IGI PUBLISHING

ITJ4173

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

This paper appears in the publication, **International Journal of Information Technology and Web Engineering**, Volume 3, Issue 1
edited by **Ghazi I. Alkhatib © 2008, IGI Global**

The Impact of Ontology on the Performance of Information Retrieval:

A Case of WordNet

Maria Indrawan, Monash University, Australia

Seng Loke, La Trobe University, Australia

ABSTRACT

The debate on the effectiveness of ontology in solving semantic problems has increased recently in many domains of information technology. One side of the debate accepts the inclusion of ontology as a suitable solution. The other side of the debate argues that ontology is far from an ideal solution to the semantic problem. This article explores this debate in the area of information retrieval. Several past approaches were explored and a new approach was investigated to test the effectiveness of a generic ontology such as WordNet in improving the performance of information retrieval systems. The test and the analysis of the experiments suggest that WordNet is far from the ideal solution in solving semantic problems in the information retrieval. However, several observations have been made and reported in this article that allow research in ontology for the information retrieval to move towards the right direction.

Keywords: *ontology; semantic information retrieval; WordNet*

INTRODUCTION

Semantic understanding is crucial to the success of many information technology applications. Much information technology research is still battling to solve the problem of semantic understanding for their research domain. Ontology adoption is currently the most popular approach taken by many researchers. The proliferation in the use of ontology to support semantic analysis has been found in many domains of information technology such as context awareness (Rack, Arbanowski, & Steglich, 2000; Yan & Li, 2006), service oriented computing (Bramantoro, Krishnaswamy, & Indrawan, 2005; Jingshan, Hunhns, 2006), and Semantic Web (Caliusco, Galli, & Chiotti, 2005; Dou, LePendu, Kim, & Qi, 2006). Some of the researchers adopt a specific built ontology whereas others investigate the use of a general purpose ontology, such as WordNet.

WordNet is an English lexical referencing system built in the early 1990s at Princeton University. Since its introduction, many researchers have used this lexical system for different purposes, such as multimedia retrieval (Benitez, Chang, & Smith; 2001), text summarization (Hachey & Grover, 2004), and automatic creation of domain-based ontology (Chen, Alahakoon, & Indrawan, 2005; Khan & Luo, 2002). In information retrieval research, the impact of WordNet has been investigated by a number of researchers. WordNet has been used to improve the performance of information retrieval systems by way of query expansion (Voorhees, 1993), semantic distance measure (Richrdson & Smeaton, 1995), and semantic indexing (Wang & Brookes, 2004) to name a few. The results showed by these studied are varied. Voorhees (1993) and Richrardson and Smeaton (1995) report that the recall and precision of the retrieval decreased with the inclusion of WordNet. Wang and Brookes (2004), on the other hand, report the opposite. We were encouraged by Wang and Brookes' report and decided to investigate further since we perceived a further improvement can be applied to their model. In addition, we also would like to explore the debate over the impact of WordNet in information retrieval researches. At the end of the investigation we would like to enrich the debate by reporting our experience and observations during the investigation. In order to achieve this, we organize this article as follows. In the next section, the article presents a short description of WordNet for those readers unfamiliar with this lexical system. In the third section, we lay out the current debate on the impact of WordNet in information retrieval. We introduce our improvement to Wang and Brookes' model in the forth section. The following section presents the experiment design and results. We conclude our discussion in the last section.

WordNet

The main construct of WordNet as a lexical system is the synonym set or synset. The synsets are divided into four major speech categories of noun, verb, adjective, and adverb. Within each of these categories, several semantic relations between synsets are defined. Included in the noun category are the *hypernym, hyponym, meronym,* and *holonym*.

Definition 1: *Semantic Relations of Synsets*

Let assume synsets $S=\{s_{i,} s_{j},\ldots,s_{n}\}$ and $L=\{l_{i}, l_{j},\ldots,l_{n}\}$ exist in the WordNet.

Hypernym: *S* is considered to be hypernym of *L*, if every *L* is a (kind-of) *S*.

Hyponym: *S* is considered to be a hyponym of *L*, if every *S* is a (kind-of) *L*.

Meronym: *S* is considered to be a meronym of *L*, if every *S* is a part-of *L*.

Holonym: *S* is considered to be a holonym of *L*, if every *L* is a part-of *S*.

□

In an example of taxonomy of synsets depicted in Figure 1, a *canine* is considered to be a hyponym of *carnivore* and a hypernym of *dog*.

Figure 2 shows that an *engine* is a meronym of a *car* and a holonym of a *piston*. The hypernym/hyponym relations are often referred as *hyponomic* relations. The meronym/holonym relations are referred as *part-whole* relations.

Figure 1. A hypernym/hyponym relations

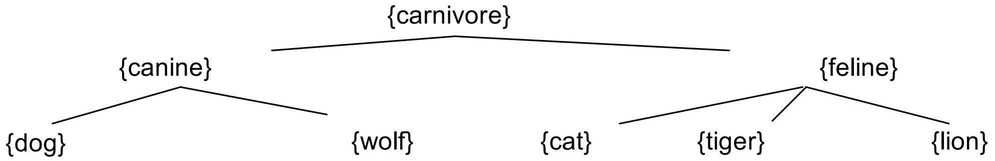
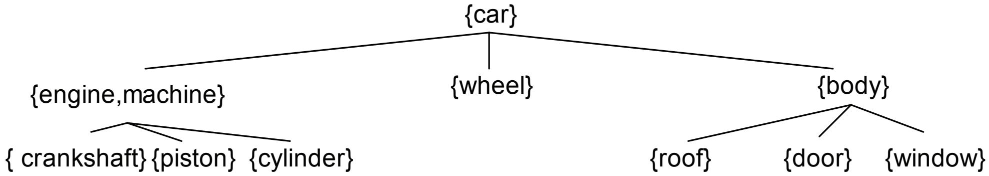


Figure 2. Holonym/meronym relations



WORDNET FOR INFORMATION RETRIEVAL SYSTEMS

WordNet has been used in several capacities to improve the performance of information retrieval systems. In this section, we explore the research problems in information retrieval and a WordNet-based model that was proposed to solve the problems. First, we present problems associated with information retrieval systems and possible techniques that may solve the problems. Subsequently, we present how the possible techniques can be developed using WordNet.

Information Retrieval Systems Limitations

Information retrieval systems are built to help users find the documents that are relevant to their information need. The information need is represented by queries posed to the system. There are a number of drawbacks inherent in the information retrieval systems. In this section, we present some of these problems and review the role of ontology in resolving these problems.

In most information retrieval systems, the queries are represented as a list of keywords. Some systems may allow users to submit a natural language query; however, in most cases, the natural language query is then processed through an indexing process resulting in a set of indexed terms. Using keywords or indexed terms for the retrieval limits the set of retrieved documents to those with the matching terms. It showed the drawback of the approach because it is possible that a document that does not contain any matching term to the query to be relevant. It is possible that the document uses synonyms of the indexed term. To avoid this problem, information retrieval researches suggest adding similar terms such as synonyms or other terms that are relevant to the document, such as hypernyms. This method is called query expansion.

Performance of information retrieval is usually measured by the precision or accuracy and recall or coverage. The recall may be improved by the query expansion as suggested earlier. To improve the precision, information retrieval needs to deduce the context of the terms used in the query since it is possible that one word has many meaning depending on the context. For example, the word *window* can mean an architectural part of a building or a name of an operating system. Without knowing the precise *sense* of the term *window* in the query, the systems will retrieve both sets of documents on *buildings* and *operating systems* while the users usually want only one set. This drawback leads to poor precision. To solve this problem, *sense disambiguation* can be employed.

Information retrieval task can be considered as finding documents that have the shortest *semantic distance* to the query. A number of distance measures have been developed, such as *cosine similarity* in vector space model. In this model, the matching of the terms in the documents and query is calculated based on finding the matching terms and the frequency analysis of the importance of the terms in discriminating the documents (*tf*idf weighting)*. It can be perceived that the semantics of the terms is given mainly by its frequency analysis rather than *linguistic* analysis. There are a number of *semantic distance* measurements that have been developed in the natural language researches

(Budanitsky, 1999). It is possible that adopting the semantic distance that is based on linguistic instead of frequency analysis may lead to a better search because it may partially solve the two previous drawbacks.

Most information retrieval systems base there operation on word- or term-level matching. There are attempts in using phrases with aim to improve precision. Similar to word matching, phrases matching can be improved by query expansion and sense disambiguation.

Having presented the problem inherent in the information retrieval systems, the next sections present the discussion on the ontological approaches that have been investigated in order to eliminate the problems.

Query Expansion

Voorhees (1994) suggests that WordNet can be used in information retrieval for query expansion. Query expansion is considered to be one of the techniques that can be used to improve the retrieval performance of short queries. Most of the indexing and retrieval methods are based on statistical methods; short queries posed challenges to this model due to the limited amount of information that can be gathered during its processing.

In expanding the query, Voorhees suggests the used of synonyms, hypernyms, hyponyms, and their combinations. The results showed that the used of synonyms, hypernyms, and hyponyms are significant in the retrieval performance for short queries, but little improvement when they are applied to the long query.

A more comprehensive study was conducted by Smeaton and Berrut (1995). They incorporated both the word sense disambiguation and query expansion into the information retrieval. However, they reported a decrease of performance instead of improvement.

The shortcoming of WordNet in supporting query expansion is also reported by Mandala, Tokunaga, and Tanaka (1999, 1998). To overcome the limitation, additional thesauri are added to complement WordNet. The results showed that the combined thesauri produced much higher recall and precision in comparison to using WordNet alone. This observation is supported by Hsu, Tsai, and Chen (2006). Their findings suggest that WordNet can improve the performance of the retrieval if the specific terms that represent the topic are found in WordNet. They called these terms as kernel words.

The impact of WordNet in query expansion shows more prominent results in the phrase-based retrieval as reported by Liu, Liu, Yu, and Meng (2004). The works previously discussed in this section base their retrieval on term or word-level. In the word-level retrieval, the query expansion based solely on WordNet does not give substantial improvement in the precision and recall.

Word Senses Disambiguation

A word may have a number of different meanings depending on its context. For example the word *arms* could mean *limbs* or *weaponry*. In WordNet, the word context is called *sense*. In most information retrieval systems, the *sense* of the words or terms is ignored during the retrieval. Researchers adopted this approach with the premise that ignoring the senses may broaden the search results which in turn will improve the recall. The increase of the recall in some cases may lead to the decrease in precision. In view to disambiguate senses for information retrieval, Voorhees (1993) proposed the use of WordNet. The proposed model used the *hyponomic* relations to determine the sense of an indexed word. Each of the indexed word were assigned a weight (*tf*idf* weight) and a sense type. During the retrieval, an indexed document term and a query term were considered similar only if they belonged to the same sense type. Hence, two terms that were usually considered to be similar in many retrieval systems may be considered to be different in this model. In other words, Voorhees's model may return relevant documents higher in the raking, that is, better precision, compared to the system without sense disambiguation. However, at the same time, it may not necessarily find all the relevant documents because the similarity matching is done with one extra condition, which is the sense type.

Indeed, their results showed that the simple stem-based retrieval without any sense disambiguation performed better than sense-based retrieval, in particular for short queries. The observations showed that short queries have very limited context information and can lead to an incorrect sense resolution. They observed that making incorrect sense resolution created deleterious affect on performance compared to making spurious matches. Hence the cost of improving the precision is not justified by the amount of degradation in the recall.

Moldovan and Milhacea (2000) propose a combination of query expansion and sense disambiguation for the Internet search. They report positive findings. However, it is difficult to directly compare their results to those of Voorhees (1994) due to the different nature of the query and data used in the experiments. Moldovan and Milhacea's sample queries are of the question and answer (Q&A) nature where users expect a specific answer. Hence the nature of *relevance* in their experiment would be different from the typical information retrieval systems. In an information retrieval system that is not Q&A, the relevance is perceived as something very broad rather than specific.

Semantic Distance

Richardson and Smeaton (1995) propose the use of *word semantic distance* in measuring similarity between documents and queries. As previously stated, most of the indexing techniques adopted by the information retrieval systems are based on the frequency analysis of the word occurrence in the document and collection. Hence, the distance between two terms are measured based on this frequency analysis rather than their actual semantic analysis. Unlike traditional systems, Richardson and Smeaton's system base the similarity measure on the word's semantic distance. To do this, they created the system as a controlled vocabulary system. WordNet was used a knowledge based containing the controlled vocabulary. Their experiments showed that the performance of their model was worst than the traditional systems based on the frequency analysis.

Semantic Indexing

In this section, we present another approach of using WordNet to improve the performance of information retrieval systems. It is similar in principle to word sense disambiguation. It attempts to infer the relations among indexed words in a query or a document. However, instead of using the sense type to represent context, it uses the weights of the indexed terms to represent the context. All terms that are related in a context are given higher weights. For example the words *plant, leaves, root, fruit*, and *flower* are related terms for a concept *flora* and are part of a lexical taxonomy. Therefore they should be given higher weights in comparison to other nonrelated words in the document. In other words, an inference is made that the document discussed the concept of *flora*.

In this model, the authors assume that the semantic content of a document would be adequately represented by the closely related terms according to a given lexical taxonomy. There could be other words that occur many times in the document, hence they may have scored high in a frequency analysis (*tf***idf* weights), however, they should not be considered to be semantically important because they lack support from other words of similar concept.

Unlike the sense disambiguation approach, the semantic weighting approach does not restrict the notion of similarity between a query and a document term to only those with similar senses. For this reason, the problem of over-fitting the model towards precision will not occur. Based on this idea Wang and Brookes (2004) propose a novel approach to semantic indexing by using WordNet to infer semantics between terms in the document. The results of the semantics inference are used to modify the term weights which were normally calculated solely based on *tf***idf*. They provide a hypothesis that words that semantically close in distance most likely represent the meaning of the document. For example, the words *car, engine, wheel,* and *brakes* may occur in a document and they are semantically close according to WordNet, then it can be assumed based on the occurrence of these words that the document contains the discussion on *automobile* or *car*. Distinct to the previous attempts in semantic indexing, such as latent semantic indexing (LSI) (Deerwester, Dumain, Furnas, Landauer, & Harshman, 1990), the semantic weight approach does not suffer from high computational expenses. Research in LSI shows that the model can improve the performance of the retrieval, but computationally, it is very expensive.

In a semantic weights model, the weights of the words or terms in a document is derived using the combination of two matrices: the document-term matrix (*W*) and the term-term matrix (*T*).

The document-term matrix (*W*) represents the *tf*idf* weight of a term in a given document. In a formal way, the document-term matrix can be represented as followed:

Definition 2. Document-Term Matrix (W): *Let i be a term in a document j, m be the total number of documents in the corpus, and n to be the total number of known term in the corpus.*

$$W = (w_{ij})_{mxn},$$

*where, w_{ij} represent the tf*idf weight of the term i in document j.*

Definition 3. Term-Term Matrix (T): *Let P be a set of known terms in the WordNet's hyponomic relations, $P = \{p_a, p_b, ..., p_z\}$.*
Let Q to be a set of known terms in the corpus, $Q = \{q_k, q_l, ..., q_n\}$.
Let X to be a taxonomy tree that has nodes made of synsets S_x.

$T = (r_{qkql})$ where

$$r_{xy} = \begin{cases} 1\, if\ k = l \\ 1\, if\ \exists q_k, q_l\ in\ X. \\ 0\ otherwise \end{cases}$$

According to the definition of the term-term matrix, the first condition determines the values of the diagonal elements in the matrix. The values of the diagonal elements are always 1 because the distance for a term from itself is always 1 regardless whether the terms exist in WordNet. Included in the second condition is a weaker condition, $q_k, q_l \in S_x$, that is, the two terms are synonyms.

Definition 4. Final-Weight Matrix: *The final weights of all the terms in all documents in the collection are given by the product of the matrix W and T.*

$$Z = W \times T$$

Wang and Brookes reported that the retrieval performance in the ADI and TIME collection is improved. Intuitively, we consider that assigning a binary value to represent an association between terms may not be optimal. Hence, we investigate the possibility of assigning an exact distance measure between terms to the element in the term-term matrix. In the forth section, we present two different possible semantic distance measures that can be adopted to improve Wang and Brookes' model.

Debate on the Role of WordNet in Information Retrieval

The impact of WordNet in improving the performance of information retrieval found to be inconsistent in the reviewed works. Table 1 shows that most of the research on query expansion and sense disambiguation in the word-level retrieval does not improve the performance, and if any, it is not significant (Mandala et Al., 1998; Richardson & Smeaton, 1995; Smeaton & Berrut, 1995; Voorhees, 1993, 1994). Improvement is achieved by either adding additional thesauri, such is that of Mandala et al. (1999), or using phrase-based retrieval (Liu et al., 2004).

The type of ontology employed may influence the performance of the retrieval. Hsu et al. (2006) compared the use of WordNet and ConceptNet (ConceptNet, n.d) and concluded that WordNet is more appropriate in finding a kernel word, that is, a specific term that is highly relevant to the topic, whereas ConceptNet is useful in finding more general words for expansion. They suggested that both ontologies will complement each other when used for retrieval. However, they never directly measured the retrieval performance; hence it is still debatable whether it will be the case since some authors have suggested that one of the main reasons that WordNet does not improve the performance is due to its lack of domain-specific terms (Mandala et al., 1998).

In supporting the linguistic analysis, research that uses WordNet as ontology has shown mixed results. Richardson and Smeaton (1995) show little improvement, while Wang and Brookes (2004) report significant improvement. We hypothesize that further improvement can be made to Wang and Brookes' model by measuring the exact semantic distance measure between terms in the term-term matrix instead of only assigning the values of 1 or 0. We present our model and discussion on the experimental results in the forth section.

USING SEMANTIC DISTANCE MEASURE IN THE TERM-TERM MATRIX

Measuring semantic relatedness or distance between words of a natural language has played an important role in many natural processing tasks such as word sense disambiguation, text summarization, speech recognition, and so forth. The most natural way of measuring the similarity of two concepts in taxonomy, given its graphical representation, is to calculate the path distance between the two compared nodes. WordNet is a lexical system with clear taxonomy; hence the path length calculation can be adapted to measure the semantic distance between two words.

There are many semantic distance measures (Budanitsky, 1999) and in this research as initial investigation, we choose two distance measures: the edge counting method and the Leacock-Chodorow (LCH) method. The edge counting is considered to be a straightforward distance measures whereas LCH is more sophisticated because it considers the depth of the tree when calculating the distance between two words. We would like to investigate whether the increase of sophistication in the level of measurement will influence the retrieval performance. Next, these two measures will be discussed in detail.

Edge Counting

The edge counting method (Budanitsky, 1999) assumed that the number of edges between terms in taxonomy is a measure of conceptual distance between terms. Following from Definition 3, p_a and p_b are two terms in the WordNet taxonomy. Let n_{edge} be the total number of edges between p_a and p_b.

The related distance of these terms is represented as:

$$dist_{edge}(p_a,p_b) = \min(n_{edge}) \qquad (1)$$

This method is simple to calculate; however, it has a problem because it does not consider the depth of the tree hence it will be sensitive to the granularity of the concept representation in the taxonomy.

Leacock-Chodorow

The Leacock-Chodorow method (Budanitsky, 1999) includes the tree depth into the calculation to remove the sensitivity to the granularity of the concept representation. The related distance of two terms p_a and p_b is measured by:

$$dist_{LCH}(p_a,p_b) = -\log\left(\frac{dist_{edge}(p_a,p_b)+1}{2xD}\right) \qquad (2)$$

In the calculation of the minimum path, this method uses the number of nodes instead of edges. This is to avoid singularities, so that synonyms are 1 unit of distance apart. The number of nodes in graph theory can be calculated as the total number of edges plus one, and hence taking Equation 1 as a base, the distance measured can be formulated as Equation 2. The D represents the maximum depth of the taxonomy.

Term-term Matrix with Semantic Dis*tance*

Based on Definition 3, Equation 1, and Equation 2, the new term-term matrix T for the edge counting and LCH methods can be represented as:

Table 1. Information retrieval systems with WordNet

Researchers	Techniques	Ontology	Results
Voorhees (1993)	Sense disambiguation	WordNet	• Improve in precision for short query. • Degradation of recall.
Voorhees (1994)	Query expansion	WordNet	• Improve the performance of short query, but not for long query.
Smeaton and Berrut (1995)	Query expansion	WordNet	• Decreased in performance.
Richardson and Smeaton (1995)	Linguistic analysis	WordNet	• Little improvement for extra effort of creating the knowledge base.
Mandala et al. (1998)	Query expansion	WordNet + automatically constructed thesauri.	• Little improvement when only WordNet is used. • Up to 98.2% improvement when WordNet is used in conjunction with other automatic generated thesauri.
Mandala et al. (1999)	Query expansion and sense disambiguation	WordNet and Roget's thesaurus	• Little improvement is achieved when either the WordNet or Roget's thesaurus is used in isolation. • Big improvement is achieved when two of the thesauri are used.
Moldovan and Mihalcea (2000)	Query expansion and sense disambiguation	WordNet	• The experiments were not conducted in the traditional test bed. • Q&A system. • Positive findings.
Liu et al. (2004)	Query expansion and sense disambiguation	WordNet	• The retrieval is based on phrases rather than word. • Improve the precision.
Wang and Brookes (2004)	Linguistic analysis	WordNet	• Improve the precision and recall.
Hsu et al. (2006)	Query expansion	WordNet and ConceptNet	• WordNet is useful in finding kernel words. • ConceptNet is useful in finding cooperative concepts words.

$$T_{edge} = (r_{qkql})_{nxn}$$

where

$$r_{xy} = \begin{cases} 1\ if\ k = l \\ dist_{edge}\ if\ \exists q_k, q_l\ in\ X \\ 0\ otherwise \end{cases} \quad (3)$$

$T_{LCII} = (r_{qkql})_{nxn}$

where

$$r_{xy} = \begin{cases} 1\, if\ k = l \\ dist_{LCH}\ if\ \exists q_k, q_l\ in\ X \\ 0\ otherwise \end{cases} \qquad (4)$$

In this approach, instead of assigning value of 1 to two related terms, the exact distance measure between the two terms is calculated and is used to populate the term-term matrix.

EXPERIMENTS AND RESULTS

We tested the performance of four different term weights: the traditional *tf*idf*, the Wang and Brookes', the edge counting, and the LCH. The SMART retrieval system was used to tokenize and identify the indexed words. The stop list and stemming were applied to all cases of term weights. In calculating the semantic distance, the WordNet 1.7 was used. We used the vector space retrieval model with cosine similarity for the retrieval. The WordNet is used to measure the semantic distances of the indexed terms in the corpus. We used the hyponomic relations of the noun speech in WordNet. The test is conducted in the ADI collection. Table 2 shows the setup of the two runs across 35 queries in ADI collection.

One issue that we encountered in using the WordNet to calculate the distance measure is the treatment of the stemmed word. We used SMART's stemming module which is based on Porter's stemming algorithm, hence the result of stemming is not necessarily a proper English word. In this case, it is possible that two words are not associated because the system could not find the matching string that represents the stemmed word in WordNet taxonomy, although the proper word may actually exist in the taxonomy. We try to reduce the impact of the stemming by using the substring matching instead of the full-word matching in the WordNet. Using this approach we managed to reduce the number of false nonmatching cases. Figure 3 and Figure 4 show the results of the experiments.

It is interesting to observe that the nonsemantic weight, *tf*idf* weighting, outperforms all the semantic indexing techniques, including the Wang and Brookes' technique. To the best

Table 2. Experiments settings

	Exact Term Run	**Substring Term Run**
Collection	ADI[1]	ADI [1]
No of documents	82	82
No of queries	35	35
Stop List	SMART stop list	SMART stop list
Stemming	Porter's stemming in SMART.	No stemming
Ontology	WordNet 1.7	WordNet 1.7
Retrieval Model	Cosine similarity in SMART.	Cosine similarity in SMART
Weights		
Traditional	Augmented *tf*idf* in SMART	Augmented *tf*idf* in SMART
Wang and Brookes	As the definition in this article.	As the definition in this article.
Edge counting	As the definition in this article.	As the definition in this article.
LCH	As the definition in this article.	As the definition in this article.

Figure 3. Exact tem matching results

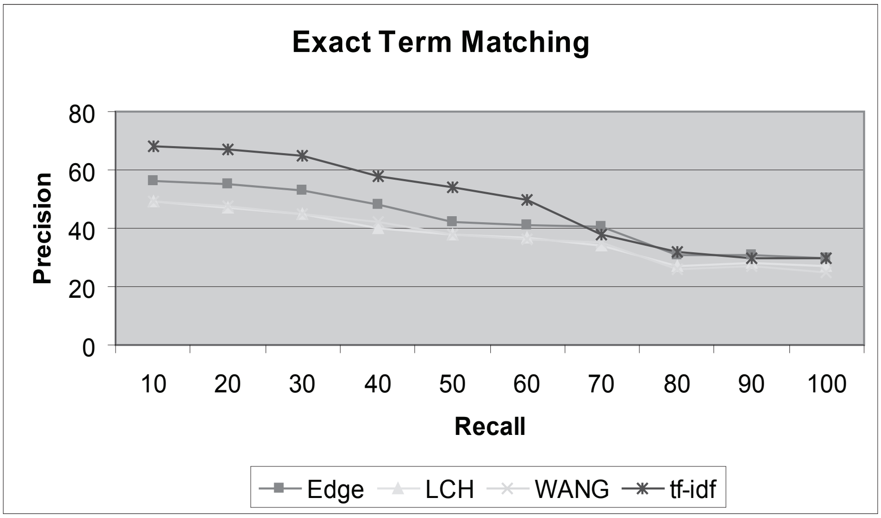
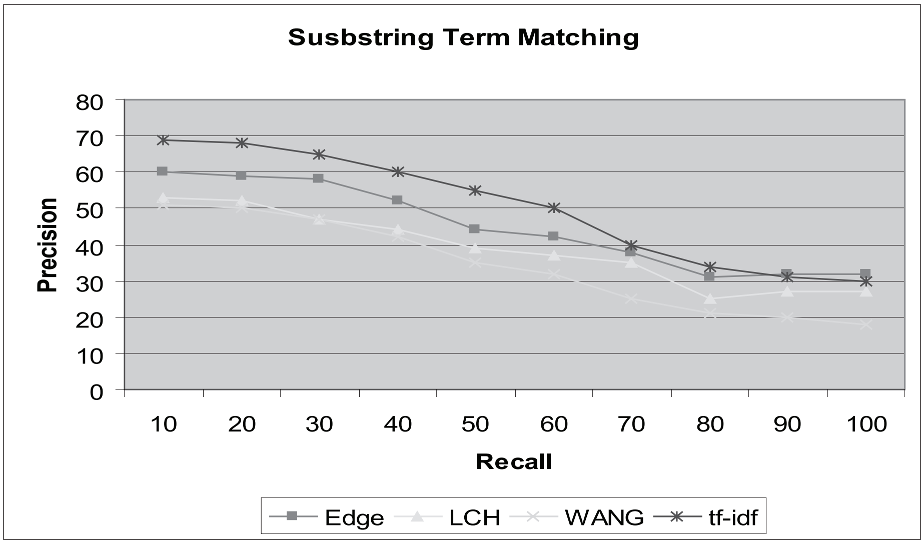


Figure 4. Substring term matching results



of our knowledge we have followed exactly the model described in their work (Wang & Brookes, 2004); however, we never managed to get the same results during the experiments. In order to make the comparison objective, we use our implementation of the Wang and Brooke's results in our discussion. We made sure that all the parameters such as stemming algorithm, stop list, and *tf*idf* weights for the document-term matrix were kept the same.

The results were not as anticipated. We expected that the semantic weightings would perform better than the traditional *tf*idf* weighting. Moreover, we also expected that the LCH would outperform the edge counting technique due to its normalization on the tree depth. The results depicted in both Figure 3 and 4 place the *tf*idf* approach to be the best.

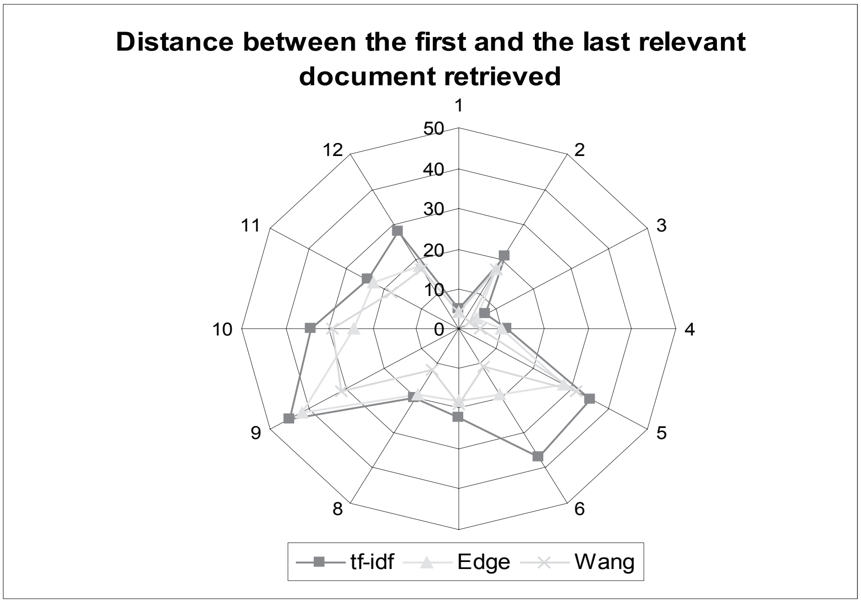
We investigated further the results to find the explanation on the poor performance of the semantic weightings. The first thought suggested that the semantic weighting may not be a good technique to improve precision because it causes the relevant documents to be closely clustered in the ranking. To observe this possibility, we calculate the distance between the position of the first relevant document and the last relevant document in the ranking. This data can show the spread of relevant documents in the ranked output. Figure 5 shows the distance of the relevant documents for a sample of 12 queries in the ADI collection.

The graph shows that the Wang and the edge models have the value of the distance closer to the central point of the graph. It shows that for all the 12 queries, the number of documents in between the first found relevant document and the last found relevant document for the *tf*idf* approach is greater than other weighting techniques. This observation shows that the semantic weighting creates a closer cluster of relevant documents compared to the output of the *tf*idf*. The cluster is created around the middle of the ranking. From this observation, it is possible that the semantic weights may actually pull the relevant documents from the top ranked into a lower rank around the middle of the ranking and pull out the lower ranked relevant documents towards the middle ranks. From the point of view of retrieval, this behaviour is not ideal.

CONCLUSION

The impact of ontologies, such as WordNet, on the performance of information retrieval has been investigated and possible improvement to the semantic indexing model has been proposed. The investigation suggests that the use of WordNet alone as an ontology to improve information retrieval performance is not appropriate. WordNet as an ontology for information retrieval has the following characteristics that may not be ideal for supporting information retrieval tasks:

Figure 5. Distance between the highest ranking and the lowest ranking of relevant documents



- WordNet contains mainly general English terms. Domain-specific terms or proper names are not represented. Hence many of the retrievals that contain domain-specific terms or proper names will not improve.
- Relations between terms are limited to a single speech. It is not possible to find relation of an adjective and a noun in WordNet. Hence, in the semantic indexing, it is not possible to derive the relation between the color "green" with the concept of "forest" for example.

The first problem stated above is exacerbated by the fact that ADI collection is a small and confined collection. There are many specific computing terms that cannot be found in WordNet. In a small collection such as ADI, improving recall has less negative impact to the level of precision. Hence, a traditional retrieval model based on Porter's stemming algorithm and vector space model produces good results without additional semantic processing using WordNet. This observation has been reported by all related work reviewed in this article. The use of WordNet to improve precision does not provide significant improvement, if not any.

Does this mean that an ontology does not have any role in information retrieval? The answer is no. Ontologies can still play a role in improving the performance of information retrieval, provided the following considerations are observed:

- The ontology used in the retrieval has to be domain or collection specific. WordNet can be used as a starting point and needs to be either combined with other ontologies or be expanded with domain-specific entries.
- Traditional test collections such as ADI, CACM, and MEDLINE tend to favor a model that has a very high recall due to its small size and narrow topic scope. The nature of information retrieval has slightly changed with the introduction of the Internet. The number of possible matching documents is very large, and hence, having a very high recall may not be necessary. It is possible that the user will be satisfied with looking at only 2% of the total matching documents. However, the user will expect that this 2% be highly relevant. Hence, we may need to investigate further in the future a way to report the results of experiments in information retrieval. Recall may not be an important factor in searching information on the World Wide Web.

The debate on the impact of ontology in information retrieval will continue; however, without finding an appropriate test bed and a way of reporting the results, the true potential of ontology in improving the performance of information retrieval may not be ever realized. It is a challenge for the information retrieval community to find a new way of reporting the results according to the evolution in the nature of the way users use the World Wide Web to find information.

REFERENCES

Benitez, A. B., Chang, S. F., & Smith, J. R. (2001, October). *IMKA: A multimedia organization system combining perceptual and semantic Knowledge*. Paper presented at the *ACM Multimedia*, Ottawa, Canada.

Bramantoro, A., Krishnaswamy, S., & Indrawan, M. (2005). *A semantic distance measure for matching Web services.* Paper presented at the Web Information Systems Engineering – WISE 2005 Workshops (pp. 217-226).

Budanitsky, A. (1999). *Lexical semantic relatedness and its application in natural language processing* (Tech. Rep. CSRG-390). University of Toronto, Computer Systems Research Group.

Caliusco, M. L., Galli, M. R., & Chiotti, O. (2005, October 31-November 2). Contextual ontologies for the Semantic Web: An enabling technology. In *Proceedings of the Third Latin American Web Congress* (*LA-WEB*) (p. 98). Washington, D.C.: IEEE Computer Society.

Chen, S., Alahakoon, D., & Indrawan, M. (2005). *Background knowledge driven ontology discovery*. Paper presented at the 2005 IEEE International

Conference on e-Technology, e-Commerce and e-Service (EEE '05) (pp. 202-207).

Deerwester, S., Dumais, S. T, Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Dou, D., LePendu, P., Kim, S., & Qi, P. (2006, April 3-7). Integrating databases into the semantic Web through an ontology-based framework. In *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)* (Vol. 00, p. 54). Washington, D.C.: IEEE Computer Society.

Hachey, B., & Grover, C. (2005, June 6-11). Automatic legal text summarization: Experiments with summary structuring. In *Proceedings of the 10th International Conference on Artificial intelligence and Law* (ICAIL '05), Bologna, Italy, (pp. 75-84). New York: ACM Press.

Hsu, M. H., Tsai, M. F., & Chen, H. H. (2006, October 16-18). Query expansion with ConceptNet and WordNet: An intrinsic comparison. In *Proceedings of the Third Asia Information Retrieval Symposium*, Singapore, (LNCS 4182, pp. 1-13).

Huang, J., Dang, J., & Huhns, M. N. (2006, September 18-22). Ontology reconciliation for service-oriented computing. In *Proceedings of the IEEE International Conference on Services Computing* (pp. 3-10). Washington, D.C.: IEEE Computer Society.

Khan, L., & Luo, F. (2002). Ontology construction for information selection. In *Proceedings 14th IEEE International Conference on Tools with Artificial Intelligence* (pp. 122- 127).

Liu, S., Liu, F., Yu, C., & Meng, W. (2004, July). An effective approach to document retrieval via utilizing wordNet and recognizing phrases. In *Proceedings of the 27th Annual International ACM SIGIR Conference*, Sheffield, UK, (pp. 266-272).

Mandala, R., Tokunaga, T., & Tanaka, H. (1998). The use of WordNet in information retrieval. In S. Harabagiu (Ed.), *Use of WordNet in Natural Language Processing Systems: Proceedings of the Association for Computational Linguistics Conference,* Somerset, NJ, (pp. 31-37).

Mandala, R., Tokunaga, T., & Tanaka, H. (1999). Complementing WordNet with Roget and corpus-based automatically constructed thesauri for information retrieval. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen.

Moldovan, D. I., & Mihalcea, R. (2000). Using WordNet and lexical operators to improve Internet searches. *IEEE Internet Computing, 4*(1), 34-43.

Rack, C., Arbanowski, S., & Steglich, S. (2006, July 23-27). *Context-aware, ontology-based recommendations.* Paper presented at the International Symposium on Applications and the Internet Workshops.

Richardson, R., & Smeaton, A. F. (1995). *Using WordNet in a knowledge-based approach to information retrieval* (Tech. Rep. CS-0395). Dublin City University, School of Computer Applications.

Smeaton, A. F., & Berrut, C. (1995). Running TREC-4 experiments: A chronological report of query expansion experiments carried out as part of TREC-4. In *Proceedings of the Fourth Text Retrieval Conference (TREC-4).* NIST Special Publication.

Voorhees, E. M. (1993, June 27-July 1). Using WordNet to disambiguate word senses for text retrieval. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proceedings of the 16th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (SIGIR '93), Pittsburgh, (pp. 171-180). New York: ACM Press.

Voorhees, E. M. (1994, July 3-6). Query expansion using lexical-semantic relations. In W. B. Croft & C. J. van Rijsbergen (Eds.), *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in information Retrieval,* Dublin, Ireland, (pp. 61-69). New York: Springer-Verlag.

Wang, B., & Brookes, B. R. (2004). *A semantic approach for Web indexing* (LNCS 3007, pp. 59-68).

www.conceptnet.org

Yan, Z., Li, Q., & Li, H. (2006). *An ontology-based model for context-aware*. Paper presented at the 1st International Symposium on Pervasive Computing and Applications (pp. 647-651).

ENDNOTE

[1] The ADI used comes as part of SMART retrieval system, ftp://ftp.cs.cornell.edu/pub/smart/

M. Indrawan is a senior lecturer of computer science in the Faculty of Information Technology, Monash University. She received her PhD in computer science from Monash University. Her current research focuses on pervasive computing, information retrieval and context-aware systems.

S. Loke is a senior lecturer of computer science in the Department of Computer Science, La Trobe University. He received his PhD in computer science from University of Melbourne. His research interests are pervasive computing, smart containers, social devices and context-aware pervasive systems.