) = \frac{1}{1 + length(c_i, c_j)} \tag{1}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

The **lch** [18] method measures the semantic similarity between concepts based on their shortest path length using a non-linear function illustrated in Eq.(2):

$$sim_{lch}(c_i, c_j) = -\log\left(\frac{length(c_i, c_j)}{2 * D}\right) \quad (2)$$

where D is the maximum depth of the concept taxonomy in a KG. The path between the root concept and a given concept through hierarchical relations is called depth, given that KGs contain concepts which can be organised as a concept taxonomy with hierarchical relations, such as WordNet taxonomy, DBpedia ontology class to name a few.

Definition 3. The $depth(c_i) = length(c_i, c_{root})$ of a concept $c_i \in V$ is defined as the shortest path length from c_i to root concept $c_{root} \in V$. For every two consecutive nodes $v_k, v_{k+1} \in P(c_i, c_{root})$, there exists an edge $e_k \in \{hypernym, subClassOf\}$.

The idea of using depth information of concepts to measure the semantic similarity lies in the property of concept taxonomies that the upper-level concepts in a taxonomy are supposed to be more general. Therefore, the similarity between lower-level concepts should be considered more similar than those concepts between upper-level concepts. For example in Fig. 1, the concept pair *scientist* and *actor* are more similar than the concept pair *person* and *product*.

The Least Common Subsumer (LCS) is the most specific concept that is a shared ancestor of the two concepts. For example, the LCS of concept *scientist* and concept *actor* is the concept *person*. Let c_{lcs} be the LCS of concepts c_i and c_j, the **wup** [19] method measures semantic similarity of given concepts using the following formula:

$$sim_{wup}(c_i, c_j) = \frac{2depth(c_{lcs})}{depth(c_i) + depth(c_j)} \quad (3)$$

The **li** method [20] combines the shortest path length and the depth of LCS. It measures semantic similarity using a non-linear functions as shown in Eq.(4).

$$sim_{li}(c_i, c_j) = e^{-\alpha length(c_i, c_j)} \cdot \frac{e^{\beta depth(c_{lcs})} - e^{-\beta depth(c_{lcs})}}{e^{\beta depth(c_{lcs})} + e^{-\beta depth(c_{lcs})}} \quad (4)$$

where e is the Euler's number and α, β are parameters that contribute to the path length and depth respectively. According to the experiment of **li** [20], the empirical optimal parameters are $\alpha = 0.2$ and $\beta = 0.6$.

Some other knowledge-based semantic similarity methods [21], [22], [32] include IC of concepts to improve performance of measuring semantic similarity. The definition of corpus-based IC proposed in [15] is presented in Definition 4.

Definition 4. The $IC_{corpus}(c_i)$ of a concept $c_i \in V$ is defined as: $IC_{corpus}(c_i) = -logProb(c_i)$, where $Prob(c_i)$ denotes the probability of encountering the set of $words(c_i)$ subsumed by concept c_i. Let $freq_{corpus}(c_i) = \sum_{w \in words(c_i)} count(w)$ be the frequency of concept c_i occurs in corpus, then $Prob(c_i) = \frac{freq_{corpus}(c_i)}{N}$ where N is the total number of concepts observed in corpus.

The quantitative characteristic of IC is that the more abstract concepts have lower value of IC and more specific concepts have higher value of IC. If two concepts share a more specific concept, it means that they share more information and thus more similar because the IC of their LCS is higher. Based on this intuition, the **res** [15] method measures the semantic similarity of two concepts using the IC of their LCS which is illustrated in Eq.(5).

$$sim_{res}(c_i, c_j) = IC_{corpus}(c_{lcs}) \quad (5)$$

The **lin** [22] method extends the **res** method by computing the similarity between concepts as the ratio between the IC of LCS and their own ICs:

$$sim_{lin}(c_i, c_j) = \frac{2IC_{corpus}(c_{lcs})}{IC_{corpus}(c_i) + IC_{corpus}(c_j)} \quad (6)$$

Similarly, the **jcn** [21] method measures the difference between concepts by subtracting the sum of the IC of each concept alone from the IC of their LCS.

$$dis_{jcn}(c_i, c_j) = IC_{corpus}(c_i) + IC_{corpus}(c_j) - 2IC_{corpus}(c_{lcs}) \quad (7)$$

It can be transformed from distance $dis_{jcn}(c_i, c_j)$ to similarity $sim_{jcn}(c_i, c_j)$ by computing the reverse of distance:

$$sim_{jcn}(c_i, c_j) = \frac{1}{1 + dis_{jcn}(c_i, c_j)} \quad (8)$$

The knowledge-based semantic similarity metrics presented above are reported having good performance in measuring the semantic similarity between concepts in WordNet. They are still applicable to measure conceptual semantic similarity in KGs as WordNet concept taxonomy has been integrated into KGs such as DBpedia, YAGO, BabelNet.

Apart from conceptual similarity, some recent works started to propose semantic relatedness or similarity metrics for instances in KGs, where wide-coverage of fine-grained semantic relations between instances are provided. The path-based semantic relatedness method [11] between instances uses a social network analysis technique to measure the effectiveness of a path connecting instances, together with the *exclusivity* metric that specifies the relative importance of a relation connecting two instances. It follows two principles in measuring semantic relatedness: 1) the shorter path between instances the higher their relatedness; 2) two instances are more related if the relations connecting them are relatively more important. The *Concept Association* [33] computes the statistical association between instances based on the occurrence and cooccurrence of nodes and edges in KGs. The Combined Information Content (CombIC) [13] proposed a information content method to derive the weights for edges (property) in DBpedia. These recently proposed semantic relatedness methods are focused on instance relatedness rather than semantic similarity between concepts. As we mentioned previously, concept similarity is usually based on taxonomical relations between concepts such as WordNet taxonomy and DBpedia ontology class. Semantic similarity between concepts in KG can be extended to instances in order to measure the semantic similarity between instances, because concepts can be viewed as semantic classes of instances. For example, the semantic similarity between the instances A and B (e.g. *dbr:Star_Wars*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

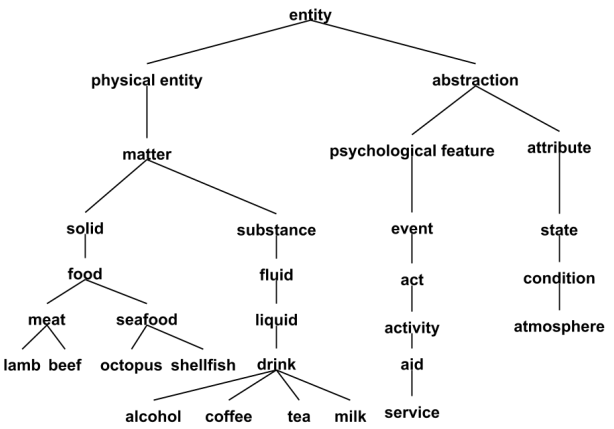


Fig. 2. A Fragment of WordNet Concept Taxonomy

and *dbr:Don_Quixote*) can be measured by calculating the semantic similarity of their respective types (e.g. dbo:Film and dbo:Book). The main goal of this paper is to propose a semantic similarity metric for concepts in KGs which is introduced in the following sections.

3 The Proposed Methods

The main idea of the wpath semantic similarity method is to encode both the structure of the concept taxonomy and the statistical information of concepts. Furthermore, in order to adapt corpus-based IC methods to structured KGs, graph-based IC is proposed to compute IC based on the distribution of concepts over instances in KGs. Consequently, using the graph-based IC in the wpath semantic similarity method can represent the specificity and hierarchical structure of the concepts in a KG. Section 3.1 presents the wpath semantic similarity method for measuring semantic similarity between concepts in KGs and Section 3.2 describes the proposed method to compute graph-based IC of concepts based on KGs.

3.1 WPath Semantic Similarity Metric

The knowledge-based semantic similarity metrics mentioned in the previous section are mainly developed to quantify the degree to which two concepts are semantically similar using information drawn from concept taxonomy or IC. Metrics take as input a pair of concepts, and return a numerical value indicating their semantic similarity. Many applications rely on this similarity score to rank the similarity between different pairs of concepts. Take a fragment of WordNet concept taxonomy in Fig. 2 as example, given the concept pairs of $(beef, lamb)$ and $(beef, octopus)$, the applications require similarity metrics to give higher similarity value to $sim(beef, lamb)$ than $sim(beef, octopus)$ because the concept *beef* and concept *lamb* are kinds of *meat* while the concept *octopus* is a kind of *seafood*. The semantic similarity scores of some concept pairs computed from the semantic similarity methods have been illustrated in Table. 2. It can be seen in this table how the row of concept pair $(beef, lamb)$ has higher similarity scores than the row of concept pair $(beef, octopus)$.

TABLE 2
The Illustration of Semantic Similarity Methods on Some Concept Pair Examples

Concept Pairs	path	lch	wup	li	res	lin	jcn	wpath
beef - octopus	0.200	2.028	0.714	0.442	6.109	0.484	0.071	0.494
beef - lamb	0.333	2.539	0.857	0.667	6.725	0.591	0.097	0.692
meat - seafood	0.333	2.539	0.833	0.659	6.109	0.760	0.205	0.662
octopus - shellfish	0.333	2.539	0.857	0.667	9.360	0.729	0.125	0.801
beef - service	0.071	0.999	0.133	0.000	0.000	0.000	0.050	0.071
beef - atmosphere	0.083	1.153	0.154	0.000	0.000	0.000	0.052	0.083
beef - coffee	0.111	1.440	0.429	0.168	3.337	0.319	0.066	0.208
food - coffee	0.143	1.692	0.500	0.251	3.337	0.411	0.095	0.260

One of the drawbacks of conventional knowledge-based approaches (e.g. path or lch) in addressing such task is that the semantic similarity of any two concepts with the same path length is the same (uniform distance problem). As illustrated in Fig. 2 and Table. 2, based on the path and lch semantic similarity methods, $sim(meat, seafood)$ is the same as $sim(beef, lamb)$ and $sim(octopus, shellfish)$ because those concept pairs have equal shortest path length. Some knowledge-based approaches (e.g. wup or li) tried to solve the drawback by including depth information in concept taxonomy. Considering that the upper level concepts are more general than the lower level concepts in hierarchy, those approaches use the depth of concepts to give higher similarity value to those concept pairs which are located deeper in taxonomy. For example, the similarity of $(beef, lamb)$ is higher than the similarity of $(meat, seafood)$ based on semantic similarity method of wup and li, because the concept *lamb* and the concept *beef* are located deeper in the concept taxonomy (*lamb* and *beef* are sub-concepts of *meat*). Though using depth has been reported performance improvement compared to pure path length methods, for a given taxonomy such as the one in Fig. 2, many concepts share the same depth (hierarchical level) resulting in same similarity. For instance, as shown in Table. 2, based on the semantic similarity methods of wup and li, $sim(lamb, beef)$ is equal to $sim(octopus, shellfish)$ because of the same depth.

In order to solve the equal path length and depth problem, some knowledge-based approaches (e.g. res, lin, or jcn) proposed to include IC because different concepts usually have different IC values (e.g, the IC of *meat* is 6.725 and the IC of *food* is 6.109) so that the $sim(lamb, beef)$ is different from $sim(octopus, shellfish)$. Note that the IC in this section is based on corpus-based IC and its implementation details is described in Section 4.1.1. IC is a statistical method to measure the informativeness of concept. General concepts have lower informativeness thus have lower value of IC, while more specific concepts would have higher value of IC. For example, the IC of *meat* is higher than the IC of *food* because *meat* is a sub-concept of *food*. The idea of using IC to compute semantic similarity is that the more information two concepts share in common, the more similar they are. Using the IC of the LCS alone in the res method can represent the common information that two concepts share, however, the problem is that the similarity of any two concepts with the same LCS is the same. For example, based on res semantic similarity, although the concept pairs $(beef, lamb)$ and $(octopus, shellfish)$ have different similarity scores, the similarity scores of concept pairs $(meat, seafood)$ and $(beef, octopus)$, $(beef, coffee)$ and $(food, coffee)$ are the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

same because the LCS of the concept pairs are concept *food* and *matter*. Other methods (e.g., lin or jcn) tried to solve the drawback by including the IC of concepts being compared. However, only using the informativeness of concepts to represent the difference between concepts may lose the valuable distance information between concepts provided by the human experts who have created the concept taxonomy. It has been shown in our preliminary experiments that the path length between concepts in a taxonomy is a very effective feature in measuring semantic similarity of concepts. Furthermore, when the LCS of the concept pairs is the root concept *entity*, the li, res, and lin methods fail by generating 0 similarity score such as concept pairs $(beef, service)$ and $(beef, atmosphere)$. In addition, the lin and jcn methods are still missing the hierarchical level information. For instance, since the concept pairs $(meat, seafood)$ are more general than $(octopus, shellfish)$, the $(meat, seafood)$ is assumed to be less similar, however, the lin and jcn methods have given higher similarity score.

Considering both advantages and disadvantages of conventional knowledge-based semantic similarity methods, we propose a weighted path length (wpath) method to combine both path length and IC in measuring the semantic similarity between concepts. The IC of two concepts' LCS is used to weight their shortest path length so that those concept pairs having same path length can have different semantic similarity score if they have different LCS. The wpath semantic similarity method is illustrated in Eq.(9):

$$sim_{wpath}(c_i, c_j) = \frac{1}{1 + length(c_i, c_j) * k^{IC(c_{lcs})}} \quad (9)$$

where $k \in (0, 1]$ and $k = 1$ means that IC has no contribution in shortest path length. The parameter k represents the contribution of the LCS's IC which indicates the common information shared by two concepts.

The proposed method aims to give different weights to the shortest path length between concepts based on their shared information, where the path length is viewed as difference and the common information is viewed as commonality. For identical concepts, their path length is 0 so their semantic similarity reaches the maximum similarity 1. As the path length between concepts in the concept taxonomy becomes bigger (bigger value of path length), the semantic similarity between concepts becomes smaller. The similarity score of the wpath is ranged in $(0, 1]$, which has improved the similarity score range in lch method and res method.

When the concepts have the same distance (equal path length), the more information two concepts share, the more similar they are. As shown in Table. 2, based on the wpath method, the similarity score of $(beef, lamb)$, $(meat, shellfish)$ and $(octopus, shellfish)$ are different based on their shared information, which shows the improvement of wpath over the path, lch, wup, and li methods. Although the wpath method is missing the depth, the LCS actually denotes the hierarchical level in taxonomy implicitly. Specifically, IC is a statistical method exploiting statistical occurrence information of concept, and the IC of LCS is similar to depth of concept indicating that the deeper level of concepts in the taxonomy are more specific, thus they are more similar. Moreover, concept's IC includes frequency of concepts so it has more various values than depth.

Since IC based metrics (e.g. res, lin and jcn) do not deal with the hierarchy of concepts, similarity scores computed by them lack of information about hierarchical levels and conceptual distance. As structural knowledge is retained in the wpath method, it is able to give higher similarity score to more specific concepts, but also give higher similarity score to those concepts sharing the same IC and located closer in taxonomy. In the example of $(beef, octopus)$, $(meat, seafood)$, since they share the same IC and $(meat, seafood)$ locates closer in the taxonomy, the wpath method has given higher similarity score to $(meat, seafood)$ than $(beef, octopus)$, which shows improvement of the wpath method over res method. The example of $(octopus, shellfish)$ and $(meat, seafood)$ shows that the wpath method has solved the hierarchical level problem of lin and jcn methods by giving higher similarity score to more specific concept pair when two concept pairs have the same path length.

In conclusion, the wpath semantic similarity method takes advantage of structure based methods (e.g. path, lch, wup and li) in representing the distance between concepts in a taxonomy, and overcomes the equal path and depth problem that would result in equal similarity scores for many concept pairs. By using the shared information (IC) between concepts to weight their path length, the wpath not only can retain the ability to show the distance between concepts based on a taxonomy, but also can acquire statistical information to tell the commonality between concepts when their conceptual structures in taxonomy are same.

The IC function in Eq.(9) denotes the general purpose IC which is used as weight for path length. According to different application scenarios, the IC function can either be the corpus-based IC (Definition 4) or the graph-based IC (Definition 5) which will be introduced in the following section.

3.2 Graph-Based Information Content

Conventional corpus-based IC requires to prepare a domain corpus for the concept taxonomy and then to compute IC from the domain corpus in offline. The inconvenience lies in the high computational cost and difficulty of preparing a domain corpus. More specifically, in order to compute corpus-based IC, the concepts in the taxonomy need to be mapped to the words in the domain corpus. Then the appearance of concepts are counted and the IC values for concepts are generated. In this way, the additional domain corpus preparation and offline computation may prevent the application of those semantic similarity methods relying on the IC values (e.g., res, lin, jcn, and wpath) to KGs, especially when the domain corpus is insufficient or the KG is frequently updated. Since KGs already mined structural knowledge from textual corpus, we present a convenient graph-based IC computation method for computing the IC of concepts in a KG based on the instance distributions over the concept taxonomy. The graph-based IC is proposed to directly take advantage of KGs while retaining the idea of corpus-based IC representing the specificity of concepts. In consequence, the IC-based semantic similarity methods

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

such as res, lin, jcn and the proposed wpath can compute the similarity score between concepts directly relying on the KG.

As we mentioned previously in Section 1, concepts in KGs are usually represented as TBox and arranged into concept taxonomies. Those concepts categorize entity instances of ABox into different types via the special relation rdf:type. For example, the concept *movie* groups all movie instances in DBpedia. Moreover, if concept A is a parent concept of concept B and concept C in the taxonomy, then the set of instances of A is the union of the instances of B and C. In other words, a concept in KG can have multiple entity instances indicating the semantic type of those entities, while an instance can have multiple concepts to describe entity categories from general to specific. For instance, a DBpedia entity instance *dbr:Tom_Cruise* can have several concepts describing its types from general to specific, *Person*, *Actor*, *AmericanFilmActor*.

Intuitively, more general concepts occur more frequently in a KG such as concepts *organization*, *person* in the tiny example of KG in Fig. 1, while more specific concepts occur less frequently such as concepts *actor*, *university*, *scientist* and many others. Therefore, the proportion of the instances belonging to a specific concept in a KG can be used to measure the specificity of the concept for the given KG, which is similar to the idea of IC. As introduced in Section 2.2, IC measures the specificity (informativeness) of a concept over a corpus. Similar to the definition of conventional corpus-based IC, we extend the definition of IC in [15] to KGs.

Definition 5. The graph-based IC in a KG is $IC_{graph}(c_i) = -logProb(c_i)$ where $Prob(c_i) = \frac{freq_{graph}(c_i)}{N}$. N denotes the total number of entities in the KG. Let $entities(c_i)$ be the function to retrieve set of entities having type of c_i, the frequency of concept c_i in the KG is defined as $freq_{graph}(c_i) = count(entities(c_i))$ where $count(x)$ is a simple counting function measuring the cardinality of a set of entities.

The above definition of graph-based IC has defined the distribution of concepts over all the instances in KG. In particular, entity instances in KG are viewed as document collections and each instance denotes a document, while a collection of concepts describing each instance are viewed as terms in a document. Then the graph-based IC is computed as the frequency of those concepts over the document collections, whose idea is similar to the idea of Inverse Document Frequency (IDF) [34] in IR, and the difference is the mathematical form. Both graph-based IC and IDF treat the less frequent concepts with higher importance. Since concepts in a taxonomy have hierarchical relations, the less frequent concepts specify more specific concepts.

Corpus-based IC methods may contain ambiguous meaning of concepts because it calculates IC by counting the occurrence of words over textual corpus, where words can be mapped to multiple concepts (ambiguous words). In comparison, graph-based IC contains specific meaning of concepts since KGs usually contain disambiguated concepts to describe types of instances. Furthermore, similar to corpus-based IC, graph-based IC can be used in semantic similarity methods which need to employ ICs such as the res, lin, jcn and wpath similarity methods. If the LCS of two concepts appears less frequently in a KG, it means that two concepts are more similar. Using graph-based IC enables semantic similarity methods to compute semantic similarity between concepts only based on the specific KG without relying on additional corpora.

Moreover, it is more convenient to compute graph-based IC than conventional IC. Since instances are linked to concepts through the rdf:type relation in a structured representation, it is convenient to retrieve all the entities in a KG belonging to a specific concept using structured query languages such as SPARQL[2]. This could be considered as online computation compared to the corpus-based IC that is required to compute in offline from textual corpus. Suppose that the SPARQL query language is implemented in the KG management system and the ontology classes are described using OWL[3], the total number of entities N in the KG can be acquired using the following SPARQL query.

```
SELECT count(?e) as ?e WHERE
{
   ?e rdf:type owl:Thing .
}
```

In addition, the function $freq_{graph}(c_i)$ can be implemented using the following SPARQL query.

```
SELECT count(?e) as ?e WHERE {
   ?e rdf:type owl:Thing .
   ?e rdf:type <concept_uri> .
}
```

The *concept_uri* denotes the URI link of the specific concept in the KG. Within the Definition 5 and the SPARQL implementation of graph-based IC, it is convenient to compute the IC of a specific concept based on a KG. Note that the above SPARQL implementation is just an illustrative example, and the similar online computation of graph-based IC can be achieved by accessing a knowledge management system. In addition, apart from being used in semantic similarity methods, graph-based IC can also be used in other KG-based applications such as selecting the most specific type of a given entity. Furthermore, the definition of graph-based IC can be applied to conventional document analysis domain where the documents are tagged with hierarchical concepts such as the Open Directory Project[4], the Medical Subject Headings[5], the ACM Term Classification[6] and many others. This paper focuses on applying graph-based IC in semantic similarity methods and leave its other applications as future work.

In summary, graph-based IC is proposed to be a possible substitution or complementary for the conventional corpus-based IC when the domain corpus is insufficient or online computation of IC is required. For those domains already containing annotated corpus such as Brown Corpus [35] for WordNet, corpus-based IC could be used if it performs well in similarity metrics for domain applications. According to our experiments, graph-based IC is as effective as corpus-based IC although it is not outperforming, thus graph-based

2. https://www.w3.org/TR/sparql11-query/
3. https://www.w3.org/TR/owl-ref/
4. https://www.dmoz.org/
5. https://www.nlm.nih.gov/mesh/
6. https://www.acm.org/publications/class-2012

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

IC could be considered as a trade off of the efficiency, convenience and effectiveness, with corpus-based IC. Having introduced the proposed semantic similarity method and graph-based IC, we then present their evaluation in the following section.

4 Experiments

The goal of our experiments is to evaluate the proposed semantic similarity method and graph-based IC in KGs. However, to best of our knowledge, currently there is no standard method and dataset to evaluate the performance of semantic similarity method and IC computation model for concepts in KG. Therefore, the commonly used word similarity datasets are used to evaluate the proposed semantic similarity method and graph-based IC based on WordNet and DBpedia. Moreover, the semantic similarity methods are evaluated in an aspect category classification task [23], [24] in order to evaluate their performance in a real application. This section presents the datasets, implementation, evaluation and provides a brief discussion about the obtained experimental results.

4.1 Word Similarity Evaluation

This section presents the evaluation of semantic similarity methods and graph-based IC in word similarity task.

4.1.1 Datasets and Implementation

We collected several publicly available gold standard datasets for evaluating word semantic similarity, which are conventionally most commonly used and some recently most updated datasets. The description of collected datasets used in experiment are listed below.

- **R&G** [36] is the first and most used dataset containing human assessment of word similarity. The dataset resulted from the experiment conducted in 1965 where a group of 51 students (all native English speakers) assessed the similarity of 65 pairs of words selected from ordinary English nouns. Those 51 subjects were requested to judge the *similarity of meaning* for two given words on a scale from 0.0 (completely dissimilar) to 4.0 (highly synonymous). It focused on semantic similarity and ignored any other possible semantic relationships between the words.
- **M&C** [37] replicated the **R&G** experiment again in 1991 by taking a subset of 30 noun pairs. The similarity between words was judged by 38 human subjects.
- **WS353** [38] contains 353 pairs of words and 13 to 16 human subjects were asked to assign a numerical similarity score between 0.0 to 10.0 (0=totally unrelated and 10=very closely related). In fact, this dataset measures general relatedness rather than similarity because it considers other semantic relations (e.g. antonyms are considered as similar). We used this dataset because it has been perhaps the most commonly-used gold standard dataset for semantic similarity recently.
- **WS353-Sim** [39] contains 203 pairs of words and is the subset of **WS353**. It has been identified by the authors to be suitable for evaluating semantic similarity specially.
- **SimLex** [40] is a recently released dataset consisting of 999 word pairs for evaluating semantic similarity specially. The dataset contains 111 adjective pairs (A), 666 noun pairs (N), and 222 verb pairs (V). We used 666 noun pairs in our experiment. Each pair of words was rated by at least 36 subjects (native English speakers) with similarity scores on scale from 0.0 (no similarity) to 10.0 (exactly mean same thing) and the average score was assigned as final human judgment score.

All the datasets described above contain a list of triples comprising two words and a similarity score denoting word similarity judged by human subjects. The human ratings on those word pairs have been proven to be highly replicable. The correlation obtained from M&C with respect to R&Gs experiment was 0.97. [15] replicated the M&C's experiment again in 1995, involving 10 computer science graduate students and post-doc researchers to assess similarity. The correlation with respect to the M&C's results was 0.96. This indicates that human assessment about semantic similarity between words is remarkably stable over a large time span and such datasets containing human ratings can be reliably used for evaluating semantic similarity methods. Since those datasets contain different coverage of word pairs, we use all the datasets for evaluation in order to present a more completed and objective experiment.

Those datasets are used for evaluating word similarity. However, the semantic similarity metrics presented in this paper are used for concepts, rather than words. We convert those concept-to-concept semantic similarity metrics into a word-to-word similarity metrics by taking the maximal similarity score over all the concepts which are the senses of the words [15], [41]. This is based on the intuition that human subjects would pay more attention to word similarities (i.e., most related senses) rather than their differences while rating two non-disambiguated words [41], which has been demonstrated in psychological studies [42]. Polysemic words can be mapped to a set of concepts. Let $s(w)$ denote a set of concepts that are senses of word w, then the word similarity measure is defined as:

$$sim_{word}(w_i, w_j) = \max_{c_i \in s(w_i), c_j \in s(w_j)} sim_{concept}(c_i, c_j) \quad (10)$$

where $sim_{concept}$ can be any semantic similarity methods for concepts presented in this paper. This function is used to compute word similarity scores for each semantic similarity method to be evaluated in this section.

Moreover, we implemented all the knowledge-based semantic similarity methods and graph-based IC using WordNet version 3.0[7] and DBpedia 2014[8]. The semantic similarity methods li , jcn and the proposed wpath method are implemented based on WordNet NLTK interface[9]. We use the default implementation of other similarity methods in NLTK which are based on the perl module of WordNet::Similarity [43]. We also use the NLTK's implementation

7. https://wordnet.princeton.edu/
8. http://dbpedia.org
9. http://www.nltk.org/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

of corpus-based IC using Brown Corpus [35]. For graph-based IC, we extracted 68423 WordNet concepts that have been used in YAGO [3] and used as DBpedia classes such as yago:Movie106613686. By computing the IC of those 68423 YAGO classes from DBpedia using the proposed graph-based IC, we can have the graph-based IC of those concepts in WordNet so that the graph-based IC can be evaluated using word similarity datasets. This graph-based IC computation is achieved by implementing a interface to compute IC using SPARQL queries which are executed in online SPARQL endpoint[10], including 4298433 entities. As a result, we developed a complete integrated framework to implement and evaluate semantic similarity methods for concepts in WordNet and DBpedia. All the implementations and resources, as well as the evaluation results, are published in Sematch[11] framework publicly.

4.1.2 Metrics and Evaluation

We follow the most established methodology for evaluating semantic similarity measures, which consists of measuring the Spearman correlation between similarity scores generated by the similarity methods and scores assessed by human. Note that both Spearman's and Pearson's correlations coefficients have been commonly used in the literatures. They are equivalent if rating scores are ordered and we use Spearman correlation coefficients in this paper for convenience. The conventional knowledge-based semantic similarity methods path [17] (Eq.(1)), lch [18] (Eq.(2)), wup [19] (Eq.(3)), li [20] (Eq.(4)), res [15] (Eq.(5)), lin [22] (Eq.(6)), jcn [21] (Eq.(8)) are used as compared methods and treated as baselines. A similarity measure is acknowledged to have better performance if it has higher correlation score (the closer to 1.0 the better) with human judgements, while it is acknowledged to be unrelated to human assessment if the correlation is 0. Since the Spearman's rank correlation coefficients produced by different semantic similarity methods are dependent on the human ratings for each dataset, we need to conduct statistical significance tests on two dependent (overlapping) correlations. We followed the Steigers Z test [44] used by Philipp et al. [45] to calculate statistical significance test between the dependent correlation coefficients produced by different semantic similarity methods, using a one-tailed hypothesis test for assessing the difference between two paired correlations. The cocor package of R [12] is used to calculate the statistical significance tests on dependent Spearman rank correlation coefficients. The statistical significance tests would determine whether the improvement in the correlation coefficient for each dataset is statistically significant.

In addition, the performance of IC is evaluated based on its performance of being used in semantic similarity methods. The IC computation method is acknowledged to be better if the semantic similarity method achieved better performance in using the IC. We compare the proposed graph-based IC to conventional corpus-based IC. The evaluation goal of graph-based IC is not to show that it outperforms the corpus-based IC, but rather to evaluate how graph-based IC can be exploited in IC-based semantic similarity metrics aiming to complement or substitute existing corpus-based IC methods in modern KGs.

10. http://dbpedia.org/sparql
11. https://github.com/gsi-upm/sematch/
12. https://cran.r-project.org/web/packages/cocor/index.html

TABLE 3
Numbers of word pairs for Evaluation Tasks. The headline denotes the numbers of word pairs in original dataset

Task	R&G (65)	M&C (30)	WS353(353)	WS353-Sim(203)	SimLex(999)
Word-Noun	65	30	348	201	666
Word-Graph	57	27	321	189	657
Word-Type	41	18	211	128	408

In order to evaluate the wpath semantic similarity method and graph-based IC, word similarity datasets have been processed and split into Word-Noun, Word-Graph and Word-Type. For evaluating the wpath similarity metric, the Word-Noun task was created by mapping words in word similarity datasets to WordNet noun concepts. The performance of graph-based IC is compared to corpus-based IC based on their performance in the similarity metrics of wpath, res, lin and jcn. To compute graph-based IC, words in word similarity datasets need to be mapped to DBpedia concepts. However, many words are not used as concepts in DBpedia such as noon, madhouse or lad to name a few. In consequence, in order to compare wpath and res, the Word-Graph was created by mapping the LCS of word pairs to DBpedia concepts, while the Word-Type was created by mapping the word pairs to DBpedia concepts for comparing wpath, lin and jcn. The more detailed dataset split criteria are described as below:

- **Word-Noun**: Word pairs are chosen from all the original word similarity datasets if both words can be mapped to WordNet concepts, while unmapped word pairs are removed from the datasets. We run all the semantic similarity methods based on WordNet and corpus-based IC. This task evaluates the performance of semantic similarity methods.
- **Word-Graph**: Word pairs are further chosen from datasets if both words can be mapped to WordNet concepts and the LCS of mapped concepts is one of the extracted 68423 WordNet concepts which are used as DBpedia type. Apart from running all the semantic similarity methods based on corpus-based IC, we also use the graph-based IC computed from DBpedia in the res method and the proposed wpath method. This task is able to evaluate the performance of the graph-based IC used in semantic similarity methods of res and wpath. This task is chosen because both res and wpath only rely on the IC of LCS.
- **Word-Type**: Word pairs are chosen if both words can be mapped to the extracted 68423 WordNet concepts used as entity type in DBpedia. We treat those mapped word pairs as DBpedia types. Then, all the semantic similarity methods are run using both corpus-based IC and graph-based IC. This task is able to evaluate the performance of graph-based IC used in semantic similarity of lin, jcn and wpath.

Table 11 shows the numbers of word pairs that are chosen from the original datasets in each task. In Word-Noun task, we generated word similarity scores of baselines and the proposed wpath (Eq.(9)) method using corpus-based

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

TABLE 4
Spearman correlations with ground truth in Word-Noun Task for proposed wpath method with different settings of k.

wpath k	R&G(65)	M&C(30)	WS353(348)	WS353-Sim(201)	SimLex(666)
k = 0.1	0.747	0.703	0.279	0.538	0.486
k = 0.2	0.746	0.696	0.326	0.621	0.497
k = 0.3	0.776	0.737	0.345	0.640	0.550
k = 0.4	0.785	**0.740**	**0.349**	0.647	0.573
k = 0.5	0.790	0.738	**0.349**	0.649	0.482
k = 0.6	0.789	0.732	0.348	0.648	0.589
k = 0.7	0.791	0.723	0.348	0.650	0.596
k = 0.8	0.794	0.728	0.344	**0.652**	**0.603**
k = 0.9	**0.795**	0.726	0.335	0.644	0.601
k = 1.0	0.781	0.724	0.314	0.618	0.584

TABLE 5
Word-Noun Task: Spearman correlations with ground truth of different semantic similarity methods.

Method	R&G(65)	M&C(30)	WS353(348)	WS353-Sim(201)	SimLex(666)
path	0.781	0.724	0.314	0.618	0.584
lch	0.781	0.724	0.314	0.618	0.584
wup	0.755	0.729	0.348	0.633	0.542
li	0.787	0.719	0.337	0.636	0.586
res(corpus)	0.776	0.733	0.347	0.637	0.535
lin(corpus)	0.784	0.752	0.310	0.609	0.582
jcn(corpus)	0.775	**0.820**	0.292	0.592	0.579
wpath(corpus)	**0.795**	0.740	**0.349**	**0.652**	**0.603**

TABLE 7
Word-Graph Task: Spearman correlations with ground truth of different semantic similarity methods.

Method	R&G(57)	M&C(27)	WS353(321)	WS353-Sim(189)	SimLex(657)
path	0.782	0.699	0.336	0.611	0.581
lch	0.782	0.699	0.336	0.611	0.581
wup	0.738	0.711	0.367	0.622	0.537
li	0.779	0.696	0.353	0.625	0.583
lin(corpus)	0.776	0.736	0.324	0.596	0.578
jcn(corpus)	0.762	**0.794**	0.308	0.589	0.576
res(corpus)	0.765	0.713	0.365	0.626	0.530
res(graph)	0.721	0.717	0.315	0.543	0.373
wpath(corpus)	**0.796**	0.714	**0.370**	**0.647**	**0.600**
wpath(graph)	0.788	0.776	0.336	0.620	0.581

IC. Furthermore, we experimented with different settings of k in range of $(0, 1]$ with interval of 0.1. The Spearman correlations between the wpath method with different k settings and human judgements are shown in Table 4. Each column denotes each dataset and each row denotes a specific k value running the wpath method. Note that the bold values in each column denotes the highest correlation score for each dataset which is also same for the following tables. The Spearman correlations between baselines and human judgements are shown in Table 5. Each row represents a semantic similarity method and the columns denote the different datasets. The row wpath shows the highest correlation score from the Table 4 for each dataset. Note that the corpus in parentheses of each method denotes that the method has used corpus-based IC. The Word-Noun is a superset of Word-Graph and Word-Type, which contains complete word pairs and human ratings. In order to evaluate whether the proposed wpath semantic similarity method outperforms other semantic similarity methods, a statistical significance test based on Steiger's Z test [44] has been carried out to analyse the results of Word-Noun task, using one tailed test and 0.05 statistical significance in each dataset. Table 6 shows the result of Steiger's Z significance test on the differences between Spearman correlations (ρ) of wpath method and other semantic similarity methods.

In Word-Graph task, we computed the word similarity scores of baselines and the wpath method with the task setting of Word-Graph. Particularly, for the res and wpath we also used graph-based IC, while the lin and jcn only used corpus-based IC. The Spearman correlations between similarity methods and human judgements for Word-Graph task are shown in Table 7. The graph in parentheses of methods denote that the method has used the graph-based IC. In Word-Type task, we computed the word similarity scores of baselines and the wpath method with the task setting of Word-Type. Apart from corpus-based IC, we also used graph-based IC for the methods of res, lin, jcn, and wpath. The difference between the methods res, wpath and methods of lin, jcn is that the previous two only use the IC of LCS while the latter two also use the IC of individual concepts. The Spearman correlations between similarity methods and human judgements for the Word-Type task are shown in Table 8. We also experimented with different k settings of wpath method in the Word-Graph task and Word-Type task for both corpus-based IC and graph-based IC. In Table 7 and Table 8 we reported the best results of wpath method for each task and each dataset, while Table 9 shows the specific settings of wpath method achieved best result in each task and each dataset. Within the evaluation of three tasks and corresponding results, we then analyse the results in the following section.

TABLE 8
Word-Type Task: Spearman correlations with ground truth of different semantic similarity methods.

Method	R&G(41)	M&C(18)	WS353(211)	WS353-Sim(128)	SimLex(408)
path	0.679	0.621	0.353	0.601	0.616
lch	0.679	0.621	0.353	0.601	0.616
wup	0.613	0.606	0.357	0.589	0.538
li	0.673	0.614	0.361	**0.612**	0.612
res(corpus)	0.667	0.679	0.355	0.595	0.540
res(graph)	0.674	0.704	0.294	0.487	0.381
lin(corpus)	0.642	0.696	0.322	0.539	0.592
lin(graph)	0.624	0.661	0.305	0.517	0.534
jcn(corpus)	0.676	**0.805**	0.342	0.546	0.594
jcn(graph)	0.309	0.324	0.241	0.440	0.331
wpath(corpus)	0.691	0.669	**0.367**	0.606	**0.625**
wpath(graph)	**0.717**	0.765	0.353	0.601	0.616

TABLE 9
Spearman correlations with ground truth in Word-Noun Task for proposed wpath method with different settings of k.

Setting	R&G	M&C	WS353	WS353-Sim	SimLex
Word-Graph IC-Corpus	k=0.9	k=0.5	k=0.7	k=0.8	k=0.8
Word-Graph IC-Graph	k=0.9	k=0.5	k=0.8	k=0.9	k=1.0
Word-Type IC-Corpus	k=0.8	k=0.5	k=0.7	k=0.9	k=0.9
Word-Type IC-Graph	k=0.6	k=0.6	k=1.0	k=1.0	k=1.0

4.1.3 Result Analysis and Discussion

Our main hypothesis in the experiments is that the proposed semantic similarity method wpath will improve over the baselines and show high correlation to human assessments. The second hypothesis is that the proposed graph-based IC computation method is effective compared to the conventional corpus-based IC, which means the graph-based IC needs to show close performance or outperforming in some cases.

Table 5, 7 and 8 show that all the semantic similarity methods have high correlation with human judgements and the proposed wpath semantic similarity method outperforms the baselines in most of cases except the M&C dataset and WS353-Sim dataset in WordType task. Moreover, from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

TABLE 6
Steiger's Z significance test on the differences between Spearman correlations (ρ) of wpath method and other semantic similarity methods using 1-tailed test and 0.05 statistical significance .

	R&G(65)		M&C(30)		WS353(348)		WS353-Sim(201)		SimLex(666)	
Method	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value
path	.982	.171	.984	.248	.967	.003	.960	.013	.955	.021
lch	.982	.171	.984	.248	.967	.003	.960	.013	.955	.021
wup	.964	.029	.946	.398	.969	.468	.959	.110	.946	.000
li	.982	.293	.978	.223	.978	.129	.974	.097	.965	.019
res	.956	.204	.943	.436	.952	.449	.948	.194	.913	.000
lin	.956	.314	.969	.353	.903	.040	.900	.038	.944	.021
jcn	.876	.296	.890	.067	.831	.026	.845	.023	.916	.029

the three tables we observed that the **jcn** method performed exceptionally best in the M&C dataset, however in other datasets it performed not as good as the one in M&C dataset. It is probably because of the small word pair sample in M&C dataset. It was also surprising that the **li** method had performed best only in WS353-Sim in Word-Type task. It may be caused by the specific subset of the dataset.

In Table 6, on R&G dataset, the significance test shows the improvement of wpath over wup (the p-value of each test is below the significance level of 0.05), while indicates no statistical significance differences with other methods. Regarding the M&C dataset, although the jcn method performs best, the result of statistical significance test indicates that no statistic significant differences between wpath and jcn (p-value > 0.05). On WS253-Sim and WS353 datasets, it is clear that the wpath has statistical significant improvement over the path, lch, jcn and lin. Finally, on SimLex dataset, the wpath has statistical significant improvement over all other semantic similarity metrics. In general, from the results of our experiments, we observed that different semantic similarity metrics have performed differently in different datasets. The wpath similarity metric has obtained the best performance in 4 out of 5 datasets (ranked as second in M&C only containing 30 word pairs). This shows that the wpath similarity metric has provided a stable performance in all datasets. Considering that SimLex is the largest dataset for semantic similarity, bigger than the combination of all other datasets, we may conclude that the wpath method has produced statistically significant improvement over other semantic similarity metrics.

From Eq.(9) we know when $k = 1$ the proposed wpath method is equivalent to path method. As the value of k becoming smaller, IC starts to have bigger influence. Even with low or high values of k, k contributes to solve the uniform distance problem of the path method illustrated in Table. 2. It has been shown in Table 4 and Table 9 that the best k has smaller value in R&G and M&C datasets. As pure IC-based semantic similarity methods also achieved better performance in those two datasets, probably the human ratings of word pairs in those datasets care more on IC or general relevance. Based on this observation, the parameter k actually defines for a given KG the balance among hierarchical structure and statistical information for calculating semantic similarity. Its values can provide insight about which metrics perform better in a given KG. For high values of k, structural metrics will provide a better result and for low values of k, IC metrics perform better.

Different KGs have different concept taxonomies and different distributions of instances over concepts. Even in a given concept taxonomy, concepts are not equally structured, such as various density of sub-concepts and different hierarchical levels of concepts. This can be shown in Fig. 2. Given that applications usually use a group of concepts from a taxonomy (e.g. restaurant domain), the specific value of k should be selected for a specific domain (e.g. a subgraph of a KG) that reflects the concept structure and IC of that domain. In consequence, the selection of k would be the optimization of k for a specific group of concepts. For those concepts having human ratings, k can be adjusted empirically or learned automatically by comparing to human ratings. For those concepts without human ratings, k should be determined according to the specific domain application, in which k can be selected empirically or learned automatically based on application performances.

Regarding to the graph-based IC, we observed that it performed better in Word-Type task than Word-Graph task. It is also shown in Table 7 and 8 that the graph-based IC has better performance in res, lin and wpath methods than jcn. It is shown in Table 8 that graph-based IC may not be suitable for the **jcn** method, and the graph-based IC achieved the best performance in R&G and M&C dataset in Word-Type task while had a similar result in other datasets compared to corpus-based IC. Consequently, we may conclude that the graph-based IC computation method is effective compared to conventional corpus-based IC in measuring word similarity but not always outperforming. Moreover, graph-based IC has a number of benefits, since it does not requires a corpus and enables online computing based on available KGs. Besides, graph-based IC metrics can benefit from the success of open linked data, and the continuous growth of available KGs.

4.2 Aspect Category Classification Evaluation

Although the word similarity correlation measure is the standard way to evaluate the semantic similarity methods, it relies on human evaluation over word pairs which may not have the same performance in real applications [46]. Therefore, we consider to evaluate the semantic similarity methods in a real application. In order to evaluate the semantic similarity methods considering structure and IC features, without involving too much other features, we choose a simple aspect category classification task in Aspect Based Sentiment Analysis (ABSA) [23], [24].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

TABLE 10
Most Frequent Words Co-occur With Each Aspect Category

Aspect Category	Frequent Feature Words
SERVICE	service, staff, waiter, waitress, wait, manager, delievery
RESTAURANT	place, restaurant, spot, pizza, femme, casa, season
FOOD	food, pizza, sushi, dish, menu, fish, chicken, meal, salad
DRINKS	wine, drink, beer, selection, bottle, martini, glass, margarita
AMBIENCE	atmosphere, place, decor, ambience, music, room, garden
LOCATION	view, location, neighborhood, city, place, outdoor, avenue

TABLE 11
Numbers of Sentences in Evaluation for Each Aspect Category

Categories	SERVICE	RESTAURANT	FOOD	DRINKS	AMBIENCE	LOCATION
Numbers	519	228	2256	54	597	752

TABLE 12
Accuracy, Precision, Recall and F-Measure of Aspect Category Classification using different semantic similarity methods.

Method	Accuracy	Precision	Recall	F-measure
path	.793	.658	.736	.680
lch	.788	.656	.704	.662
wup	.769	.630	.685	.637
li	.783	.659	.701	.667
res	.723	.560	.679	.558
lin	.731	.575	.674	.567
jcn	.732	.606	.702	.609
wpath	**.800**	**.664**	**.741**	**.689**

4.2.1 Aspect Category Classification

ABSA is an evaluation task of the SemEval workshop that provides benchmark datasets of reviews and a common evaluation framework. In SemEval 2015 and 2016, the task sentence-level ABSA has defined a subtask so-called Aspect Category Detection, whose aim is to identify every entity E and attribute A pair, towards which an opinion is expressed in the given text [23]. Specifically, given an input sentence such as "The food was delicious", ABSA needs to detect the E and A pair (category=FOOD#QUALITY) for the target word "food" and to estimate its sentiment either positive or negative. The English dataset has been provided for two domains: Laptops and Restaurants. We have chosen the latter for this evaluation. In the restaurant domain, SemEval predefines a set of entities SERVICE, RESTAURANT, FOOD, DRINKS, AMBIANCE and LOCATION, which can be viewed as general aspect categories. Our task of aspect category classification consists in assigning a general aspect category to opinion target words. For example, words such as wine, beverage and soda are classified into DRINKS, while words such as bread, fish, and cheese are classified into FOOD. Note that only entity E (FOOD) is used as general aspect category and the attribute QUALITY is not considered for simplicity.

This task challenges semantic relatedness methods, especially for corpus-based methods. For instance, in restaurant review corpora, those target words such as fish and wine would appear in same surrounding contexts (e.g. "the fish is delicious and the wine is great"). Since corpus-based methods are based on calculating co-occurrences of terms in a corpus, they can hardly discriminate terms from different categories that are frequently collocated (e.g. fish and wine). In such scenario, knowledge-based methods are useful to include the structural knowledge from domain taxonomy. As illustrated in a fragment of WordNet in Fig. 2, lamb, beef, and seafood are sub-concepts of FOOD category, while coffee, tea and milk are sub-concepts of DRINKS category. Intuitively, semantic similarity methods can be used to measure the taxonomical similarity between target words and aspect category in order to classify the target words into correct aspect category.

The most frequent target words of a category are used as features for representing that category. Features of different aspect categories are illustrated in Table 10. Formally, we use $A = \{a_1, \ldots, a_n\}$ to denote a set of aspect categories, and $f(a_i)$ to denote the feature words of a category a_i. For a feature word $w_k \in f(a_i)$, we use $weight(w_k) = \frac{count(w_k)}{N_{a_i}}$ to denote the weight of the feature word w_k. The $N_{a_i} = \sum_{w_k \in f(a_i)} count(w_k)$ denotes the total count of feature words of category a_i. The counts of feature words are derived from the annotated datasets from SemEval ABSA [23], [24]. We can define a simple aspect category classification framework based on the word semantic similarity method defined in Eq.(10), in which different semantic similarity methods are used. Given a sequence of new target words $T = \{w_1, \ldots, w_k\}$, we chose the aspect category $\hat{a}$ that maximizes the following similarity function as the correct category of the T.

$$\hat{a} = \underset{a_i \in A}{argmax} \max_{w_j \in T} \sum_{w_k \in f(a_i)} sim_{word}(w_j, w_k) * weight(w_k) \tag{11}$$

Given an aspect category a_i, the formula sums the semantic similarity scores between the target words and the feature words. The highest similarity score of the target word is chosen to represent the similarity score between T and a_i. The aspect category with the highest similarity score would be chosen as the correct aspect category. By using different semantic similarity methods in this simple framework, we are able to show the effectiveness of similarity methods in a simple real world application without involving too much additional features. We then report the evaluation results in the following Section.

4.2.2 Evaluation Results

We use the restaurant review datasets of ABSA in SemEval-2015 and SemEval-2016 [23], [24]. Both datasets contain annotated target words and corresponding category. We have converted the specific categories into general categories, and collected a list of target words and category pairs. As a result, we got a dataset containing 4406 tuples in form of target words and category pairs such as (shellfish, FOOD). The numbers of pairs belong to each category are shown in Table 11. Since the dataset contains 6 classes, we use multi-class classification metrics accuracy, macro-average of precision, recall and f-measure as the performance metrics to evaluate the semantic similarity methods.

We have implemented the semantic similarity based aspect category classification system and evaluated the classification system in the dataset with different semantic similarity methods. The evaluation results are reported in Table 12. We have experimented with different k values and the best $k = 0.9$ is chosen for the proposed wpath method.

This k can be treated as the optimized setting for wpath method in calculating semantic similarity between concepts in the restaurant domain. As we mentioned previously, the k value can provide insight about which metrics perform better in a given group of concepts. Since k has shown higher value in this restaurant domain, the structure information of concept taxonomy is relatively more important. It is also shown in Table 12 that the structure based semantic similarity methods, path, lch, wup, and li are performing better than IC based methods res, lin and jcn. Moreover, the proposed wpath method has achieved the best accuracy, precision, recall and F-measure score among other semantic similarity methods. We could conclude that combining the statistical IC with the structure information can improve the performance of structure based semantic similarity methods in the task of aspect category classification where the hierarchical structure is considered to be important.

5 Conclusions and Future Work

Measuring semantic similarity of concepts is a crucial component in many applications which has been presented in the introduction. In this paper, we propose wpath semantic similarity method combining path length with IC. The basic idea is to use the path length between concepts to represent their difference, while to use IC to consider the commonality between concepts. The experimental results show that the wpath method has produced statistically significant improvement over other semantic similarity methods. Furthermore, graph-based IC is proposed to compute IC based on the distributions of concepts over instances. It has been shown in experimental results that the graph-based IC is effective for the res, lin and wpath methods and has similar performance as the conventional corpus-based IC. Moreover, graph-based IC has a number of benefits, since it does not requires a corpus and enables online computing based on available KGs. Based on the evaluation of a simple aspect category classification task, the proposed wpath method has also shown the best performance in terms of accuracy and F score.

In this paper, we evaluated the proposed method in the word similarity dataset and simple classification using the most established evaluation method. More evaluation of semantic similarity methods in other applications considering the taxonomical relation could be useful and can be one of our future works. Furthermore, this paper mainly discussed semantic similarity rather than general semantic relatedness. Therefore, another future work could be in studying the combination of knowledge-based methods with the corpus-based methods for semantic relatedness. Finally, since we combined WordNet and DBpedia together in this paper, we would further explore using the proposed approaches for measuring the entity similarity and relatedness in KGs.

Acknowledgments

This work is supported by the Spanish Ministry of Economy and Competitiveness under the R&D projects SEMOLA (TEC2015-68284-R) and EmoSpaces (RTC-2016-5053-7), by the Regional Government of Madrid through the project MOSI-AGIL-CM (grant P2013/ICE-3019, co-funded by EU Structural Funds FSE and FEDER), and by the European Union through the project MixedEmotions (Grant Agreement no: 141111).

References

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 1247–1250.

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "Dbpedia-a crystallization point for the web of data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 154 – 165, 2009, the Web of Data.

[3] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "Yago2: A spatially and temporally enhanced knowledge base from wikipedia (extended abstract)," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI '13. AAAI Press, 2013, pp. 3161–3165.

[4] I. Horrocks, "Ontologies and the semantic web," *Commun. ACM*, vol. 51, no. 12, pp. 58–67, Dec. 2008. [Online]. Available: http://doi.acm.org/10.1145/1409360.1409377

[5] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[6] R. Navigli and S. P. Ponzetto, "Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[7] E. Hovy, R. Navigli, and S. P. Ponzetto, "Collaboratively built semi-structured content and artificial intelligence: The story so far," *Artificial Intelligence*, vol. 194, pp. 2 – 27, 2013, artificial Intelligence, Wikipedia and Semi-Structured Resources.

[8] R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 2, p. 10, 2009.

[9] A. Moro, A. Raganato, and R. Navigli, "Entity linking meets word sense disambiguation: a unified approach," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 231–244, 2014.

[10] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum, "Kore: Keyphrase overlap relatedness for entity disambiguation," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, ser. CIKM '12. New York, NY, USA: ACM, 2012, pp. 545–554.

[11] I. Hulpus, N. Prangnawarat, and C. Hayes, "Path-based semantic relatedness on linked data and its use to word and entity disambiguation," in *International Semantic Web Conference*, 2015.

[12] J. Pound, I. F. Ilyas, and G. Weddell, "Expressive and flexible access to web-extracted data: A keyword-based structured query language," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '10. New York, NY, USA: ACM, 2010, pp. 423–434.

[13] M. Schuhmacher and S. P. Ponzetto, "Knowledge-based graph document modeling," in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, ser. WSDM '14. New York, NY, USA: ACM, 2014, pp. 543–552.

[14] S. Shekarpour, E. Marx, A.-C. N. Ngomo, and S. Auer, "Sina: Semantic interpretation of user queries for question answering on interlinked data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 30, pp. 39 – 51, 2015, semantic Search.

[15] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 448–453.

[16] P. D. Turney, P. Pantel *et al.*, "From frequency to meaning: Vector space models of semantics," *Journal of artificial intelligence research*, vol. 37, no. 1, pp. 141–188, 2010.

[17] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 1, pp. 17–30, 1989.

[18] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," *WordNet: An electronic lexical database*, vol. 49, no. 2, pp. 265–283, 1998.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2610428, IEEE Transactions on Knowledge and Data Engineering

[19] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ser. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994, pp. 133–138.

[20] Y. Li, Z. Bandar, and D. Mclean, "An approach for measuring semantic similarity between words using multiple information sources," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, no. 4, pp. 871–882, 2003.

[21] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *Computational Linguistics*, vol. cmp-lg/970, no. Rocling X, p. 15, 1997.

[22] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning*, ser. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296–304.

[23] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "Semeval-2015 task 12: Aspect based sentiment analysis," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado*, 2015, pp. 486–495.

[24] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. D. Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jimnez-Zafra, and G. Eryiit, "SemEval-2016 task 5: Aspect based sentiment analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation*, ser. SemEval '16. San Diego, California: Association for Computational Linguistics, June 2016.

[25] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI*, vol. 6, 2006, pp. 775–780.

[26] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Comput. Linguist.*, vol. 16, no. 1, pp. 22–29, Mar. 1990.

[27] 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*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 767–776.

[28] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[29] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis." in *IJCAI*, vol. 7, 2007, pp. 1606–1611.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[31] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, vol. 12, 2014, pp. 1532–1543.

[32] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of Artificial Intelligence Research*, vol. 11, no. 95, pp. 95–130, 1999.

[33] L. Han, T. Finin, and A. Joshi, "Schema-free structured querying of dbpedia data," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, ser. CIKM '12. New York, NY, USA: ACM, 2012, pp. 2090–2093.

[34] K. Church and W. Gale, "Inverse document frequency (idf): A measure of deviations from poisson," in *Natural Language Processing Using Very Large Corpora*, ser. Text, Speech and Language Technology, S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, Eds. Springer Netherlands, 1999, vol. 11, pp. 283–295.

[35] W. N. Francis and H. Kucera, "Brown corpus manual," *Brown University*, 1979.

[36] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Commun. ACM*, vol. 8, no. 10, pp. 627–633, Oct. 1965.

[37] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and cognitive processes*, vol. 6, no. 1, pp. 1–28, 1991.

[38] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," *ACM Trans. Inf. Syst.*, vol. 20, no. 1, pp. 116–131, Jan. 2002.

[39] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, "A study on similarity and relatedness using distributional and wordnet-based approaches," in *NAACL*, Stroudsburg, PA, USA, 2009, pp. 19–27.

[40] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *arXiv preprint arXiv:1408.3456*, 2014.

[41] D. Sánchez, M. Batet, D. Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach," *Expert Systems with Applications*, vol. 39, no. 9, pp. 7718–7728, 2012.

[42] A. Tversky, "Feaures of similarity," *Psychological Review*, vol. 84, pp. 327–352, 1977.

[43] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet::similarity: Measuring the relatedness of concepts," in *Demonstration Papers at HLT-NAACL 2004*, ser. HLT-NAACL–Demonstrations '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004, pp. 38–41. [Online]. Available: http://dl.acm.org/citation.cfm?id=1614025.1614037

[44] J. H. Steiger, "Tests for comparing elements of a correlation matrix." *Psychological bulletin*, vol. 87, no. 2, p. 245, 1980.

[45] P. Singer, T. Niebler, M. Strohmaier, and A. Hotho, "Computing semantic relatedness from human navigational paths: A case study on wikipedia," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 9, no. 4, pp. 41–70, 2013.

[46] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

Ganggao Zhu is currently a PhD candidate in the lab of intelligent systems at the Universidad Politécnica de Madrid, Spain. He graduated in computer science from Northwestern Polytechnical University, P.R.China in 2010 and obtained master degree in software and system from the Universidad Politécnica de Madrid in 2012. His current research interests are semantic analysis, semantic search and question answering.

Carlos A. Iglesias is an associate professor at the Universidad Politécnica de Madrid, Spain. His research interests are in multiagent systems, service engineering, and Web engineering. Iglesias has a PhD in telecommunications engineering from the Universidad Politécnica de Madrid.