

A Graph-based Approach to Measuring Semantic Relatedness in Ontologies

Ahmad Hawalah
School of Computer Science and
Electronic Engineering
University of Essex
Colchester, UK
ayhawa@essex.ac.uk

Maria Fasli
School of Computer Science and
Electronic Engineering
University of Essex
Colchester, UK
mfasli@essex.ac.uk

ABSTRACT

The need of determining the degree of semantic similarity, relatedness or distance between two concepts within the same ontology or two different ontologies is becoming an increasingly important task in the field of Information Retrieval. Although a great attention has been paid to design semantic similarity/relatedness methods based on taxonomies, there has been little discussion about the design of semantic similarity/relatedness methods based on ontologies. In this paper we introduce a novel graph-based semantic relatedness approach to calculate semantic relatedness considering both hierarchical and non-hierarchical concepts in an ontology. In addition, our approach considers some important properties such as different relation types, concepts' depth and distance that play an essential role in measuring semantic relatedness. Three experimental studies are provided to first illustrate that our approach give a different results than other measures in the literature. Then, we compare our approach against existed methods using a benchmark dataset, and finally, evaluate our approach by using a real ontology and compare the predictions of our semantic relatedness approach against the human-subject judgment. The results in all the experiments show a considerable improvement against traditional taxonomy-based measures.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Information filtering, Relevance feedback; H.1.2 [User/Machine Systems]: Human information processing; I.1.11 [Distributed Artificial Intelligence]: Multiagent systems

General Terms

Measurement, Experimentation.

Keywords

Semantic similarity; semantic relatedness; ontology; web mining

1. INTRODUCTION

The rapid expansion of heterogeneous and independent data networks and repositories have caused information overload to

such an extent that the process of finding a specific piece of information or a suitable product may often become frustrating and time-consuming for users. Therefore, determining the degree of semantic similarity, relatedness or distance between two concepts in a structured network is becoming increasingly important in the field of information retrieval and ontology-engineering to enhance user search and browsing experience. Furthermore, semantic similarity measures have been used in different applications such as query expansion [1, 2] where user queries are modified to retrieve more relevant information, community mining [3, 4] and natural language processing [5, 6].

However, much of the focus has been on designing a semantic similarity or distance measures based on taxonomies where information about a concept inclusion hierarchy is the only considered information [5, 7, 8].

However, the main focus in this paper is to develop a general semantic relatedness method to measure the relatedness between concepts in an arbitrary ontology where different relation types exist. Furthermore, we address three questions in this paper which are: i) how to measure semantic relatedness between concepts in an ontology where various relation types exist, ii) how to exploit the common understanding of the information expressed in an ontology, iii) and finally what properties should be considered when computing the semantic relatedness that could impact of the overall relatedness performance.

In this paper, we attempt to answers these questions by first introducing important properties which play an essential role in measuring semantic relatedness, and then introducing a novel graph-based semantic relatedness approach to calculate semantic relatedness considering both hierarchical and non-hierarchical concepts in an ontology.

This paper is structured as follows. The following section discusses related work. Next we discuss the important properties that are going to be considered in our semantic relatedness approach. Section four presents the details of our proposed graph-based semantic relatedness for ontological concepts and its stages. The next section presents a set of experiments that have been conducted and discusses the found results. The paper ends with the conclusions and pointers to future work.

2. PREVIOUS WORK

In the last few years, the term of ontology has started to be used frequently in connection with modelling and representing user profiles. According to Razmerita and Lytras [9] "An ontology enables the conceptualization of the domain knowledge for an application". In other words, ontologies provide a highly

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WIMS'11, May 25-27, 2011 Sogndal, Norway.
Copyright © 2011 ACM 978-1-4503-0148-0/11/05... \$10.00"

expressive ground for describing keywords or concepts and a rich variety of interrelations among them. Unlike simple methods of representing information such as weighted keywords and semantic network profiles, an ontology provides a more powerful, deeper and broader concept hierarchy representation for user profiles. Some of the most important advantages of developing and using ontologies are summarized as follows: (i) sharing of common understanding of the information; (ii) the ability to reuse a domain ontology; (iii) support for context reasoning; (iv) powerful information and context management and inference mechanisms; (v) reduction of ambiguity and solving inconsistency of the information.

Many semantic measures have been presented in the literature in the past few years, which can be divided into three main models. The first model is the edge-based model which is based on finding the shortest path between two concepts. The second model is the information content model that is based on shared information contents of the most common super-concept of two concepts. The last model is the information-theoretic model that is based on estimating the similarity between concepts in an ontology. All of the measures in the first and second models are based on taxonomies and enclosed to just taxonomic relations such as is-a relation. The third model however, consists of more sophisticated measures that measure semantic similarities for concepts in an ontology. All of these three models are discussed as follows:

2.1 Shortest Path Length (Edge-based) Model

Early research focused on using taxonomy which is a hierarchal representation of information to improve the information retrieval. Such a taxonomy consists of concepts and relations between these concepts. The main assumption in this group is that the shorter distance between two concepts, the more similar the concepts are. Many research proposals have adopted this hypothesis to compute similarity or distance between two concepts in a network representation. One of the first and simplest measures in this group was proposed by Rada [8]. In this paper, Rada et al. measure the semantic distance between two concepts based on the number of edges between them. However, this approach is limited to just taxonomic relations, and it assumes that all the taxonomy relations have the same weight.

Many other approaches, however, suggested considering more characteristics when calculating the weight between two concepts including the concept depth, the network density, the type and the strength of the relation that connects two concepts [10]. Sussna [11] for example, measures semantic similarity taking into the account the length between two concepts and the network density. According to [12], Sussna's approach was the first approach that consider the depth of the network representation built upon the hypothesis that the deep concepts in the tree are more related than those in higher levels. Although this approach has been reported to be robust, the restriction on the relation types impact negatively on the overall performance [13].

Another work based on concept-depth approach was proposed by Wu and Palmer [14]. They suggested measuring semantic similarity between two concepts in a taxonomy taking into account the least common super-concept between those two concepts.

One major criticism of all the above studies is the underlying assumption that all the relations in taxonomy have the same value.

This problem has been discussed by many researchers including [13, 15, 16]. These studies suggested weighting the relations and the edges between concepts in a taxonomy to overcome the earlier problem.

2.2 Information content model

Another hypothesis has been introduced in the literature based on the observation that the more the information is shared between two concepts, the more similar they are. Resnik [5] proposed a new semantic measure based on this hypothesis as an alternative to the edge – based model. In his approach, Resnik argued that similarity between two concepts can be expressed based on the information shared between these two concepts. The information shared between these two concepts can be estimated by the information content for the highly specific super-concept that subsumes them both. Jiang and Conrath [13] also proposed another approach that is also based on the information content. Unlike Resnik's approach that uses just the information content for the most specific common super-concept, Jiang and Conrath considered also the information content for each concept. They compute the distance between two concepts by first summing up all the information content of both concepts. Then the final similarity is computed as the deference between the total sum and the information content of the most specific common super-concept. Another similar approach to Jiang and Conrath [13] is proposed by Lin [17]. In this study, Lin introduced a semantic similarity measure for a taxonomy and corpus. In this approach Lin considered the information content for the most common super-concept as in [5], and the information content for both concepts as in [13]. In addition, Lin uses the same components as [13] measurement but instead of the difference he uses the ratio.

All of the above measures in the both Edge-based and information content models were proposed basically to work with taxonomies and limited to just one relation (is-a) or some other taxonomic relations. However, many issues arise when attempting to apply these measures on a more complex network representation such as an ontology which has more complex and non-uniform relations. In the next section, more sophisticated approaches that estimate semantic similarity for concepts in generalized ontologies are presented.

2.3 Information-theoretic model

A few works that measures semantic similarity for ontological concepts representations were proposed in the literature. Unlike previous works that only consider taxonomic relations, these works consider different types of relations including hierarchical and non-hierarchical concepts. Maguitma and Menczer [18], proposed a graph-based similarity approach to deal with this problem. The main goal of this work is to detect semantic similarity for hierarchical and non-hierarchical concepts in an ontology where the relations between these concepts are dense and varied. In this works, [18] proposed a semantic similarity algorithm that is based on Open Directory Project (ODP) ontology [19]. This ontology which known to be the largest user-edited directory of the Web has various types of hierarchical and non-hierarchical relations. The process of measuring semantic similarity yields to four main stages. Firstly, for each relation's type, [18] assigned a pre-defined weight which represents the degree of membership between two concepts. Then the membership weights are then used to measure the transitive relations for all concepts in the ontology. A fuzzy membership

value for each concept in the ontology is then defined using the MaxProduct fuzzy composition function. After calculating the fuzzy membership matrix, the semantic similarity for two concepts in an ontology is calculated.

Maguitma and Menczer [18] reported that their algorithm has achieved a significantly greater correlation accuracy to human judgment than tree based approaches. Although this algorithm has achieved good results, it does not take into the account other important properties such as the depth of the ontology and the dense of concepts.

Song et al. [20], suggested another semantic similarity approach which is similar to the above one called fuzzy semantic similarity. In this approach, the depth of the ontology, the density, and the different relation types have been considered. Similar to [18], Song et al. [20] have utilized ontology graphs, which are represented as adjacency matrices. The process of measuring semantic similarity in this approach consists of four main stages which are: compute direct relations; compute non-hierarchical relations, transitive relations and finally computing the semantic similarity between all concepts. The reported results showed that this fuzzy similarity measure achieved better results when it was compared to other approaches such as [14] approach that is based on edge-count model.

Although both of these approaches achieved good results, many important properties have not been explored in their approaches. In the next section we attempt to explore new properties in more detail and how they may affect the semantic relatedness performance.

3. MEASURING SEMANTIC RELATEDNESS OF ONTOLOGICAL CONCEPTS

In this section we introduce a novel approach to calculate semantic relatedness considering both hierarchical (i.e. taxonomic) and non-hierarchical (i.e. non-taxonomic relations such as cross relations between two concepts) relations in an ontology. This approach is motivated by the following points:

- Take advantage of the ontology's unique structure which contains hierarchal and non-hierarchal concepts.
- Having different types of relations including hierarchal and non- hierarchal (semantic) relations.
- Considering different important aspects the impact on the relatedness.
- The ability of applying proposed approaches to large and general domains.

Motivated by the above points, we identified eight essential properties that are derived from a set of assumptions, and serve as a basis for our semantic relatedness approach. In addition, these properties are tied to the structures and relations of ontologies that have hierarchical and non hierarchical concepts (Figure 1 illustrates a fragment example of such an ontology). These properties are:

Property 1: Identity Property

The relatedness between two identical concepts where the distance between these two concepts is zero is the highest possible value for relatedness.

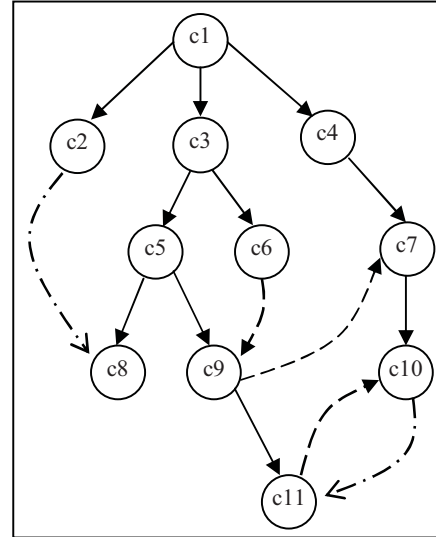


Figure 1. A fragment of an ontology with different relation types.

Property 2: Symmetrical and Non-Symmetrical Properties

Many researches in the literature argued that semantic similarity or relatedness between two concepts should be symmetrical [5, 8]. However, according to [12], this argument is not valid when using taxonomies or ontologies which are represented as directed graphs. In other words, each edge for each concept in a taxonomy or an ontology has a relationship in the direction of the edge. Therefore, in our approach the relatedness between two concepts are both Symmetrical and Non-Symmetrical. Semantic relatedness between two concepts is symmetrical when these concepts have just one relation between them, and hence the ordering is disregarded. To illustrate, if a user has two interesting concepts let say football and swimming. What is important for our approach is to measure the relatedness between these two concepts, and for this reason (football $\rightarrow$ swimming) and (swimming $\rightarrow$ football) are the same. On the other hand, in some cases semantic relatedness for two concepts could be Non-Symmetrical if there is more than one relation between these two concepts. This case occurs mostly in ontologies where different types of semantic relations are available. Therefore, each relation and direction may have different weight, and hence each direction between two concepts may have different relatedness value.

Property 3: Positive Property

The relatedness between two concepts c1 and c2 is non negative and it is between [0,1].

Property 4: Hierarchal and Non-hierarchal Relations Property

Each type of relations including hierarchal or non-hierarchal relations has different meaning. Therefore, in our approach each relation should have different weight to reflect the meaning of the relation.

Property 5: Minimal Distance Property

Minimum Distance is the shortest length of a path between two concepts in an ontology. The relatedness between two concepts c1 and c2 is higher if they are closer, while the relatedness is lower if they are far. In our approach, this property should be satisfied by

considering different types of relations which have different weights as it was mentioned above.

Property 6: Depth-Relative Property

Depth-Relative Property states that concepts deep in an ontology are more closely related than concepts higher in the ontology. Unlike all the above mentioned approaches that use the depth of the ontology (i.e. the maximum level of the root concept in an ontology) to satisfy this property; in our approach we use the maximum depth not for the root concepts, but for each concept based on its descendants. This point is explained in depth in the following section.

Property 7: Multiple Paths Property

Two concepts $c1$ and $c2$ could have more than one relation (or path) between them. In our approach, we consider just the relation with the maximum relatedness value.

Property 8: Transitivity property

Transitivity property states that if a concept x has sub-concepts, then x is to some extent related to all of its descendants. In our approach semantic relatedness is transitive to all the inclusion relations for any concept including hierarchal and just the first level of non-hierarchal concepts. In other words, for non-hierarchal concepts, we just consider hierarchal concepts but not non-hierarchal ones. To illustrate, in figure 1, concept $c6$ is related to $c9$ and $c11$ but not $c10$. $C6$ and $c9$ are related because they have a non-hierarchal relation between them. $C6$ is also related to $c11$ because $c11$ is hierarchically related to $c9$, and $c9$ is the first level of non-hierarchal concepts of $c6$. On the other hand, $c6$ is neither related to $c7$ nor $c10$, because they have non-hierarchal relations with $c9$ and $c11$. Hence, they are not related to $c6$ as we only consider the first level of non-hierarchal concepts. Without this restriction, considering all relations' types would make the classification and the organization of the ontology meaningless by making all concepts related to almost every other concept. Therefore, by limiting the transitivity to just hierarchal concepts and first level of non-hierarchal concepts allows us to preserve feasibility and meaning.

Property 9: A Concept's Maximum Depth Property

In property 6, we identified a Depth-Relative which is the depth of a concept in an ontology. In this property, however, we introduce a new property that as far as we know has not been explored before. The concept's maximum depth is basically the maximum depth of its most specific leaf concept (sub-concept). Similar to the previous Transitivity property, when calculating the maximum depth of a concept we consider just the hierarchal and just the first level of non-hierarchal concepts, and just the hierarchal concepts under the non-hierarchal concepts. We add this restriction to avoid having a loop when a concept for example is linked semantically to one of its parents. However, this property is very important when we compute the semantic relatedness between concepts as it reflects the density of such a concepts. Furthermore, instead of considering just the global depth of the whole ontology, we consider each concept's maximum depth and in this way we can explore more context information about each concept.

3.1 Graph-based Semantic Relatedness for Ontological Concepts

In this section we introduce a graph-based semantic relatedness measure based on ontologies (GBSRO). In this approach an ontology O is represented as a graph of concepts c , and relations r between these concepts $O=(c, r)$. Where c is both hierarchal and non-hierarchal concepts and r are both hierarchal and semantic relations. Figure 24, illustrates a fragment of an abstract ontology consisting of concepts and three different relations' types. The O ontology is defined by the sets of $n=\{c1, c2, c3, c4, c5, c6, c7, c8, c9, c10, c11\}$, the hierarchal relations (solid arrows) $H=\{(c1, c2), (c1, c3), (c1, c4), (c3, c5), (c3, c6), (c4, c7), (c5, c8), (c5, c9), (c7, c10), (c9, c11)\}$, the first type of non-hierarchal relations (dash arrows) $S1=\{(c6, c9), (c9, c7), (c11, c10)\}$ and the second type of non-hierarchal relations (dash-dot arrows) $S2=\{(c2, c8), (c10, c11)\}$.

In order to represent semantic relatedness between concepts, six stages are proposed where each stage deals with a particular aspect of relatedness. Furthermore, in each stage semantic relatedness between concepts is presented as an adjacency matrix. In the final stage, the entire matrices are combined into one matrix to represent the final semantic relatedness between all concepts. However, for the illustration purposes, in each stage we measure the semantic relatedness between all the concepts in the ontology's fragment in figure 1. In addition, to be more realistic, this fragment will be considered as a fragment from the Open Directory Project (ODP) ontology. We use ODP as an example because it is known to be the largest user-created ontology that contains various relations. In next section we introduce all the six stages in detail.

3.1.1 Stage one: Concepts' relations types

Each type of relations has different meaning. In order to differentiate the importance of each type, we assign different weights to different relations explicitly. These weights reflect the membership degree between two concepts. A definition of relations' weights is given as follows:

3.1.1.1 Definition 1: relations' weights

As we mentions earlier, each relation r is assigned different weights w , where $w \in [0,1]$. Considering the ontology fragment in figure 1, we assign different weights for different relations as follows: $w=\alpha$ for each $H \in r$, $w=\beta$ for each $S1 \in r$ and $w=\gamma$ for each $S2 \in r$. It is important to emphasize that the weight does not reflect the semantic relatedness between concepts, but simply to distinguish between different relation types.

In the ODP for example, three relation types exist. The first type H is a hierarchal relation which treated as is-a relation. The second type $S1$ is a *symbolic* relation for multi-classification between concepts. Because concepts' duplications are not allowed in ODP, symbolic relations are presented to solve the multiple-membership problem. The last relation $S2$ is a *related* relation that represents "see also" concepts. All of these relations have different meaning, so we assign different weight for each relation as follows:

$$\text{Relations' weights} = \begin{cases} w = .9 & \text{for each } H \in r \\ w = .8 & \text{for each } S1 \in r \\ w = .5 & \text{for each } S2 \in r \end{cases}$$

Because the H relation connects hierarchal concepts, we set H to be .9. We assume that the highest possible weight between any two concepts is .9, while we assign a weight of 1 to concepts that are identical where the distance between these two concepts is zero. Regarding symbolic relations, we set $S1=.8$ because a symbolic relation seems to be slightly weaker than a hierarchal relation. Finally, we set $S2=.5$ because see-also relation seems to be weaker than the former relations, as this relation suggests further concepts that might be interesting to users, so we assigned a lower weight to it. Next, we represent the weights of all the relations between two direct concepts as an adjacency matrix W :

$$W = \begin{matrix} & \begin{matrix} c1 & c2 & c3 & c4 & c5 & c6 & c7 & c8 & c9 & c10 & c11 \end{matrix} \\ \begin{matrix} c1 \\ c2 \\ c3 \\ c4 \\ c5 \\ c6 \\ c7 \\ c8 \\ c9 \\ c10 \\ c11 \end{matrix} & \begin{bmatrix} 1 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & .5 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & .9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & .9 & .9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & .8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & .9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .8 & 0 & 1 & 0 & .9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .8 & 1 \end{bmatrix} \end{matrix}$$

3.1.2 Stage 2: Computing Semantic Relatedness for Direct-related Concepts

Direct-related concepts are defined as concepts that have one hierarchal or non-hierarchal relation between them. To estimate semantic relatedness for direct-related concepts, we introduce a simple equation that reflects some of the earlier mentioned properties in section 3 (i.e. properties: 1,2,3,6,8 and 9). We identify the semantic relatedness between two concepts as a matrix $D_{c1,c2}$ as follows:

$$D_{ci,cj} = \begin{cases} 1 & \text{if } ci = cj \\ \sigma & \text{if } ci \neq cj \text{ and } (ci, cj) \in H, S1 \text{ or } S2 \\ 0 & \text{if otherwise} \end{cases}$$

Where σ is the semantic relatedness for two direct-related concepts, and it is estimated as follows:

$$Rel_{Direct}(ci, cj) = \frac{w(ci, cj) \times c_{j_level}}{c_{j_max_depth}}$$

Where: ci is the super-concept of cj and cj is a sub-concept of ci . $w(ci, cj)$ is the weight of the relation between ci and cj and it is identified in stage one. c_{j_level} is the hierarchal level of cj . $c_{j_max_depth}$ is the maximum depth of the concept cj (max_depth was identified in Property 9 section3).

In order to measure the semantic relatedness between ci and cj , we first consider the weight $w(ci, cj)$ based on the relation between them, then we calculate the level of the sub-concept (c_{j_level}) using the following formula:

$$c_{j_level} = \sum |(\text{root} \rightarrow ci)| + |(ci \rightarrow cj)|$$

c_{j_level} is calculated by summing up the number of concepts between $(\text{root} \rightarrow ci)$ and between $(ci \rightarrow cj)$. For example, when calculating the relatedness between concepts $c7$ and $c10$ in figure 1 the level of $c10$ is 4, while when calculating the relatedness between $c11$ and $c10$ the level of $c10$ is 5 (because $|(root \rightarrow c11)| = 4$, and $|(c11 \rightarrow c10)| = 1$).

The $c_{j_max_depth}$ in the equation which is the maximum depth of the most specific leaf concept connected to concept cj . In addition, when calculated the max_depth we consider all the cj 's hierarchal concepts, and all the hierarchal concepts under non-hierarchal concepts. For illustration, the maximum depth of a concept $c2$ is 3 because $c2$ is connected to $c8$, and there are no concepts under $c8$, while the maximum depth of a concept $c3$ is 6 as we consider both the hierarchal concept $c11$ and the first non-hierarchal sub-concept $c10$.

Finally, by dividing the concept level by its maximum level depth, we satisfy the Hierarchal and Semantic Relations Property, Depth-Relative Property, and Minimal Distance Property. Therefore, deeper and closer concepts are more related than higher and far concepts. Next Matrix (Matrix D) shows the semantic relatedness for all the direct-related concepts in figure 1.

$$D = \begin{matrix} & \begin{matrix} c1 & c2 & c3 & c4 & c5 & c6 & c7 & c8 & c9 & c10 & c11 \end{matrix} \\ \begin{matrix} c1 \\ c2 \\ c3 \\ c4 \\ c5 \\ c6 \\ c7 \\ c8 \\ c9 \\ c10 \\ c11 \end{matrix} & \begin{bmatrix} 1 & .6 & .3 & .36 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ .6 & 1 & 0 & 0 & 0 & 0 & 0 & .5 & 0 & 0 & 0 \\ .3 & 0 & 1 & 0 & .45 & .54 & 0 & 0 & 0 & 0 & 0 \\ .36 & 0 & 0 & 1 & 0 & 0 & .54 & 0 & 0 & 0 & 0 \\ 0 & 0 & .45 & 0 & 1 & 0 & 0 & .9 & .6 & 0 & 0 \\ 0 & 0 & .54 & 0 & 0 & 1 & 0 & 0 & .533 & 0 & 0 \\ 0 & 0 & 0 & .54 & 0 & 0 & 1 & 0 & .66 & .72 & 0 \\ 0 & .5 & 0 & 0 & .9 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .6 & .533 & .66 & 0 & 1 & 0 & .75 \\ 0 & 0 & 0 & 0 & 0 & 0 & .72 & 0 & 0 & 1 & .5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .75 & .8 & 1 \end{bmatrix} \end{matrix}$$

3.1.3 Stage 3: Computing Semantic Relatedness for Transitive-related Concepts

Concepts ($c1, c2$) are transitive-related when $c1$ and $c2$ are not direct-related concepts, and $c1$ is a super-concept of $c2$. The relatedness of transitive-related concepts is calculated considering the *Transitivity property* in section 3. This property states that if a concept x has sub-concepts, then x is to some extent related to all of its descendants. In our approach semantic relatedness between two concepts is restricted to all hierarchal and non-hierarchal concepts, but not non-hierarchal under non-hierarchal concepts. Similar to the previous stage (stage 2) we add this rule to preserve feasibility and meaning of the ontology. In addition to this property, we also consider a number of important properties that were discussed in section 3. We first consider the Minimal Distance Property where closer concepts are more related than far ones. However, relying solely on this property is not sufficient to represent relatedness between two concepts as we need also to incorporate the depth of concepts. Thus, we incorporate the Depth-Relative Property which states that concepts deep in an ontology are more closely related than higher concepts.

Considering the ontology example in figure 1, ($c1, c2$) and ($c5, c8$) for example, are not transitive-related concepts because they are direct-related concepts. Similarly, ($c6, c7$) also are not transitive-related concepts because $c6$ and $c7$ are connected by two non-hierarchal relations. On the other hand, concepts such as ($c6, c11$) are transitive-related as $c6$ is connected to $c9$ by a non-hierarchal relation, and $c9$ is connected to $c11$ by a hierarchal relation which

is allowed. To compute the relatedness of transitive-related concepts, we use the following formula:

$$Rel_{Transitive}(c_i, c_j) = \frac{\sum DR(c_i \rightarrow c_j) \times c_{i_level}}{c_{j_level} \times \sum |(c_i \rightarrow c_j)|}$$

Where $\sum DR(c_i \rightarrow c_j)$ is the sum of the semantic relatedness weights between c_i and c_j . If a c_i is not directly related to c_j e.g. $c_i \rightarrow c_x \rightarrow c_j$ then the $\sum DR(c_i \rightarrow c_j) = DR(c_i, c_x) + DR(c_x, c_j)$. c_{i_level} is the total number of concepts between the root and concept c_i , and it is given as follows:

$$c_{i_level} = \sum |(root \rightarrow c_i)|$$

c_{j_level} , however, is the total number of concepts between the root and the super-concept c_i and the number of concepts between c_i and c_j :

$$c_{j_level} = \sum |(root \rightarrow c_i)| + |(c_i \rightarrow c_j)|$$

Finally, $\sum |(c_i \rightarrow c_j)|$ is the number of relations between c_i and c_j .

This formula is sensitive to both levels of c_i and c_j as well as to the distance between them. We compute the level of concepts by summing up the number of concepts between two concepts rather than selecting the real depth. This is essential to overcome non-uniform relations in an ontology as some super-concepts in deep levels may be linked to the same or even higher level sub-concepts (e.g. as concepts c_6 and c_7 in figure 1). Thus, computing the level this way allows us to more accurately estimate the relatedness values between hierarchal and non-hierarchal concepts.

Next Matrix (Matrix **TG**) shows the transitive-graph for all transitive concepts in figure 1. The followed matrix (matrix **T**) shows the semantic relatedness for all the transitive-related concepts in figure 1 as a fragment of ODP.

		c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}
$TG=$	c_1	0	0	0	0	1	1	1	1	1	1	1
	c_2	0	0	0	0	0	0	0	0	0	0	0
	c_3	0	0	0	0	0	0	1	1	1	1	1
	c_4	0	0	0	0	0	0	0	0	0	1	1
	c_5	0	0	0	0	0	0	1	0	0	1	1
	c_6	0	0	0	0	0	0	0	0	0	1	1
	c_7	0	0	0	0	0	0	0	0	0	0	1
	c_8	0	0	0	0	0	0	0	0	0	0	0
	c_9	0	0	0	0	0	0	0	0	0	1	0
	c_{10}	0	0	0	0	0	0	0	0	0	0	0
	c_{11}	0	0	0	0	0	0	0	0	0	0	0

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}
$T=$	c_1	1	0	0	.125	.14	.15	.183	.112	.135	.105
	c_2	0	1	0	0	0	0	0	0	0	0
	c_3	0	0	1	0	0	.228	.337	.262	.216	.24
	c_4	0	0	0	1	0	0	0	0	.315	.234
	c_5	.125	0	0	0	1	0	.378	0	0	.358
	c_6	.14	0	0	0	0	1	0	0	0	.384
	c_7	.15	0	.228	0	.378	0	1	0	0	.366
	c_8	.183	0	.337	0	0	0	0	1	0	0
	c_9	.112	0	.262	0	0	0	0	0	1	.51
	c_{10}	.135	0	.216	.315	.358	0	0	0	.51	1
	c_{11}	.105	0	.24	.234	.405	.384	.366	0	0	0

3.1.4 Stage 4: Computing Semantic Relatedness for Sibling-related Concepts

Sibling-related concepts are defined as concepts that share at least one common super-concept, and none of these concepts is a super-concept of the other. In order to measure the semantic relatedness for sibling-related concepts, we first consider Minimal Distance Property (property 5) that states that closer concepts are more related than far ones. The depth of concepts is also considered as it impacts on the overall relatedness as the concepts deep in an ontology are more related than higher concepts (see property 6 in section 3). We also consider the property 9 which compute the maximum depth of a concept.

To measure the semantic relatedness for sibling-related concepts, we borrow the original approach for measuring semantic similarity suggested by Wu and Palmer [14], as follows:

$$sim_{WP} = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3}$$

However, unlike [14] who only considered the length between the path between two concepts and the depth of the common super-concept, we further incorporate all the relation types, the maximum depth of the super-concept, and the depth of sibling concepts as follows:

$$Rel_{Sibling}(c_i, c_j) = \frac{2 \times N_{c_level}}{N_i + N_j + 2 \times N_{c_level}}$$

$$N_i = (c_{i_max_depth} - (\frac{c_{i_max_depth} - N_{c_level}}{c_{i_depth}})) \times (2 - Rel(c_{common}, c_i))$$

$$N_j = (c_{j_max_depth} - (\frac{c_{j_max_depth} - N_{c_level}}{c_{j_depth}})) \times (2 - Rel(c_{common}, c_j))$$

Where N_{c_level} is the depth of the common super-concept, $c_{i_max_depth}$ is the maximum depth of c_i , $c_{j_max_depth}$ is the maximum depth of c_j . Both of these maximum depths are calculating considering the restriction that was set in stage 2. In this restriction we just consider all the hierarchal concepts, and all the hierarchal concepts under the first level of non-hierarchal concepts. c_{i_depth} is the depth of the concept c_i while c_{j_depth} is the depth of the concept c_j . Finally, $Rel(c_{common}, c)$ is the semantic relatedness between the super-concept and c_i or c_j . Additionally, if (common concept, c_i) are direct-related concepts then the semantic relatedness is calculated using stage2, while if these concepts are transitive-related then the equation in the stage 3 is used.

This measure satisfies: the positive, the Hierarchal and Semantic Relations, the Minimal Distance, Depth-Relative and concepts' maximum depth properties. The Hierarchal and Semantic Relations property is satisfied because the (c_{common}, c) is computed using either the direct-related or transitive-related equations. The Minimal Distance property is also satisfied as shorter distance between sibling-concepts and a common concept means higher relatedness. Next matrix (matrix **SG**) represents the all the sibling-related concepts in figure 1, while the matrix **S**

identifies the semantic relatedness for all the sibling-related concepts in figure 1 which is treated as an ODP ontology.

$$SG = \begin{matrix} & \begin{matrix} c1 & c2 & c3 & c4 & c5 & c6 & c7 & c8 & c9 & c10 & c11 \end{matrix} \\ \begin{matrix} c1 \\ c2 \\ c3 \\ c4 \\ c5 \\ c6 \\ c7 \\ c8 \\ c9 \\ c10 \\ c11 \end{matrix} & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \end{matrix}$$

$$S = \begin{matrix} & \begin{matrix} c1 & c2 & c3 & c4 & c5 & c6 & c7 & c8 & c9 & c10 & c11 \end{matrix} \\ \begin{matrix} c1 \\ c2 \\ c3 \\ c4 \\ c5 \\ c6 \\ c7 \\ c8 \\ c9 \\ c10 \\ c11 \end{matrix} & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .186 & .20 & .154 & .172 & .172 & 0 & .145 & .163 & .14 \\ 0 & .186 & 0 & .155 & 0 & 0 & .135 & 0 & 0 & 0 & 0 \\ 0 & .20 & .155 & 0 & .132 & .145 & 0 & .161 & .125 & 0 & 0 \\ 0 & .154 & 0 & .132 & 0 & .234 & .118 & 0 & 0 & .113 & 0 \\ 0 & .172 & 0 & .145 & .234 & 0 & .128 & .255 & 0 & .122 & 0 \\ 0 & .172 & .135 & 0 & .118 & .128 & 0 & .317 & 0 & 0 & 0 \\ 0 & 0 & 0 & .161 & 0 & .255 & .317 & 0 & .343 & .313 & .32 \\ 0 & .145 & 0 & .125 & 0 & 0 & 0 & .343 & 0 & 0 & 0 \\ 0 & .163 & 0 & 0 & 0 & .122 & 0 & .313 & 0 & 0 & 0 \\ 0 & .14 & 0 & 0 & 0 & 0 & .14 & .32 & 0 & 0 & 0 \end{bmatrix} \end{matrix}$$

3.1.5 Stage 5: Computing Semantic Relatedness for Parent-related Concepts

Parent-related concepts are defined as concepts that share at least one common sub-concept, and none of these concepts is a super-concept of the other (e.g. c2 and c5). Obviously any parent-related concepts are at the same time sibling-related, while the reverse is not true. Therefore, if two concepts are both parent-related and sibling-related, the final semantic relatedness value would be the one that is related to the shortest path between the two concepts. For example in figure 1, concepts (c2, c3) are both parent-related in c8 with a length of 3, and sibling-related in c1 with a length of 2. Because the length between (c2, c3) for the sibling-relation is lower than in parent-relation, the semantic relatedness resulted from the sibling-related stage is chosen. However, if the length of both relation types is the same, then the highest semantic relatedness value is chosen.

In order to measure the semantic relatedness for parent-related concepts, we apply a similar equation as in stage 4 with slight modifications as follows:

$$Rel_{parent}(c_i, c_j) = \frac{N_{c_level}}{N_i + N_j}$$

$$N_i = (N_{i_max_depth} - (\frac{N_{i_max_depth} - N_{c_level}}{c_{i_depth}})) \times (2 - Rel(c_{common}, c_i))$$

$$N_j = (N_{j_max_depth} - (\frac{N_{j_max_depth} - N_{c_level}}{c_{j_depth}})) \times (2 - Rel(c_{common}, c_j))$$

Where N_{c_level} is the higher depth of the common sub-concept. As a common sub-concept in this stage has two super-concepts, this means that this sub-concept may have two different levels based on each super-concept. For example, in the figure1, c2 and c5 are parent-related concepts because they have one common sub-concept c8. However, the sub-concept c8 has two depths: (c1→c2→c8) = 3, and (c1→c8→c3→c5→8) = 4. In order to solve this problem we select the higher level to keep the relatedness value from being inflated (e.g. $N_{c8_level} = 3$ in the previous

example). Clearly we made some modifications in this equation. The main reason why we modified this equation is that the N_{c_level} which is the depth of the common sub-concept, would obviously has a deeper level than both c_i and c_j . If we use the same equation as in stage 4, the relatedness value would be too large as we multiply the N_{c_level} by 2. Therefore, in this stage we solve this problem by normalizing the equation so the final semantic relatedness value would be realistic.

Next matrix (matrix **PG**) represents the all the parent-related concepts in figure 1, while the matrix **P** represents the semantic relatedness for all the parent-related concepts in figure 1 which is treated as an ODP ontology.

$$PG = \begin{matrix} & \begin{matrix} c1 & c2 & c3 & c4 & c5 & c6 & c7 & c8 & c9 & c10 & c11 \end{matrix} \\ \begin{matrix} c1 \\ c2 \\ c3 \\ c4 \\ c5 \\ c6 \\ c7 \\ c8 \\ c9 \\ c10 \\ c11 \end{matrix} & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{matrix}$$

$$PG = \begin{matrix} & \begin{matrix} c1 & c2 & c3 & c4 & c5 & c6 & c7 & c8 & c9 & c10 & c11 \end{matrix} \\ \begin{matrix} c1 \\ c2 \\ c3 \\ c4 \\ c5 \\ c6 \\ c7 \\ c8 \\ c9 \\ c10 \\ c11 \end{matrix} & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .29 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .2 & 0 & 0 & 0 & .22 & 0 & 0 \\ 0 & .29 & 0 & .2 & 0 & .279 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .279 & 0 & .307 & 0 & 0 & .32 & 0 \\ 0 & 0 & 0 & 0 & 0 & .307 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .22 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .32 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{matrix}$$

3.1.6 Stage 6: Aggregating Adjacency Matrices

After measuring semantic relatedness for direct-related, transitive-related, sibling-related and parent-related concepts, we then aggregate all the adjacency matrices using a composition function $\otimes$ as follows:

$$AAM = [Rel(DR) \otimes Rel(TR) \otimes Rel(SR) \otimes Rel(PR)] = \max(Rel(DR) \cdot Rel(TR) \cdot Rel(SR) \cdot Rel(PR)).$$

Where:

AAM= is the Aggregated Adjacency Matrices that represents the final semantic relatedness matrix for all related concepts.

Rel(DR) = direct_related matrix.

Rel(TR) = transitive_related matrix.

Rel(SR) = sibling_related matrix.

Rel(PR) = parent_related matrix.

If there is more than one relation between two concepts, the relation with the maximum weight is selected for the **AAM**, which satisfies the Multiple Paths Property. Next matrix shows the last semantic relatedness for all concepts in figure 1.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11
c1	1	.6	.3	.36	.125	.14	.15	.183	.112	.135	.105
c2	.6	1	.186	.2	.29	.172	.172	.5	.145	.163	.14
c3	.3	.186	1	.155	.45	.54	.228	.337	.262	.216	.24
c4	.36	.2	.155	1	.2	.145	.54	.161	.22	.315	.234
c5	.125	.29	.45	.2	1	.279	.378	.9	.6	.358	.405
c6	.14	.172	.9	.145	.279	1	.307	.255	.533	.32	.384
c7	.15	.172	.228	.54	.378	.307	1	.371	.66	.72	.366
c8	.183	.5	.337	.161	.9	.255	.317	1	.343	.313	.32
c9	.112	.145	.262	.22	.6	.533	.66	.343	1	.51	.75
c10	.135	.163	.216	.315	.358	.32	.72	.313	.51	1	.5
c11	.105	.14	.24	.234	.405	.384	.366	.32	.75	.8	1

4. EVALUATING THE SEMANTIC RELATEDNESS APPROACH

We conduct three experiments to evaluation our semantic relatedness approach. According to Maguitman and Mencze [18] the first question that needs to be answered when evaluating any semantic similarity or relatedness approach is whether the proposed approach detects different results from the other semantic similarities approaches. Furthermore, we add another important question of whether the proposed approach is sensitive to different properties including the distance and the depth of concepts or not. Therefore, in the first experiment we try to address these questions. In the second experiment, we perform a comparative evaluation by comparing the performance of our approach against other proposed approaches using a benchmark dataset. Finally, we validate our approach by using the ODP ontology, and apply the proposed approach on data collected from real users. All of these experiments are listed as follows:

4.1 Experiment1: Analysis of differences

In this section, we try to prove that our proposed approach is sensitive to all the proposed properties in section 3 by comparing the semantic relatedness results from our approach with three common approaches namely: Wu and Palmer [14], Leacock [21] and Maguitma and Menczer [18]. For the purpose of this experiment, we use the ontology fragment in figure1 as an ODP fragment. Next we examine different properties and how much our approach is sensitive to them.

Property1: considering different relation types:

Concepts	$sim_{Wu\&Palmer}$	$sim_{Leacock}$	$sim_{M\&M}$	<i>Our approach</i>
Rel(c6,c9)	.85	.69	.30	.533
Rel(c5,c9)	.85	.69	.34	.6

Both $sim_{Wu\&Palmer}$ and $sim_{Leacock}$ have not consider different relations, so both Rel(c6,c9) and Rel(c5,c9) scored the same weight. On the contrary, both $sim_{M\&M}$ and our approach consider all types of relations, and hence Rel(c6,c9) and Rel(c5,c9) have different weights.

Property2: general concepts are less related than specific concepts:

This property states that concepts deep in an ontology are more related than concepts in higher levels.

Concepts	$sim_{Wu\&Palmer}$	$sim_{Leacock}$	$sim_{M\&M}$	<i>Our approach</i>
Rel(c3,c4)	0	.60	.27	.155
Rel(c2,c4)	0	.60	.27	.20
Rel(c6,c5)	.5	.60	.31	.234

Rel(c8,c9)	.67	.60	.37	.343
------------	-----	-----	-----	------

In $sim_{Leacock}$ approach, the concepts' depths are ignored, so that all concepts have the same weight. While, $sim_{Wu\&Palmer}$, $sim_{M\&M}$ and our approach satisfy the condition of this property as they are sensitive to the depth of the concepts. However, because in our approach we consider the maximum depth of each concept instead of the maximum depth of the whole ontology, we have more precise and more sensitive measurement.

Property3: closer concepts are more related than far ones:

Concepts	$sim_{Wu\&Palmer}$	$sim_{Leacock}$	$sim_{M\&M}$	<i>Our approach</i>
Rel(c6,c5)	.5	.60	.31	.234
Rel(c6,c8)	.4	.47	.26	.128

All the approaches satisfy the condition of that closer concepts are more related than far concepts.

Property4: in the transitivity-related: closer and deep concepts are more related than far and higher concepts:

Concepts	$sim_{Wu\&Palmer}$	$sim_{Leacock}$	$sim_{M\&M}$	<i>Our approach</i>
Rel(c3,c9)	.5	.60	.29	.262
Rel(c4,c10)	.5	.60	.25	.315
Rel(c3,c11)	.4	.47	.25	.24
Rel(c4,c11)	.4	.47	.22	.234

Both $sim_{Wu\&Palmer}$ and $sim_{Leacock}$ satisfy the property 4. However, both approaches failed to distinguish between the weights of (c3,c9) and (c4,c10), and between (c3,c11) and (c4,c11). $sim_{M\&M}$ also satisfy property 4, but it also failed to distinguish between (c3,c11) and (c4,c11). This is due to the fact that they only consider the distance and the global depth of the ontology, but not the depth of each concept as in our approach. Therefore, in our approach all the weights are different as they are sensitive to the maximum depth of each concept.

Property5: in transitivity-related: deep concepts are more related than higher ones:

Concepts	$sim_{Wu\&Palmer}$	$sim_{Leacock}$	$sim_{M\&M}$	<i>Our approach</i>
Rel(c1,c5)	0	.60	.12	.125
Rel(c5,c11)	.66	.60	.14	.405

Wu and Palmer and partly satisfied this property by measuring the similarity of (c5,c11). On the other hand, Leacock failed to satisfy this property as both (c1,c5) and (c5,c11) received the same weight. However, both Maguitma and Menczer and our approach managed to distinguish between the weight of (c5,c11) and (c1,c5), though our approach assigned more weight to (c5,c11) compared to Maguitma and Menczer. This shows that our approach is more sensitive in handling this property, and hence is capable to reflect that a deeper concept is more related than a higher level concept.

Property6: Symmetry

Concepts	$sim_{Wu\&Palmer}$	$sim_{Leacock}$	$sim_{M\&M}$	<i>Our approach</i>
Rel(c10,c11)	.85	.77	.42	.5
Rel(c11,c10)	.88	.77	.45	.8
Rel(c9,c7)	.85	.77	.44	.66
Rel(c7,c9)	0	.30	.44	.112
Rel(c2,c4)	0	.60	.27	.20
Rel(c4,c2)	0	.60	.27	.20

Unlike all approaches, our approach is both symmetrical for concepts that have one relation between them, while non-symmetrical for concepts with more than one relation connected them. In the above table, $(c10,c11) \neq (c11,c10)$ because there are two types of relations between them. Similarly, $(c9,c7) \neq (c7,c9)$ as there are two paths between them. On the other hand, because there is just one path between $c4$ and $c2$, $Rel(c4,c2)$ and $(c2,c4)$ is symmetrical and that $Rel(c4,c2) = (c2,c4) = .20$.

From the above evaluation, we showed that our approach managed to satisfy all the aspects that associated to the unique structure of an ontology. In addition, our approach proves that it can estimate more sensitive semantic relatedness weights considering properties such as the maximum depth of concepts and the distance.

4.2 Experiment2: Comparative study

In the previous experiment we evaluated the sensitiveness of our approach using the ontology fragment in figure 1. However, in this stage we found that it is necessary to evaluate the performance of our approach based on a benchmark dataset, so the results of our approach would be comparable with other previous studies in the same field. For this reason, in this experiment we use Miller and Charles [22] dataset which consists of 30 pairs of nouns extracted from the WordNet [23]. In their experiment 38 undergraduate students were asked to rate each pair of noun on a scale from 0 (no similarity) to 4 (perfect synonym). The average rating given by all students for each noun pair represents a good estimate of the similarity degree between two words. The correlation between subject ratings was reported to be .90, and hence .90 is used as the upper bond ratio.

4.2.1 Experiment process:

In this experiment we attempt to compute the correlation between the ratings of our semantic relatedness approach and the mean ratings reported by Miller and Charles. Furthermore, we compare our results against six other semantic similarity approaches namely: Wu and Palmer [14], Resnik [5], Leacock [21], Jiang & Conrath [13], Lin [17] and Hirst and St-Onge [24]. For the purpose of this evaluation, we assign a weight of .9 to all the relation types in the WordNet.

4.2.2 Results:

Table 1 lists the complete semantic relatedness results estimated by our approach and Miller and Charles mean ratings.

Table 1. Semantic relatedness by word

Word Pairs		Proposed approach	M&C mean
car	automobile	1	3.92
gem	jewel	1	3.84
journey	voyage	0.83	3.84
boy	lad	0.9	3.76
coast	shore	0.771	3.7
asylum	madhouse	0.9	3.61
magician	wizard	1	3.5
midday	noon	1	3.42
furnace	stove	0.2477	3.11
food	fruit	0.178	3.08
bird	cock	0.818	3.05
bird	crane	0.513	2.97
tool	implement	0.48	2.95
brother	monk	0.9	2.82
crane	implement	0.246	1.68
lad	brother	0.288	1.66
journey	car	0.0393	1.16
monk	oracle	0.264	1.1
cemetery	woodland	0.1809	0.95
food	rooster	0.268	0.89
coast	hill	0.2679	0.87
forest	graveyard	0.1809	0.84
shore	woodland	0.2256	0.63
monk	slave	0.228	0.55
coast	forest	0.2208	0.42
lad	wizard	0.315	0.42
chord	smile	0.198	0.13
glass	magician	0.114	0.11
noon	string	0.053	0.08
rooster	voyage	0.04	0.08

Table 2 shows the correlation coefficient results between seven componential approaches and Miller and Charles ratings mean.

Table 2. Correlation coefficient values for 7 methods

Measure	Correlation with M&C mean
Wu&Palmer	0.74
Resnik	.774
Leacock	.816
Jiang & Conrath	.8484
Lin	.8224
Hirst & St-Onge	.744
Our approach	.853

It can be clearly seen that the semantic relatedness for our approach outperformed all the listed approaches including edge-based methods (Wu&Palmer [14], Leacock [21] and Hirst and St-Onge [24]) and information content (Resnik [5], Jaing and Conrath [13] and Lin [17]). Unlike all the listed methods, in our approach we considered different properties. Furthermore, the good correlation value for our approach is also resulted from considering a one unique property which is the maximum depth of each concept when measuring the semantic relatedness. It is worth noting that the correlation ratio of our approach boosts to .895 which is very close to the upper bond .90 if we disregard Food-Fruit word pair. This is because the semantic relation between

these two nouns is not exploited in WordNet, which lead our approach to estimate a relatively lower semantic relatedness between them.

In this experiment we showed that our proposed approach achieved a good correlation value to the human-subject rating reported by Miller and Charles. However, in the next experiment we attempt to validate our approach by using the ODP ontology, and apply the proposed approach on data collected from real users.

4.3 Experiment3: Validation by user study

In the previous experiment we evaluated our approach based on a benchmark dataset reported by Miller and Charles. However, in this experiment, we think it is important to evaluate our approach when using an ontology that has different relation types and on a real user data we collected ourselves. For this reason, we conducted a human-subject experiment where real users are asked to conduct a survey. A survey of 15 questions (i.e. each question has two parts) about relatedness between web pages were created, and distributed to 19 participants. For each question, firstly users are given a target web page and asked to select the more related web page from two candidate web pages. Then users are asked to estimate the relatedness between the target web page and the candidate web pages selecting a rate from 0 to 10, where 0 is not related and 10 is matched (see figure 2). Each web page in each question represents a single concept from the ODP ontology. These concepts were selected randomly from the ODP.

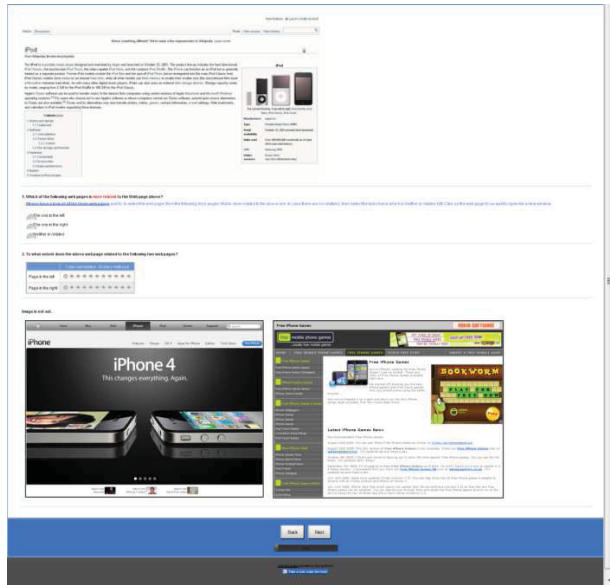


Figure 2. An example of a question in the survey

4.3.1 Experiment process:

In this experiment, we first try to find whether our approach is capable of producing accurate judgments regarding the question of which candidate web page is more related to the target web page. Then, we compare the performance of our approach in terms of the estimated relatedness weights between web pages and the user ratings. For both experiments, we compare user predictions to three different approaches namely: our approach, Wu & Palmer [14], Leacock [21] and Maguitma and Menczer [18].

4.3.2 Results:

In order to find whether our approach is capable of producing accurate judgments regarding the question of which candidate web page is more related to the target web page, we test four different approaches to see which one is better predictor of users' judgments. Table 3 summarizes the accuracy results for all the four approaches using the following accuracy equation:

$$Accuracy = \frac{\text{number of correct answers}}{\text{total number of questions}}$$

Where *number of correct answers* is the total number of predictions the match user judgments, and the *total number of questions* is 15.

Table 3 shows that applying our approach of measuring semantic relatedness of web pages and concepts from ODP achieved a good accuracy of 66%. Furthermore, our approach yields a substantial improvement over the tree-based approaches: Wu & Palmer and Leacock, that both achieved an accuracy of 46%, and noteworthy improvement over the Maguitma & Menczer approach who achieved 53%. These results showed that our approach captures some semantics that are not captured in the other measures.

Table 3. Total accuracy results for all approaches

Approach	Our approach	Wu & Palmer	Leacock	Maguitma & Menczer
Accuracy	66%	46%	46%	53%

In the next experiment, the relatedness values for each of the 30 pairs of web pages were computed by all the four approaches and compared to the users' estimations that were used as the base/reality values (see figure 3). It is clear that Wu and Palmer and Leacock techniques are not accurate to compute semantic relatedness as a great deal of the predictions is completely different than the users' judgments. However, Wu and Palmer approach shows better predictions than Leacock approach. The reason for this is that Wu and Palmer consider the depth of concepts when computing similarity, while Leacock considers just the global ontology depth. Maguitma and Menczer, however, shows better predictions compared to the former two approaches due to considering more important properties such as property 4. On the other hand, our approach shows a much better accuracy compared to the former approaches. In addition, there are number of questions namely: 1,4,11,12,15,24 and 26 where the relatedness of our approach estimates almost identical relatedness to the users' judgments.

Table 4, illustrates the correlations results between the predictions computed four approaches and the user ratings. Wu and Palmer approach shows a better correlation (.753) than Leacock approach (.74). The reason for this is that Wu and Palmer in their measure they consider the depth of concepts when computing similarity, while Leacock considers just the global ontology depth. Maguitma and Menczer, however, show better correlation than the former two approaches (.765). This is mainly due to that Maguitma and Menczer consider more properties when computing the similarity between two concepts. On the other hand, our proposed method achieves 9.8% better correlation over Maguitma and Menczer, 11.5% over Wu&Palmer and 15% over Leacock state-of-the-art methods. This is mostly due to

considering more factors that affect the relatedness between concepts such as the maximum depth of each concept.

Table 4. Correlation results for four approaches

Approach	Proposed approach	Wu & Palmer	Leacock	Maguitma & Menczer
Correlation	.84	.753	.73	.765

Finally, we can safely conclude that our measurement consistently returns more accurate results in comparison with other semantic similarity measurements.

5. CONCLUSIONS AND FUTURE WORK

The process of retrieving information in the Internet becomes more and more difficult as a result of the rapid growth of heterogeneous and independent data networks and repositories. One way to process this type of data is by exploiting the structure and relations that form such a network. In this paper, we particularly focused on ontologies as the main knowledge

resource which advantageous from having a more complex structure than taxonomies as well as the ability of handling more relation types including hierarchal and non-hierarchal ones. Additionally, we proposed a semantic relatedness approach to measure the relatedness between concepts in an ontology taking into the consideration different important properties that impact positively on the overall performance. We also conducted three different experiments to evaluate different aspects in our proposed approach. The obtained results demonstrate the effectiveness of using our approach to measure semantic relatedness between concepts in an ontology over taxonomy-based approaches.

As part of further work, we will attempt to integrate the proposed approach with a dynamic user profile that captures user interests. The main goal is to infer concepts (from a reference ontology) that are related semantically to the user actual interests. Furthermore, the semantic relatedness approach can also be used to group related concepts together. These related interests would be given a higher weight if they appear together in the user profile than if they appear in an isolated way.

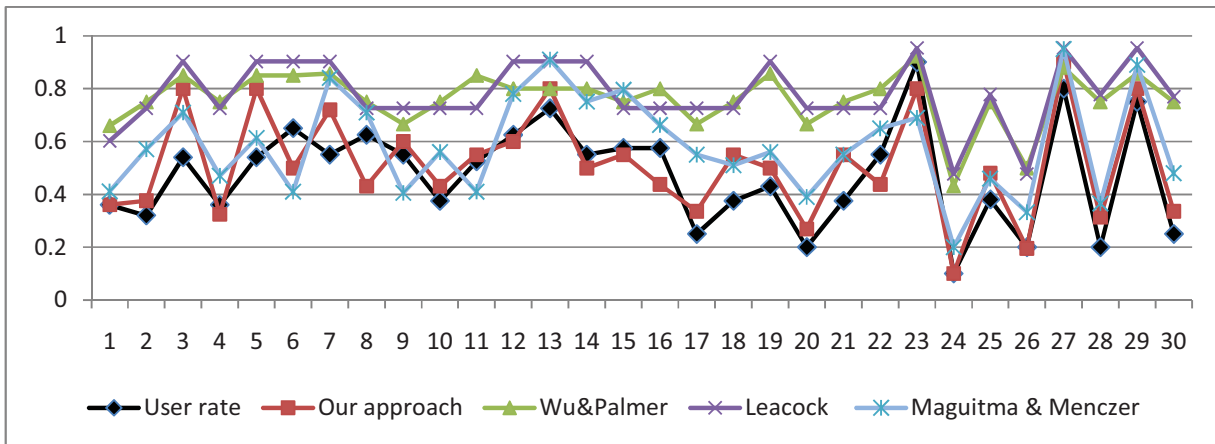


Figure 3. Comparing the estimated values suggested by three approaches

6. REFERENCES

- [1] Mitra, M., Singhal, A. and Buckley, C. 1998. Improving automatic query expansion. In *Proc. of 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. pp. 206-214.
- [2] Vlez, B., Wiess, R., Sheldon, M. And Gifford, D. 1997. Fast and effective query refinement. In *Proc. of 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. pp. 6-15.
- [3] Matsuo, Y., Mori, J., Hamasaki, M., Ishida, K., Nishimura, T., Takeda, H., Hasida, K. and Ishizuka, I. 2005. Polyphonet: An advanced social network extraction system. In *Proc. of 15th International World Wide Web Conference*.
- [4] Mika, P. 2005. Ontologies are us: A unified model of social networks and semantics. In *Proc. of ISWC2005*.
- [5] Resnik, P. 1999. Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*. pp. 95-130.
- [6] Rosenfield, R. 1996. A maximum entropy approach to adaptive statistical modelling. *Computer Speech and Language*. pp. 187-228.
- [7] Ge, J. and Qiu, Y. 2008. Concept Similarity Matching Based on Semantic Distance. In *Proc. of 4th International Conference on Semantic, Knowledge and Grid*.
- [8] Rada, R., Mili, H., Bicknell, E. and Blettner, M. 1998. Development and Application of a Metric on Semantic Nets. *IEEE Trans. on Systems, Man, and Cybernetics*, 19(1). pp.17-30.
- [9] Razmerita, L. and Lytras, M. 2008. Ontology-Based User Modelling Personalization: Analyzing the Requirements of a Semantic Learning Portal. *Lecture Notes in Artificial Intelligence*, Vol. 5288, pp. 354 – 363.

- [10] Cross, V. 2004. Fuzzy semantic distance measures between ontological concepts. *Fuzzy Information*. 04, IEEE Annual Meeting of the Volume 2.
- [11] Sussna, M. 1993. Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network. *Proceedings of the Second International Conference on Information and Knowledge Management*, CIKM'93. pp. 67-74.
- [12] Knappe, R. 2005. *Measures of Semantic Similarity and Relatedness for Use in Ontology-based Information Retrieval*. PhD. Roskilde University
- [13] Jiang J. and Conrath, D. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *In Proceedings of International Conference on Research in Computational Linguistics*, Taiwan.
- [14] Wu, Z. and Palmer, M. 1994. Verb Semantics and Lexical Selection. *32nd Annual Meeting of the Association for Computational Linguistics*. pp. 133-138.
- [15] Richardson, R., Smeaton, A. F., and Murphy, J. 1994. *Using WordNet as a knowledge base for measuring semantic similarity between words*. Working paper CA 1294, Dublin City University, School of Computer Applications, Dublin, Ireland.
- [16] Kim, Y. and Kim, J. 1990. *A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph*, Journal of Documentation, Vol. 46 (2). pp.113–136.
- [17] Lin, D. 1998. An Information-Theoretic Definition of Similarity. *15th International Conference on Machine Learning*, pp. 296-304.
- [18] Maguitman, A. and Mencze, F. 2005. Algorithmic detection of semantic similarity. *In Proceedings of the 14th International Conference on World Wide Web (WWW 2005)*. Chiba, Japan.
- [19] Open Directory Project (ODP). 2010. <http://www.dmoz.org/> September 2010.
- [20] Song, L., Ma, J., Liu, H., Lian, L. and Zhang, D. 2007. Fuzzy Semantic Similarity Between Ontological Concepts. *In Advances and Innovations in Systems, Computing Sciences and Software Engineering*. pp.275-280.
- [21] Leacock, C. and Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. *In Fellbaum, C., ed.: WordNet: An electronic lexical database*. MIT Press. pp. 265–283.
- [22] Miller, G. and Charles, W. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*. pp. 1–28.
- [23] WordNet. 2010. <http://wordnet.princeton.edu/>. November 2010.
- [24] Hirst, G. and St-Onge, D. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. *In Fellbaum*. pp. 305–332.