Paradox in Applications of Semantic Similarity Models in Information Retrieval

Hai Dong, Farookh Khadeer Hussain, and Elizabeth Chang

Digital Ecosystems and Business Intelligence Institute, Curtin University of Technology,
GPO Box U1987 Perth, Western Australia 6845, Australia
{hai.dong,farookh.hussain,elizabeth.chang}@cbs.curtin.edu.au

Abstract. Semantic similarity models are a series of mathematical models for computing semantic similarity values among nodes in a semantic net. In this paper we reveal the paradox in the applications of these semantic similarity models in the field of information retrieval, which is that these models rely on a common prerequisite – the words of a user query must correspond to the nodes of a semantic net. In certain situations, this sort of correspondence can not be carried out, which invalidates the further working of these semantic similarity models. By means of two case studies, we analyze these issues. In addition, we discuss some possible solutions in order to address these issues. Conclusion and future works are drawn in the final section.

Keywords: information retrieval, semantic net, semantic similarity models.

1 Introduction

Semantic similarity models are a series of mathematical models for computing semantic similarity values among nodes in a semantic net [7]. These models are broadly applied in the field of information clustering and retrieval. For their applications in the field of information retrieval, a common characteristic of these models' working procedures can be concluded as follows:

- First of all one or more nodes in a semantic net (normally the component words of the query) are identified by the literal content of a user query.
- Then the semantic similarity values of other nodes in the semantic net to these identified nodes are computed, and those semantically similar nodes are determined and returned based on the values and a threshold.

Thus, the foundation of these theories is built upon the first group of nodes in a semantic net identified by a given user query. However, as a matter of fact, some user queries are ambiguous or over-particular, which do not have corresponding nodes in a semantic net. In other words, the first group of nodes in a semantic net cannot be identified by the user queries. Without the first group of nodes, the semantically similar nodes cannot be determined and returned. As can be seen from the above argument, there is a paradox in these semantic similarity measure models that these

M. Ulieru, P. Palensky, and R. Doursat (Eds.): IT Revolutions 2008, LNICST 11, pp. 60–68, 2009.

modes could be invalid for the ambiguous or over-particular query situations. The objective of this paper is to introduce the paradox in detail.

In the following sections, first of all, we will review the literature with regards to semantic nets and the applications of semantic similarity models in information retrieval. Next, by means of a case study, we will introduce the paradox of these models' applications in information retrieval. Conclusion and future works are drawn in the final section.

2 Related Works

In the section, we briefly review the current literature with respect to semantic nets and semantic similarity models.

2.1 Semantic Nets

A semantic network (net) is a graphic notation for representing knowledge in patterns of interconnected nodes and arcs. It is a directed or undirected graph consisting of vertices, which represent concepts, and edges, which represent semantic relations between the concepts [8].

An example of a semantic network is WordNet©, a lexical database of English. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. Some of the most common semantic relations defined are meronymy (A is part of B, i.e. B has A as a part of itself), holonymy (B is part of A, i.e. A has B as a part of itself), hyponymy (or troponymy) (A is subordinate of B; A is kind of B), hypernymy (A is superordinate of B), synonymy (A denotes the same as B) and antonymy (A denotes the opposite of B).

2.2 Semantic Similarity Models

The semantic similarity models can be categorized into three main types – edge (distance)-based models, node (information content)-based models and hybrid models.

Edge (Distance)-based Models. Edge-based model is based on the shortest path between two nodes in a definitional network, which is a type of hierarchical semantic net in which all nodes are linked with is-a relations. The model is based on the assumption that all nodes are evenly distributed and of similar densities and the distance between any two nodes are equal. It also can be applied to a network structure [6].

The definition is provided by Rada, which is described below:

Let A and B be two concepts represented two nodes a and b, respectively, in an is-a semantic network. A measure of the conceptual distance between a and b is given by

$$\text{Distance}\ (A,B) = \text{minimum number of edges seperating}\ a\ \text{and}\ b \qquad (1)$$

and the similarity between a and b is given by

$$\mathrm{sim}(A,B) = 2\times\mathrm{Max}\text{ - Distance}(A,B) \tag{2}$$

where Max is maximum depth of the definitional network.

Leacock et al. [5] consider that the number of edges on the shortest path between two nodes should be normalized by the depth of a taxonomic structure [5], which are

$$\text{Distance }(A,B) = \frac{\text{minimum number of edges seperating } a \text{ and } b}{2\times\mathrm{Max}} \tag{3}$$

and the similarity between a and b is given by

$$\mathrm{sim}(A,B) = \text{-log Distance }(A,B) \tag{4}$$

The model is based on the assumption that all nodes are evenly distributed and of similar densities and the distance between any two nodes are equal. Additionally, obviously, this model only can be used in a tree-like structure.

For their applications in the field of information retrieval, a document and a query can be represented by two sets of concepts (nodes) respectively in a semantic network. Meanwhile, the query can be transformed into its Disjunctive Norm Form (DNF), which is a group of conjunctive concepts. Thus, the semantic similarity between the document and the query can be measured by computing the distance between the two set of nodes.

The limitations of the edge-based models can be concluded as follows:

- In normal taxonomic or ontological structure, the network density is not regular, which is opposite to the premise of distance-based approach (and);
- The scope of the model only limits in definitional networks, which does not consider some link-types such as part-of, antonyms and so forth.

Node (Information content)-based Model. Information content model is used to judge semantic similarity between concepts in a definitional network, based on measuring their similarity probabilities based on their information content. This model can avoid the defect of edge counting approach which cannot control variable distances in a dense definitional network [7].

The information shared by two concepts can be indicated by the concept which subsumes the two concepts in the taxonomy. Then we define

$$sim(c_1,c_2) = \max_{c\in S(c_1,c_2)}[-\log \mathrm{P}(c)] \tag{5}$$

Where, $S(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2, and P(c) is the possibility of encountering an instance of concept c.

Similar to the applications of the edge-based models in information retrieval, semantic similarity values between a document and a query also can be converted as the measure of two sets of nodes in a semantic net.

The limitations of the node-based models can be concluded as follows:

- It ignores the information that may be useful (and);
- Many synonyms may have exaggerated content value (and);

- Due to the fact information content values are calculated for synsets as opposed to individual words, it is possible for the information content value to be over-exaggerated in situations, where synsets are made up of a number of commonly occurring ambiguous words

Hybrid Model. Jiang et al. [2] developed a hybrid model that uses node-based theory to enhance the edge-based model, which also considers the factors of local density, node depth and link types [2]. The weight between a child *c* and its parent concept *p* can be measured as

$$wt(c,p)=(\beta+(1-\beta)\frac{\bar{E}}{E(p)})(\frac{d(p)+1}{d(p)})^{\alpha}(IC(c)-IC(p))T(c,p) \tag{6}$$

where $d(p)$ is the depth of node p, $E(p)$ is the number of edges in the child links, $\bar{E}$ is the average density of the whole hierarchy, $T(c, p)$ is the factor of link type, and α and β ($\alpha \geq 0$, $0 \leq \beta \leq 1$) are the control parameters of the effect of node density and node depth towards the weight.

The distance between two concepts is defined as follows:

$$\text{Distance}(c_1,c_2)=\sum_{c\in\{path(c_1,c_2)-LS(c_1,c_2)\}} wt(c,p(c)) \tag{7}$$

where $path(c_1, c_2)$ is the set that contains all the nodes in the shortest path from c_1 to c_2, and $LS(c_1, c_2)$ is the lowest concept that subsume both c_1 and c_2.

In some special cases such as only link type is considered as the factor of weight computing (α=0, β=1, and $T(c, p)$ =1), the distance algorithm can be simplified as

$$\text{Distance}(c_1,c_2)=\text{IC}(c_1)+\text{IC}(c_2)\text{ - }2\times sim(c_1,c_2) \tag{8}$$

where $\text{IC}(c_1)=\text{-log P}(c)$, and $sim(c_1,c_2)=\max_{c\in LS(c_1,c_2)}[\text{-log P}(c)]$.

Finally, the similarity value between two concepts is measured by converting the semantic distance as follows:

$$\text{Sim}(c_1,c_2)=1\text{ - Distance}(c_1,c_2) \tag{9}$$

The testing results show that the parameter α and β do not heavily influence the weight computing [2]. The application of this hybrid model in information retrieval is similar to the edge-based models.

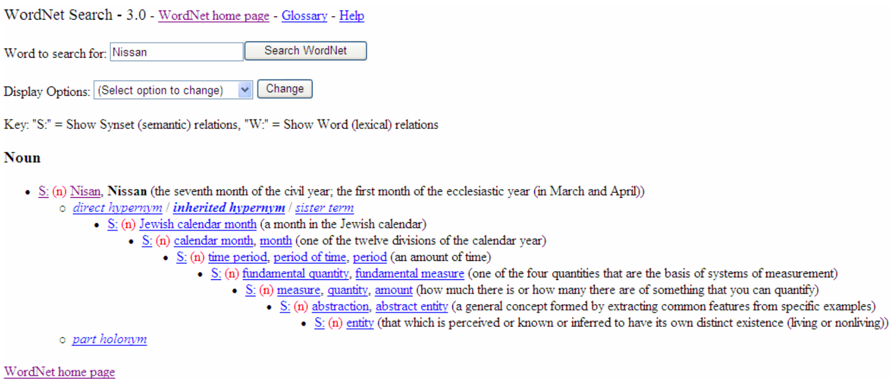
It can be observed that the distance computing between two concepts double the information content difference value between their lowest subsumer and the two concepts. However, for instance, for two high level (low information content value) nodes, their lowest subsumer's information content value may be slightly less than or equal to the nodes' values, thus their computed distance is close to zero, and they can be regarded as similar. However, there is a possibility that the two high level concepts could be hugely different. In other words, the computing result could be contradictive to the fact. Therefore, the hybrid model could meet troubles when measuring the similarity between high level nodes, which can be considered as a defect of this model. In addition, as it is the integration of edge-based and node-based models, some defects of these models also appear in the hybrid model, such as exaggerated information content values of synonyms.

3 Case Study for Analysing the Paradox

In this section, we will use two case studies to respectively introduce the invalidity of the semantic similarity models in the scenarios of query ambiguity and over-particularness situation. We choose WordNet© as the semantic net environment.

3.1 Case Study I – Query Word Sense Ambiguity

Nissan® is a well-known Japanese automobile company name, and we want to retrieve the word's meaning by WordNet© in this case study. Once we enter the query term "*Nissan*" into the WordNet© search engine, the search engine can return its glosses and synset relations. In this case, the synset relations include its direct hypernyms, inherited hypernyms and sister terms (Fig. 1). These relations are displayed as a tree-like structure where the "*Nissan*" node is the one of the tree's leaves (because there are no direct hypernyms of the node), its sister terms are the other leaves of the tree which has same joint, and its inherited hypernyms are the branches connecting it to the tree's trunk. Thus, these terms (nodes) and relations construct a semantic net together. In normal cases, the models mentioned above can be used for computing the semantic similarity values between the "*Nissan*" node and the other nodes in the semantic net.



WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: Nissan [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) Nisan, **Nissan** (the seventh month of the civil year; the first month of the ecclesiastic year (in March and April))
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) Jewish calendar month (a month in the Jewish calendar)
      - S: (n) calendar month, month (one of the twelve divisions of the calendar year)
        - S: (n) time period, period of time, period (an amount of time)
          - S: (n) fundamental quantity, fundamental measure (one of the four quantities that are the basis of systems of measurement)
            - S: (n) measure, quantity, amount (how much there is or how many there are of something that you can quantify)
              - S: (n) abstraction, abstract entity (a general concept formed by extracting common features from specific examples)
                - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
  - *part holonym*

WordNet home page

Fig. 1. Retrieved results of "*Nissan*" in WordNet© search engine

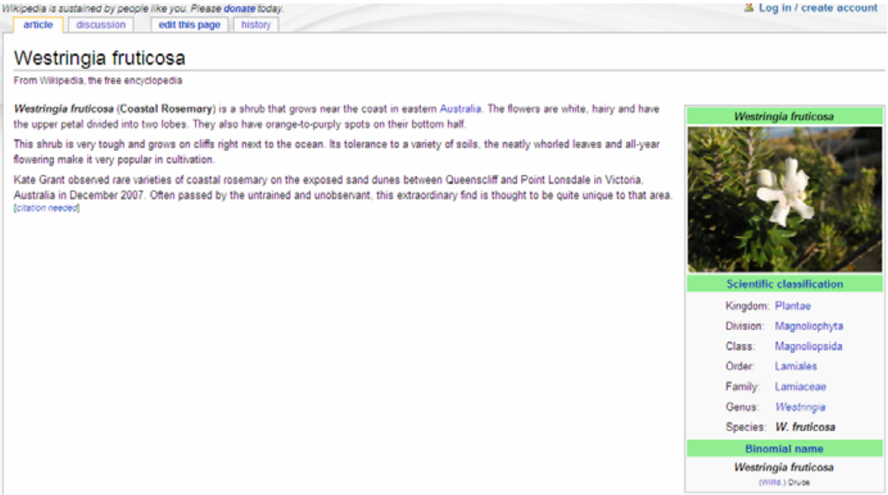
However, the returned glosses of "*Nissan*" from WordNet© indicate that it is a religious word (Fig. 1). Obviously this is not the correct meaning of the word that we want to query. Thus, the user-queried "*Nissan*" node cannot be located in the semantic net. Furthermore, those models cannot be applied for finding semantically similar nodes without locating the user-queried node.

The reason why this problem occurs is that there is an ambiguity between the word "*Nissan*" in the field of automobile and in the field of religion. Wordnet© only denotes the religious acceptation of "*Nissan*", which results in the problem of node mislocation in the semantic net.

In conclusion, this case study illustrates that semantic similarity models cannot be applied for the situation when query words are ambiguous for semantic nets.

3.2 Case Study II – Query Word Over-Particularness

(Westringia) fruiticosa is a sort of Australia's unique plant (Fig. 2). In this case study, we want to query the word "*fruiticosa*" in WordNet©. However, the return result shows that this word cannot be retrieved (Fig. 3). Similarly, the semantic similarity models cannot work in such situation. This is because the word "*fruiticosa*" is so particular that there is no record of this word stored in the WordNet© knowledge- base.


Wikipedia is sustained by people like you. Please donate today.
Log in / create account
article
discussion
edit this page
history

Westringia fruticosa

From Wikipedia, the free encyclopedia

Westringia fruticosa **(Coastal Rosemary)** is a shrub that grows near the coast in eastern Australia. The flowers are white, hairy and have the upper petal divided into two lobes. They also have orange-to-purply spots on their bottom half.

This shrub is very tough and grows on cliffs right next to the ocean. Its tolerance to a variety of soils, the neatly whorled leaves and all-year flowering make it very popular in cultivation.

Kate Grant observed rare varieties of coastal rosemary on the exposed sand dunes between Queenscliff and Point Lonsdale in Victoria, Australia in December 2007. Often passed by the untrained and unobservant, this extraordinary find is thought to be quite unique to that area. [citation needed]

Westringia fruticosa

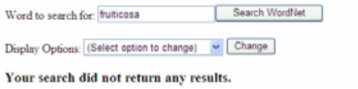
Scientific classification

Kingdom:	Plantae
Division:	Magnoliophyta
Class:	Magnoliopsida
Order:	Lamiales
Family:	Lamiaceae
Genus:	*Westringia*
Species:	***W. fruticosa***

Binomial name

Westringia fruticosa
(Willd.) Druce

Fig. 2. Interpretation of "fruiticosa" in Wikipedia®



Word to search for: fruiticosa

Display Options: (Select option to change)

Your search did not return any results.

Fig. 3. Retrieved results of "fruiticosa" in WordNet© search engine

Based on the above case studies, we analyse the paradox of semantic similarity models in two primary situations, which are query word ambiguity and over-particularness. These situations contribute to the issue that these models cannot locate the first group of nodes in a semantic net and thus cannot proceed further. There is no methodology provided by the authors of these models to solve this issue. In the next section we will provide some possible solutions to this issue.

4 Possible Solutions

In this section, we will propose possible solutions for the issue of query word sense ambiguity and over-particularness.

4.1 Possible Solutions for Query Word Sense Ambiguity

For the issue of query word sense ambiguity, the most popular approach is to use supervised machine learning approaches, which normally use pre-defined ontologies [4]. There are two possible solutions which can address this issue, as described below:

- One possible solution is to use generic ontologies, which provides common senses of generic words. One significant instance is Cyc knowledge-base, which is a general-purpose repository of common sense concepts and facts [3]. Its application – OpenCyc stores over 47,000 concepts and over 306,000 facts in its knowledge-base (www.opencyc.org).
- Another possible solution is to use domain-specific ontologies, which provide senses for domain-specific terms. One example is Gene Ontology (GO), which uses ontology to annotate gene terms. Gene Ontology database stores over 20,000 terms in the genic field (www.geneontology.org).

These ontologies can provide multiple senses for ambiguous query words in order to reduce the word sense ambiguity. By means of a question-answering module, user can choose the most appropriate sense of a query word, and thus correctly locate the first group of nodes in a semantic net.

However, this methodology has limitations described below:

- Most of these ontologies need to be manually created, which is a labour-intensive and time consuming task and may not necessarily cover the ambiguous words in all domains [1].
- There could be no node in a semantic net that can match the user-selected sense. For instance, in our first case study, although user can choose an appropriate sense of "*Nissan*" – an automobile company's name from an ontology, WordNet© cannot return the relevant result due to the fact that the sense is not stored in its knowledge-base.

4.2 Possible Solutions for Query Word Over-Particularness

For the issue of query word over-particularness, the possible solution could be to use online dictionaries to enrich semantic nets' content. There are two approaches below:

- One possible solution is to use online dictionary APIs to find synonyms for over-particular query words. Then the synonyms are matched with words in semantic nets.
- Another possible solution is to use online dictionary to manually enlarge the glossary volume of semantic nets.

Obviously the first approach is more cost-saving. However, it is found that most online dictionaries do not provide their APIs. Therefore, the first approach may not be feasible during implementation.

5 Conclusions and Further Works

In this paper, we point out and discuss the existing paradox (research issues) in the applications of the semantic similarity models in the field of information retrieval. It is observed that the process of these models involves two common steps: firstly, the nodes which correspond to a user query are located; then the models start to work and compute the semantic similarity values between these identified nodes and other nodes in a semantic net. Thus, locating nodes corresponding to the query word is a pre-requisite of these models. However, as pointed out in this paper, in certain scenarios the process of locating the corresponding nodes to the user query cannot be carried out, and thus the semantic similarity models cannot process further. In order to shed further light on this paradox, we review the three main categories of the semantic similarity models, which are edge (distance)-based models, node (information content)-based models and hybrid models. For each category of these models, we survey their applications and analyze their limitations in the field of information retrieval. Next, we use two case studies to illustrate the issues in detail – *query word sense ambiguity* and *query word over-particularness* which trouble the semantic similarity models. The two case studies are implemented in WordNet© – a typical semantic net environment. In the first case study, we use a query word which can be located but cannot be disambiguated by WordNet©. Due to this reason that the query word's sense cannot be disambiguated, as a result of which, the semantic similarity models could calculate wrong semantic similarity scores. In the second case study, we use a query word which is not stored in the WordNet© knowledge-base. These models cannot work further without node locating.

We discuss several possible solutions for the two issues. For solving the query word sense ambiguity, supervised machine leaning approaches with ontologoies are popular. For solving query word over-particularness, online dictionaries could be used to address this issue. However, by means of detailed analysis, we found that every solution has its own limitations, which cannot ultimately solve the two issues. Therefore, we assert that in order to solve this paradox further and deep research needs to be carried out in the field of semantic similarity models and semantic nets.

References

1. Andreopoulos, B., Alexopoulou, D., Schroeder, M.: Word Sense Disambiguation in Biomedical Ontologies with Term Co-occurrence Analysis and Document Clustering. Int. J. Data Mining and Bioinformatics 2, 193–215 (2008)
2. Jiang, J.J., Conrath, D.W.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: International Conference on Research in Computational Linguistics (ROCLING X), Taiwan, pp. 19–33 (1997)
3. Curtis, J.C., Baxter, D.: On the Application of the Cyc Ontology to Word Sense Disambiguation. In: The 19th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2006). AAAI Press, Melbourne Beach (2006)
4. Joshi, M., Pedersen, T., Maclin, R., Pakhomov, S.: Kernel Methods for Word Sense Disambiguation and Acronym Expansion. In: The 21st National Conference on Artificial Intelligence (AAAI 2006). AAAI, Boston (2006)
5. Leacock, C., Chodorow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification. In: WordNet: An Electronic Lexical Database, pp. 265–283. MIT Press, Cambridge (1998)
6. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and Application of a Metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19, 17–30 (1989)

7. Resnik, P.: Semantic Similarity in A Taxonomy: An Information-based Measure and Its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
8. Sowa, J.F.: Semantic Networks. In: Shapiro, S.C. (ed.) Encyclopedia of Artificial Intelligence. Wiley, Chichester (1992)