International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.8 (2014), pp.41-52
http://dx.doi.org/10.14257/ijmue.2014.9.8.04

Obtaining Information Content of Concepts: An Overview[1]

Lingling Meng[1], Runqing Huang[2] and Junzhong Gu[3]

[1]Department of Educational Information Technology, East China Normal University, Shanghai, 200062, China
[2]Shanghai Municipal People's Government, Shanghai, 200003, China
[3]Computer Science and Technology Department, East China Normal University, Shanghai, 200062, China
[1]llmeng@deit.ecnu.edu.cn, [2]runqinghuang@gmail.com, [3]jzgu@ica.stc.sh.cn

Abstract

Semantic similarity measure between concepts is a generic issue for many applications of computational linguistics and artificial intelligence. Information Content (IC) of concept is an important dimension of accessing semantic similarity between two concepts. Recently it has attracted great concern and become a hot topic. The paper gives a general overview of usage of information content in semantic similarity computing and then focuses on how to obtain information content. It reviews and analyses state-of-art IC models, including Corpus-dependent and Corpus-independent IC approach. Hyponym-based IC Model, leaves-based IC Model, concept's topology structure based IC Model and relation-based IC Model are discussed respectively in detail. The important related issues are described. Finally further research is outlined for the improvement of IC.

Keywords: *information content, hyponym-based model, leaves-based model, concept's topology based model, relation-based model*

1. Introduction

Semantic similarity measure between concepts is a generic issue for many applications of computational linguistics and artificial intelligence, such as information extraction [1-2], text segmentation [3], natural language processing [4], word sense disambiguation [5], question answering [6], recommender system [7], information retrieval [8-9]. Recently the measures based on WordNet have drawn great concern and become a hot issue. Generally speaking, the measures based on WordNet can be divided to two groups: path based measures and information content based measures. Information Content (IC) of concept is an important dimension in assessing the similarity of two concepts or two words in information content based semantic similarity measures. It provides an estimation of its general or specialty, which contributes to better understand concepts' semantic. Now IC has been successfully applied in semantic similarity computation [10-15]. Resnik first proposed an information content-based similarity measure in 1995 following information theoretic approach [12]. He assumed that similarity between two concepts depended on the extent to which they share information in common. The more information two concepts share, the more similar they are. In practice, it is indicated by the specific subsumer in the taxonomy.

$$sim_{\mathrm{Re}snik}(c_1, c_2) = -\log p(lso(c_1, c_2)) = IC(lso(c_1, c_2)) \tag{1}$$

Where, lso(c_1,c_2) is the most specific common subsumer of c_1 and c_2.

[11] The work in the paper was supported by Shanghai Industry-University Cooperation Foundation (Grant No. Shanghai CXY-2013-84) and Shanghai Scientific Development Foundation (Grant No.11530700300).

In 1998, Lin proposed another information content-based metric. He assumed that the similarity between concept c_1 and c_2 depended on not only their shared information, but also their information respectively, expressed by [13]:

$$sim_{Lin}(c_1, c_2) = \frac{2 * IC(lso(c_1, c_2)))}{IC\ (c_1) + IC\ (c_2)} \tag{2}$$

Based on Lin's method, Meng presented an improved measure, which is defined as [14]:

$$sim_{meng}(c_1, c_2) = e^{sim_{lin}} - 1 = e^{(\frac{2*IC(lso(c_1,c_2))}{IC\ (c_1)+IC\ (c_2)})} - 1 \tag{3}$$

Contrary to the above measures, in 1997 Jiang proposed a measure from different perspective by calculating semantic distance to obtain semantic similarity [15]. We get similarity by considering the opposite of the distance.

$$dis_{Jiang}(c_1, c_2) = (IC(c_1) + IC(c_2)) - 2IC(lso(c_1, c_2)) \tag{4}$$

We can see that IC is an important parameter in information content based semantic similarity measures. How to obtain IC values? Some measures have been proposed. In terms of whether to depend on corpus, all the measures can be divided in two groups: corpus-dependent IC model and corpus-independent IC model. This paper gives an overview about the different models that have been used for semantic similarity and highlights important related issues.

The structure of the paper is as follows. The measures discussed are all based on WordNet, so firstly we provide the background information regarding WordNet in Section 2. Corpus-dependent IC models are outlined in Section 3. In Section 4 we discussed the different Corpus-independent IC models respectively. The further research is described in Section 5 and a summary is given in Section 6.

2. WordNet

WordNet is the product of a research project at Princeton University [16]. It is a large lexical database of English. WordNet focuses on the word meanings instead of word forms. In WordNet Nouns, verbs, adverbs and adjectives are organized by a variety of semantic relations into synonym sets (synsets), which represent one concept. Examples of semantic relations used by WordNet are synonymy, autonomy, hyponymy, member, similar, domain and cause and so on. For example, a car is a vehicle (is-a) and keyboard is part of computer (part-of). Hyponym/hypernym (is-a) is the most common relations, which accounts for close to 80% of the relations. Words and words are interconnected via is-a relations to form a hierarchy structure, which makes it a useful tool for obtaining word sense similarity. An example of is-a relation in WordNet is shown as Figure 1. In the taxonomy the deeper concept is more specific and the upper concept is more general. Therefore C_7 is more general than $C_{16.}$ C_{16} is more general than C_{23}. C_{23} is more general than C_{32}. C_1 is the most general concept. C_{40}, $C_{41}$and C_{42} are the most general concept.

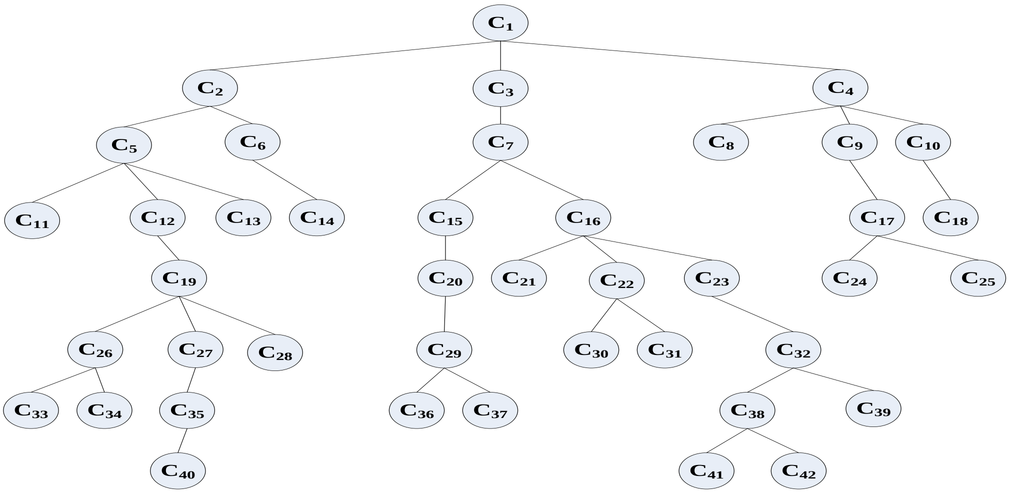


Figure 1. An Example of Is-a Relation

3. Corpus-dependent IC Model

Corpus-dependent IC model computes IC values through statistical analysis of noun in Brown corpus, which was first proposed by Resnik in 1995 [12]. It assumes that, for a concept c in the taxonomy, let p(c) be the probability of encountering and instance of concept c. According to the definition from information theory, the information content of c is expressed by:

$$IC(c) = -\log p(c) \tag{5}$$

Probability of a concept was estimated as:

$$p(c) = \frac{freq(c)}{N} \tag{6}$$

Where N is the total number of nouns, and freq(c) is the frequency of instance of concept c occurring in the taxonomy.

When computing freq(c), each noun or any of its taxonomical hyponyms that occurred in the given corpus was included. That is to say, each individual occurrence of any noun in the corpus is recursively counted as an occurrence of each of its taxonomic ancestors.

$$Freq(c) = \sum_{w \in W(c)} count(w) \tag{7}$$

Where W(c) is the set of words subsumed by concept c.

From formula (5) ~ (7), we can see that,

Firstly, IC(c) is inversely proportional to p(c). When p(c) increases, IC(c) decreases.

Secondly, the more general a concept, the more hyponyms it has, and the smaller its information content value.

Thirdly, it uses statistical methods to obtain IC value by calculating the probability of noun. Therefore word segmentation and Part-of-Speech tagging will make the work increase.

Finally, it relies on corpus analysis. IC value is directly related to corpus, and depends on the richness of the Corpus. For one concept, it will get different IC value in different corpus. If a concept is not included in the corpus, then the IC value of this concept will be assigned to 0, which is clearly unreasonable.

4. Corpus-independent IC Model

Corpus independent IC models view WordNet as a statistical resource. They assumes that concept in WordNet is numerous and rich. Recent years the models have drawn great concern. Some researchers have proposed some models, and the typical model including hyponym-based model, leaves-based IC model, concept’s topology based model and relation-based model.

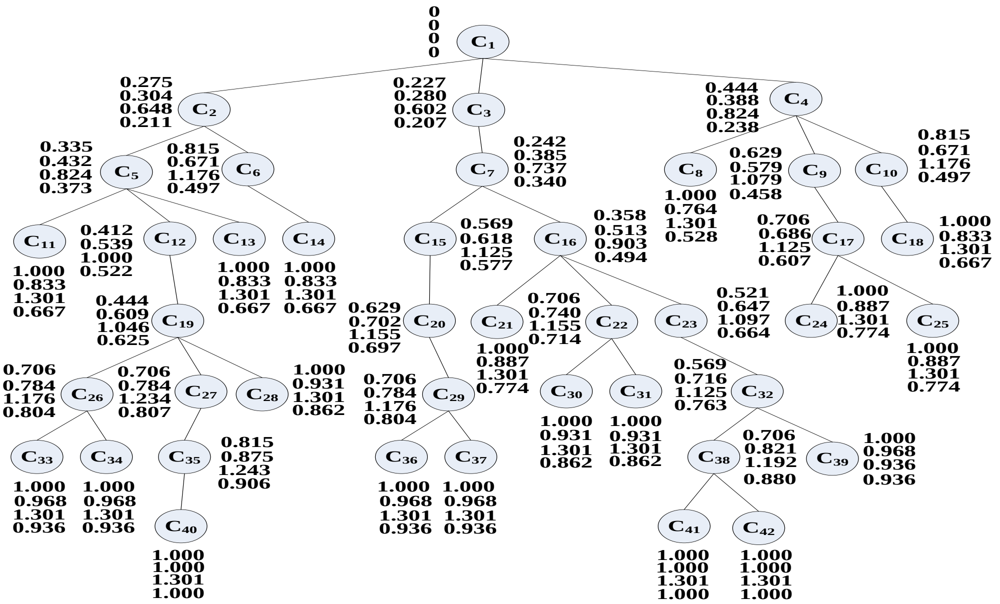
4.1. Hyponym-based IC Model

4.1.1. Nuno Model: Hyponym-based IC Model was proposed by Nuno. It assumes that a concept’s IC value depends on its hyponyms in the taxonomy of WordNet. The more hyponyms of a concept, the higher probability of the concept being encountered, and the less information it conveys. That is to say, a concept with more hyponyms expresses less information than the concepts with less ones. For a concept, the more hyponyms it has, the more general it is. Inversely, the less hyponyms it has, the more specified it is .It implies that the leaves nodes have no hyponyms and they are most specified. So the information they convey is maximal. The root node has the maximal hyponyms and it is the most general. Thus it expresses the minimal information. Therefore IC value of a concept is the function of the hyponyms it as, formally [17]:

$$IC(c) = \frac{\log(\frac{hypo(c)+1}{\max_{wn}})}{\log(\frac{1}{\max_{wn}})} = 1 - \frac{\log(hypo(c)+1)}{\log(\max_{wn})} \tag{8}$$

Where the function hypo(c) returns the number of hyponyms of a given concept c. And, $\max_{wn}$ is a constant value which is set to the maximum number of concepts that exist in the taxonomy.

Next, in order to state clearly, take Figure 1 as an example, and the IC values with different models are presented in Figure 2.


C1: 0 0 0 0
C2: 0.275 0.304 0.648 0.211
C3: 0.227 0.280 0.602 0.207
C4: 0.444 0.388 0.824 0.238
C5: 0.335 0.432 0.824 0.373
C6: 0.815 0.671 1.176 0.497
C7: 0.242 0.385 0.737 0.340
C8: 1.000 0.764 1.301 0.528
C9: 0.629 0.579 1.079 0.458
C10: 0.815 0.671 1.176 0.497
C11: 1.000 0.833 1.301 0.667
C12: 0.412 0.539 1.000 0.522
C13: 1.000 0.833 1.301 0.667
C14: 1.000 0.833 1.301 0.667
C15: 0.569 0.618 1.125 0.577
C16: 0.358 0.513 0.903 0.494
C17: 0.706 0.686 1.125 0.607
C18: 1.000 0.833 1.301 0.667
C19: 0.444 0.609 1.046 0.625
C20: 0.629 0.702 1.155 0.697
C21: 1.000 0.887 1.301 0.774
C22: 0.706 0.740 1.155 0.714
C23: 0.521 0.647 1.097 0.664
C24: 1.000 0.887 1.301 0.774
C25: 1.000 0.887 1.301 0.774
C26: 0.706 0.784 1.176 0.804
C27: 0.706 0.784 1.234 0.807
C28: 1.000 0.931 1.301 0.862
C29: 0.706 0.784 1.176 0.804
C30: 1.000 0.931 1.301 0.862
C31: 1.000 0.931 1.301 0.862
C32: 0.569 0.716 1.125 0.763
C33: 1.000 0.968 1.301 0.936
C34: 1.000 0.968 1.301 0.936
C35: 0.815 0.875 1.243 0.906
C36: 1.000 0.968 1.301 0.936
C37: 1.000 0.968 1.301 0.936
C38: 0.706 0.821 1.192 0.880
C39: 1.000 0.968 0.936 0.936
C40: 1.000 1.000 1.301 1.000
C41: 1.000 1.000 1.301 1.000
C42: 1.000 1.000 1.301 1.000

Line 1: resulted with hyponym-based IC metric(Nuno model)
Line 2: resulted with hyponym-based IC metric(zhou model)
Line 3: resulted with leaves-based IC metric
Line 4: resulted with concept’ s topology structure IC metric (Meng model)

Figure 2. IC Values based on Different Models

From formula (8) and Figure 2, it is noted that,

Firstly, its main advantage is that IC does not rely on corpus analysis, and avoid the sparse data problem.

Secondly, IC is inversely proportional to the number of hyponyms that a concept has。

Thirdly, if a concept is root node in the taxonomy, it has the maximum number of hyponyms. For example in Figure 2, C_1 is root. hypo（C_1） is 41. Thus:

$$IC(C_1)=1-\log(41+1)/\log 42=1-1=0.$$

Fourthly, if a concept is a leaf node in the taxonomy, it has the minimum number of hyponyms. In Figure 2, C_{40} is a leaf node, hypo（C_{40}） is 0. Thus:

$$IC(C_{40})=1-\log(0+1)/\log 42 =1-0=1。$$

Beside this, all leaves will have the same IC values.

Fifthly, the IC values are range from 0 to1.

Finally, hypo(C_4) and hypo(C_{19}) have the same value of 7, which resulted IC(C_4) and IC(C_{19}) have the same IC value of 0.444. This is because that the model only take hyponyms into considerate. A directly result is that the concepts with the same number of hyponyms will have the same IC values. But the depth of concepts in the taxonomy is different and the deeper one is more specific. For example in Figure 2, deep(C_{19}) is 5 and deep(C_4) is 2. C_{19} should convey more information that C_4, and have the bigger IC value.

4.1.2. Zhou Model: Zhou takes the depth of each concept into account and presents an improved model, expressed by [18]:

$$IC(c)=k(1-\frac{\log(hypo(c)+1)}{\log(node_{\max})})+(1-k)(\frac{\log(deep(c))}{\log(deep_{max})}) \tag{9}$$

Where $node_{max}$ is the maximum number of concepts that exist in the taxonomy, deep(c) returns the depth of concept c in the taxonomy, $deep_{max}$ is the max depth of the taxonomy, k is a changeable factor so as to adjust the weight of the two items. In his experiment, k is 0.5.

From formula (9) and Figure 2, it is noted that:

Firstly, Zhou model solves the problem that generated in Nuno model. If two concepts have the same number of hyponyms, the deeper one in the taxonomy will have the bigger IC value.

Take Figure 2 as an example,

$$IC(C_4)=0.5*(1-\log(7+1)/\log 42)+(1-0.5)*(\log 2/\log 8)= 0.388$$
$$IC(C_{19})=0.5*(1-\log(7+1)/\log 42)+(1-0.5)*(\log 5/\log 8)= 0.609$$

In spite of hypo(C_4) is equal to hypo(C_{19}), however deep(C_{19}) is 5 and deep(C_4) is 2, and deep(C_{19}) is deeper than deep(C_4).Hence IC(C_{19}) > IC(C_4)。

Secondly, root node has the maximum number of hyponyms in the taxonomy and its IC value is 0. If a leaf is the deepest node in the taxonomy, its IC value is 1. The IC values are range from 0 to1.

Thirdly, if a leaf is not the deepest node in the taxonomy, its IC value is proportional to its depth. Take C_{37} as an example, hypo(C_{37}) is 0.

$$IC(C_{37})=0.5*(1-\log(0+1)/\log 42)+(1-0.5)*\log 7/\log 8 =0.5+0.5*\log 7/\log 8=0.968。$$

Fourthly, if two concepts with the same number of hyponyms have the same depth in the taxonomy, they IC values will be equal. For example,

$$hypo(C_{26})= hypo(C_{27})=2 \text{ and } deep(C_{27})= deep(C_{27})=6,$$

Then,

$$IC(C_{26})= IC(C_{27})= 0.5*(1-\log(2+1)/\log 42)+(1-0.5)*(\log 6/\log 8)= 0.784$$

Finally, although two concepts have the same number of hyponyms or leaves, their hyponyms are arranged in a different way. Take C_{26}, C_{27} as an example, the hyponyms of C_{26} and C_{27} are 2. The two hyponyms of C_{26} are arranged as siblings side by side. However the two hyponyms of C_{27} are arranged in a line. Therefore, C_{26} and C_{27} should convey different information.

4.2. Concepts' Topology-based IC Model

Concepts' Topology-based IC Model takes each concept's topology into account, such as the concept's depth, hyponyms, sibling concepts, and so on.

4.2.1. Sebti model: Sebit model is based on the assumption that taxonomic structure of WordNet is organized in a meaningful and principled way, where concepts in higher depths and having more sibling concepts in the taxonomy structure are more informative and their values are bigger. Figure 3 represents this method for computing IC value for a fragment of concepts in WordNet [19].

According to Figure 3, we can see that Entity has nine sibling concepts, and then the information content of Entity can be calculated by:

$$IC(Entity) = -\log(\frac{1}{9}) = 2.1972$$

According to the same rule:

$$IC(Object) = -\log(\frac{1}{9} * \frac{1}{10}) = 4.4998$$

$$IC(Aritifact) = -\log(\frac{1}{9} * \frac{1}{10} * \frac{1}{36}) = 8.0833$$

$$IC(Instrumentality) = -\log(\frac{1}{9} * \frac{1}{10} * \frac{1}{36} * \frac{1}{42}) = 18.8210$$

$$IC(Covering) = -\log(\frac{1}{9} * \frac{1}{10} * \frac{1}{36} * \frac{1}{42}) = 18.8210$$

$$IC(Container) = -\log(\frac{1}{9} * \frac{1}{10} * \frac{1}{36} * \frac{1}{42} * \frac{1}{13}) = 14.3859$$

$$IC(Clothing) = -\log(\frac{1}{9} * \frac{1}{10} * \frac{1}{36} * \frac{1}{42} * \frac{1}{34}) = 15.3474$$

$$IC(Box) = -\log(\frac{1}{9} * \frac{1}{10} * \frac{1}{36} * \frac{1}{42} * \frac{1}{13} * \frac{1}{49}) = 18.2778$$

$$IC(Dress) = -\log(\frac{1}{9} * \frac{1}{10} * \frac{1}{36} * \frac{1}{42} * \frac{1}{34} * \frac{1}{31}) = 18.7813$$

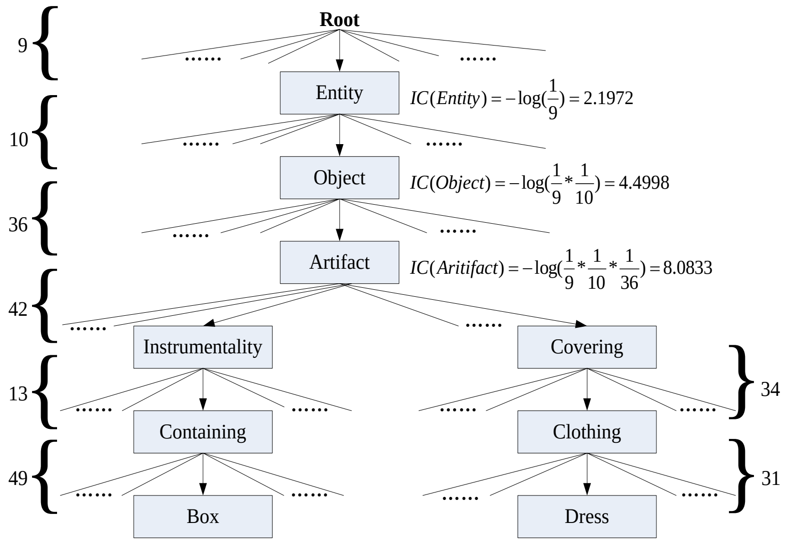

Root
9
......
......
Entity
$IC(Entity)=-\log(\frac{1}{9})=2.1972$
10
Object
$IC(Object)=-\log(\frac{1}{9}*\frac{1}{10})=4.4998$
36
Artifact
$IC(Aritifact)=-\log(\frac{1}{9}*\frac{1}{10}*\frac{1}{36})=8.0833$
42
Instrumentality
Covering
13
34
Containing
Clothing
49
31
Box
Dress

Figure 3. Example of Computing IC Values for a Fragment of Concepts

It can be seen that,

Firstly, for two concepts in the same depth, the more sibling concepts, the bigger IC values. For example, Clothing has 34 sibling concepts and Containing has 13 sibling concepts. Therefore, IC(Clothing) > IC(Containing).

Secondly, the IC values will become bigger with the increasing of depth.

Finally, the sibling concepts will have the same IC values. For example, Instrumentality and Covering are sibling concepts and IC(Instrumentality) = IC(Covering) = 18.8210.

4.2.2. Meng Model: Meng model argues that each concept is unique in the taxonomy. For a given concept c, IC is the function of itself and its hyponyms' arrangements, including concept's depth, the number of its hyponyms, and the depth of each hyponym. It is defined as [20]:

$$IC(c)=\frac{\log(deep(c))}{\log(deep\_\max)}*(1-\frac{\log(\sum_{a\in hypo(c)}\frac{1}{deep(a)}+1)}{\log(node\_\max)}) \tag{10}$$

For a given concept c, where deep(c) denotes the depth of concept c in the taxonomy, deep_max presents the max depth of the taxonomy; a is a concept in the taxonomy, which satisfies a∈hypo(c); node_max is the maximum number of concepts that exist in the taxonomy.

If c is root, deep (root) is 1.Then log (deep(c)) =log (1) =0.

If c is a leaf, hypo(c) is 0. Then,

$$\sum_{a\in hypo(c)}\frac{1}{deep(a)}=0 \tag{11}$$

And,

$$IC(c)=\frac{\log(deep(c))}{\log(deep\_\max)}) \tag{12}$$

From formula (10) (11) (12) and Figure 2, we can see that,

Firstly, IC (root) is 0. If a leaf is the deepest node in the taxonomy, its IC value is 1.

Secondly, The IC values are range from 0 to1.

Thirdly, each concept's IC value is unique. Even if two concepts with the same number of hyponyms are in the same depth in the taxonomy, they will have different IC values. Take C_{27} and C_{29} as an example, IC (C_{27}) is bigger than IC (C_{29}), because deep (C_{35}) is equal to deep (C_{36}), and deep (C_{40}) is bigger than deep (C_{37}). If two concepts have the same number of leaves and subsumers, their IC values are different too. For example, both C_{17} and C_{15} have two leaves, but hypo (C_{15}) is 2 and hypo (C_{15}) is 4. Therefore IC (C_{17})> IC (C_{15}).

Thirdly, leaves in different depth will convey different information. For a specific version of WordNet, deep_max is a fixed value. The deeper of a leaf, the more information it expresses.

4.3. Leaves-based IC Model

Leaves-based IC Model was proposed by David in 2011. It starts from number of leaves of concepts to calculate the IC value. Leaves-based IC model assumes that taxonomical leaves represent the semantic of the most specific concepts of a domain, and they are enough to differentiate one concept from other ones, regardless the amount of inner concepts incorporated in the taxonomy [21]. Leaves-based IC model argues that the more leaves a concept has the less information it expresses. That is to say, a concept with more leaves is more general. Besides this, the depth of concept in taxonomy has been taken into account. Here, depth is instead by the number of subsumers from a different view. Formally [21],

$$IC(c) = -\log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{\max\_leaves+1}\right) \tag{13}$$

Where, let C be the set of concepts of the ontology, for a given concept c,

leaves(c)={l∈C|l∈hyponyms(c) ∧ l is a leaf}.

Max_leaves represents the number of leaves corresponding to the root node of the hierarchy. subsumers(c) returns the set of subsumers.

Subsumers(c)={a∈C | c≤a }∪{c}, c≤a means that c is a hierarchical specialization of a.

From formula (13) and Figure 2, it is noted that:

Firstly, for a given version of WordNet, max_leaves is a fixed value. In Figure 2, leaves_max=19.

Secondly, concept's IC is inversely proportional to the amount of leavers. The more leaves a concept contains, the smaller of its IC is. The less leaves a concept contains, the bigger of its IC is.

Thirdly, concept's IC is directly proportional to its number of taxonomical subsumers.

Finally, if two concepts have the same number of leaves and subsumers, they will have same IC values. Take C_{15} and C_{17} as an example,

|subsumers (C_{15})| = |subsumers (C_{17})| = 4, and |leaves (C_{15}) | = |leaves (C_{17}) |= 2.

Therefore,

$$IC(C_{15}) = IC(C_{17}) = 1.125.$$

However, hypo (C_{15}) is 4 and hypo (C_{17}) is 2. IC (C_{15}) and IC (C_{17}) should convey different information.

4.4. Relation-based IC Model

Relation-based IC Model was proposed by Md. Hanif Seddiqui in 2010. It is based on the assumption that every concept is defined with sufficient semantic embedding with the organization, property functions, property restrictions and other logical assertions. Relation based IC model is defined as [22]:

$$IC(c) = \rho \cdot IC_{rel}(c) + (1-\rho) \cdot IC_{nuno}(c)) \tag{14}$$

$$IC_{rel}(c) = \frac{\log(rel(c)+1)}{\log(total\_rel+1)} \tag{15}$$

$$\rho = \frac{\log(total\_rel+1)}{\log(total\_rel) + \log(total\_concept)} \tag{16}$$

Where rel(c) is the number of relations of concept c. And total_rel denotes the total number of relations, while total_concept represents the maximum of concepts in the ontology. IC_{nuno}(c) is defined with formula (8).

From formula (14) (15) (16), it is noted that:

Firstly, the model took concept, properties and their relations into account.

Secondly, IC_{rel}(c) is proportional to the number of properties it is related to, the more rel(c), the higher IC_{rel}(c).

Besides these, it can be applied in not only a simple taxonomy, but also a complex ontology with concept-properties relations.

Different IC models above is defined from different views, table 1 presents the characteristic respectively.

Table 1. Comparison of Different IC Models

Whether to consider the factors	characteristic and the result						
	Corpus dependent IC Model	Corpus independent IC Model					
	Resnik model	Hyponym-based IC Model		Leaves-based IC Model	Concepts’ topology-based IC Model		Relation-based IC Model
		Nuno model	Zhou model	David model	Meng model	Sebti model	Md. Hanif model
Whether to rely on corpus	Yes	No	No	No	No	No	No
Sparse data problem	Yes	No	No	No	No	No	No
Depth(c) increase	No	No	Yes, IC increase	Yes IC increase	Yes, implicitly IC increase	Yes, IC increase	Yes, implicitly IC increase
Hypo(c) increase	No	Yes, IC decrease	Yes, IC decrease	No	Yes IC decrease	No	Yes IC decrease
Leaves(c) increase	No	No	No	Yes IC decrease	No	No	No
Rel(c) increase	No	No	No	No	No	No	Yes IC increase
Concept’s topology	No	No	No	No	Yes	Yes	No
Sibling concepts increase	No	No	No	No	No	Yes, IC increase	No

5. Discussion and Further Research

IC plays an important role in semantic similarity computing. It is necessary to obtain highly effective IC model. Many researchers proposed some IC models from different view. On the whole, the IC models can be grouped into two categories: Corpus-dependent IC model and Corpus-independent IC model. From the discussion above, we noted that there is sparse data problem in Corpus-dependent IC model. Corpus-independent IC model still is a hot topic. Hyponym-based IC model, leaves-based IC model, concepts' topology-based IC model and relation-based IC model are all Corpus-independent IC model. In hyponym-based IC model, depth is not been taken into considered, which directly results that the concepts with the same number of hyponyms will have the same IC values. Zhou makes some improvements. However, in Zhou's work if two concepts have the same depth and hypo(c), their IC will be equal. Leaves-based IC model argues that leaves are enough to describe and differentiate the concept from any other one, regardless of the inner-detail of the hierarchy. But it is noted that concepts with the same leaves and subsumers will have the same IC value too. Concepts' Topology-based IC Model takes each concept's topology into account, such as the concept's depth, hyponyms, sibling concepts, and so on. Sebti model is based on the assumption that concepts in higher depths and having more sibling concepts in the taxonomy structure are more informative and their values are bigger. But we noticed that the sibling concepts will have the same IC values. Meng model assumes that each concept is unique. IC is the function of itself and its hyponyms' arrangements, including concept's depth, the number of its hyponyms, and the depth of each hyponym. Relation-based IC model take concepts' hyponyms, property, restrictions and other logical assertions into account. Part of the model comes from Nuno model.

Our aim is to effectively distinguish different concepts and make us obtain the most accurate IC value so that the semantic similarity between words can be more appropriate description. In further work, it may be a good idea to take link type included. Besides these, if we design IC model according information entropy, maybe we will have an unexpected discovery.

6. Summary

This paper gives an overview about various information content computing methods. Corpora-dependent IC model and Corpora-independent IC model are illustrated. Hyponym-based model, leaves-based model, concept's topology structure based model and relation-based model are discussed in detail. The important related issues are presented. What's more, a comparative study has been done. The result is shown in table 1. Finally the paper gives some suggestions of the area in further research.

References

[1] J. Atkinson, A. Ferreira and E. Aravena, "Discovering implicit intention-level knowledge from natural-language texts", Knowl.-Based Syst, vol. 22, no. 7, (**2009**).

[2] M. Stevenson and M. A. Greenwood, "A semantic approach to IE pattern induction", Proceedings of 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan, USA, (**2005**) June 25-30.

[3] H. Kozima, "Computing Lexical Cohesion as a Tool for Text Analysis," doctoral thesis, Computer Science and Information Math., Graduate School of Electro-Comm., Univ. of Electro- Comm., 1994.

[4] Y. Li, Z. A. Bandar and D. McLean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", IEEE Transactions on Knowledge and Data Engineering, vol. 4, no. 15, (**2003**).

[5] S. Patwardhan, S. Banerjee and T. Pedersen, "Using measures of semantic relatedness for word sense disambiguation", Proceedings of 4th International Conference on Computational Linguistics and Intelligent Text Processing and Computational Linguistics, (**2003**) February 16-22, Mexico City, Mexico.

[6] A. G. Tapeh and M. Rahgozar, "A knowledge-based question answering system for B2C eCommerce", Knowl.-Based Syst., vol. 21, no. 8, (**2008**).

[7] Y. Blanco-Fernández, J. J. Pazos-Arias, A. Gil-Solla, M. Ramos-Cabrer, M. López- Nores, J. García-Duque, A. Fernández-Vilas, R.P. Díaz-Redondo and J. Bermejo-Muñoz, "A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems", Knowl.-Based Syst., vol. 21, no. 4, (**2008**).

[8] R. K. Srihari, Z. F. Zhang and A. B. Rao, "Intelligent indexing and semantic retrieval of multimodal documents", Information Retrieval, vol. 2, (**2000**).

[9] 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, (**1989**).

[10] A. Formica, "M. S: Concept similarity in formal concept analysis: an information content approach", Knowl.-Based Syst., vol. 21, no. 1, (**2007**).

[11] G. Pirró, "M. S: A semantic similarity metric combining features and intrinsic information content", Data Knowl. Eng, vol. 68, no. 11, (**2009**).

[12] P. Resnik, "Using information content to evaluate semantic similarity", Proceedings of the 14th International Joint Conference on Artificial Intelligence (1995) August 20-25, Montréal Québec, Canada.

[13] D. Lin, "An information-theoretic definition of similarity", Proceedings of the 15th International Conference on Machine Learning, (1998) July 24-27, Madison, Wisconsin, USA.

[14] L. Meng and J. Gu, "A New Method for Calculating Word Sense Similarity in WordNet", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 3, (**2012**).

[15] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", Proceedings of International Conference on Research in Computational Linguistics, (**1997**) August 22-24, Taipei, Taiwan.

[16] C. Fellbaum, "WordNet: An electronic lexical database", Language, Speech, and Communication, MIT Press, Cambridge, USA, (**1998**).

[17] N. Seco, T. Veale and J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet", Proceedings of the 16th European Conference on Artificial Intelligence, (**2004**) August 22-27, Valencia, Spain.

[18] Z. Zhou, Y. Wang and J. Gu, "New Model of Semantic Similarity Measuring in Wordnet", Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering, (**2008**) November 17-19, Xiamen, China.

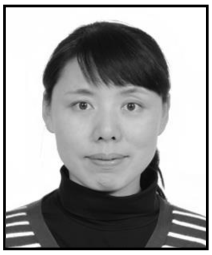
[19] A. Sebti and A. A. Barfrouch, "A new word sense similarity measure in WordNet", Proceedings of the International Multiconference on Computer Science and Information Technology, (2008) October 20–22, Wisa, Poland.

[20] L. Meng, J. Gu and Z. Zhou, "A New Model of Information Content Based on Concept's Topology for Measuring Semantic Similarity in WordNet", International Journal of Grid and Distributed Computing, vol. 3, (**2012**).

[21] D. Sánchez, M. Batet and David Isern, "Ontology-based information content computation", Knowl.-Based Syst., vol. 24, no. 2, (**2011**).

[22] Md.Hanif Seddiqui and M. Aono, "Metric of intrinsic information content for measuring semantic similarity in an ontology", Proceedings of 7th Asia-Pacific Conference on Conceptual Modeling, (**2010**) January 18-21, Brisbane, Australia.

Authors



Lingling Meng, she is an associate professor of Department of Educational Information Technology in East China Normal University. Her research interests include intelligent information retrieval, ontology construction and knowledge engineering.

Runqing Huang, he has a Ph.D. from Shanghai Jiao Tong University. He works in Shanghai Municipal People's Government, P. R. China. His present research interests include modeling strategic decisions, economic analysis, electronic government and Logistics.

Junzhong Gu, he is Supervisor of PhD Candidates, full professor of East China Normal University, head of Institute of Computer Applications and director of Lab, Director of Multimedia Information Technology (MMIT). His research interests include information retrieval, knowledge engineering, context aware computing, and data mining.