Information Retrieval Based on Semantic Similarity Using Information Content

Kishor Wagh [1], Satish Kolhe [2]

[1] Department of Information Technology, Government College of Engineering, Amravati, India

[2] Department of Computer Science, North Maharashtra University, Jalgaon, India

Abstract

Evaluating semantic similarity of concepts is a problem that has been extensively investigated in the literature in different areas, such as artificial intelligence, cognitive science, databases and software engineering. Semantic similarity relates to computing the similarity between conceptually similar but not necessarily lexically similar terms. Currently, it is growing in importance in different settings, such as digital libraries, heterogeneous databases and in particular the Semantic Web. In such contexts, very often concepts are organized according to taxonomy (or a hierarchy). We investigate approaches to compute the semantic similarity between natural language terms. This paper presents new approach for measuring semantic similarity between words and hierarchical structure is used to present information content. In this paper, we present a search engine using Google API that expands the user query based on similarity scores of each term of user's query. Users query words are replaced with synonyms discovered from the similarity measures and input to the Google search API.

Keywords: *Search engine, Concept, Information content similarity.*

1. Introduction

Web contains very large amount of information, which are scattered and dynamic as well as diverse in terms of content and nature. Since people with different background, knowledge, and expectation organize the information in web, users query is not adequate to represent the information they want to retrieve. Keyword matching technique fails to retrieve semantically or lexically related document thus retrieving more irrelevant results. Such techniques are constrained by attempting to match the user keyword to the source document and present information to the user with documents that matched the user keyword. Our method uses the Information content approach to calculate similarity between two concepts in the taxonomy to discover the related concepts, which are not implicit in the query. For example a s earch query seeking for the information on given term would return hits containing the specified term but would fail to retrieve the document that is described by its synonymy term.

In this paper, we presented an approach for capturing similarity between words that is concerned with the syntactic similarity of two strings. Semantic similarity is a confidence score that reflects the semantic relation between the meanings of two sentences. It is difficult to gain a high accuracy score because the exact semantic meanings are completely understood only in a particular context. Some dictionary-based algorithms are available to capture the semantic similarity between two words.

Context used in search query is of great importance in retrieving relevance information thus finding the meaning of the each word used in query is essential. For this similarity score of the concepts represented by each word in the query is computed. The pair of concept that has higher similarity value is chosen as the concept described by the words. This discovered concept is used to supplement users query with its synonyms based on relatedness score. In this paper, we will present a method for retrieving information which uses the algorithm given in section 4.

The paper is organized as follows. Section 2 describes the background. Section 3 reviews the WordNet taxonomy, database, word senses, and semantic similarity measurement. Section 4 briefly describes the search engine algorithm. Section 5 describes the results. And conclusion is given in section 6.

2. Background

WordNet connects concepts or senses, but most words have more than one sense. Until now, several approaches for computing similarity between concepts have been proposed.Acording to the parameter used in the similarity approaches, they can be classified into three main categories, including the edge-based approach, the

information content approach and the hybrid approach which combines the first two methods. Leacock and Chodorow [9] proposed a semantic similarity measure that typifies the edge-based approach. In their measure, the similarity is determined by the length of shortest path that connects two concepts in the WordNet taxonomy.Wu and Palmer's method [16] calculates similarity by considering the depths of the two concepts in the WordNet hierarchy, along with the depth of the lowest super-ordinate.Resnik [5] introduced the first similarity measure to combine corpus statistic with a conceptual taxonomy. The key intuition in Resnik's measure is that for any two concepts, the most specific concept that subsumes them both in the conceptual taxonomy represents the information that the concepts share in common. He determines similarity by calculating the information content of the shared subsume.

3. WordNet Taxonomy

WordNet is a lexical database for the English language [1]. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The specific meaning of one word under one type of POS is called a sense. Each synset [12][13] has a gloss that defines the concept it represents. For example, the words night, nighttime, and dark constitute a single synset that has the following gloss: the time after sunset and before sunrise while it is dark outside. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. The database can also be browsed online. WordNet [6] was created and is being maintained at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George A. Miller. Development began in 1985. WordNet's latest version is 3.0.

Synsets are connected to one another through explicit semantic relations. Some of these relations (hypernym, hyponym for nouns, and hypernym and troponym for verbs) constitute is-a-kind-of (holonymy) and is-a-part-of (meronymy for nouns) hierarchies. For example, tree is a kind of plant, tree is a hyponym of plant, and plant is a hypernym of tree. Analogously, trunk is a part of a tree, and we have trunk as a meronym of tree. While semantic relations apply to all members of a synset because they share a meaning but are all mutually synonyms, words can also be connected to other words through lexical relations, including antonyms (opposites of each other) which are derivationally related, as well. WordNet also provides the *polysemy count* of a word: the number of synsets that contain the word. If a word participates in several synsets (i.e. has several senses) then typically some senses are much more common than others.

3.1 WordNet Database

For each syntactic category, two files represent the WordNet [6] database — index.pos and data.pos, where pos is either noun, verb, adj or adv. The database is in an ASCII format that is human- and machine-readable, and is easily accessible to those who wish to use it with their own applications. The index and data files are interrelated. The WordNet morphological processing function, morphy(), handles a wide range of morphological transformations. During WordNet development synsets are organized into forty-five lexicographer files based on syntactic category and logical groupings. grind() processes these files and produces a database suitable for use with the WordNet library, interface code, and other applications. A file number corresponds to each lexicographer file. File numbers are encoded in several parts of the WordNet system as an efficient way to indicate a lexicographer file name. The file lexnames lists the mapping between file names and numbers, and can be used by programs or end users to correlate the two.

The syntactic categories in WordNet are– noun, verb, adjective and adverb. Each lexicographer file consists of a list of synonym sets (*synsets*) for one part of speech. Although the basic synset syntax is the same for all of the parts of speech, some parts of the syntax only apply to a particular part of speech. Each filename specified is of the form:

pathname/pos**.**suffix

where pathname is optional and pos is either noun, verb, adj or adv. suffix may be used to separate groups of synsets into different files, for example noun.animal and noun.plant. One or more input files, in any combination of syntactic categories, may be specified. A list of the lexicographer files used to build the complete WordNet database.grind**( )** produces the following output files:

Table 1: Files of WordNet

Filename	**Description**
Index.pos	Index file for each syntactic category
Data.pos	Data file for each syntactic category
Index.sense	Sense index

The WordNet sense index provides an alternate method for accessing synsets and word senses in the WordNet database. It is useful to applications that retrieve synsets or other information related to a specific sense in WordNet, rather than all the senses of a word. It can also be used with tools like **grep** and Perl to find all senses of a word in

one or more parts of speech. A specific Word-Net sense, encoded as a *sense_key*, can be used as an index into this file to obtain its WordNet sense number, the database byte offset of the synset containing the sense, and the number of times it has been tagged in the semantic concordance texts.

A *sense_key* is the best way to represent a sense in semantic tagging or other systems that refer to WordNet senses. *sense_key*s are independent of WordNet sense numbers and *synset_offset*s, which vary between versions of the database. Using the sense index and a *sense_key*, the corresponding synset (via the *synset_offset*) and WordNet sense number can easily be obtained.The sense index file lists all of the senses in the WordNet database with each line representing one sense. The file is in alphabetical order, fields are separated by one space, and each line is terminated with a newline character.Each line is of the form:

sense_key synset_offset sense_number tag_cnt

sense_key is an encoding of the word sense. Programs can construct a sense key in this format and use it as a binary search key into the sense index file. synset_offset is the byte offset that the synset containing the sense is found at in the database "data" file corresponding to the part of speech encoded in the sense_key. synset_offset is an 8 digit, zero-filled decimal integer, and can be used with fseek to read a synset from the data file.

sense_number is a d ecimal integer indicating the sense number of the word, within the part of speech encoded in sense_key, in the WordNet database. tag_cnt represents the decimal number of times the sense is tagged in various semantic concordance texts. A tag_cnt of **0** indicates that the sense has not been semantically tagged. All of the WordNet noun synsets are organized into hierarchies, headed by the unique beginner synset for entity in the file noun.Tops.

3.2 WordNet as an ontology

The hypernym/hyponym relationships among the noun synsets can be interpreted as specialization relations between conceptual categories. In other words, WordNet can be interpreted and used as a lexical ontology [2] in the computer science sense.

The WordNet dictionary contains the senses of words. The frequency of particular sense is given in parenthesis and “n” indicate the noun (n in parenthesis).According to WordNet dictionary [6], the word “person” has three senses:

- sense 1: (6833)S: (n) **person**, individual, someone, somebody, mortal, soul (a human being) *"there was too much for one person to do"*
- sense 2:(1)S: (n) **person** (a human body (usually including the clothing)) *"a weapon was hidden on his person"*
- sense 3 :S: (n) **person** (a grammatical category used in the classification of pronouns, possessive determiners, and verb forms according to whether they indicate the speaker, the addressee, or a third party) *"stop talking about yourself in the third person"*

The word “student” has one sense:

- sense 1 : (67)S: (n) **student**, pupil, educatee (a learner who is enrolled in an educational institution)

student,pupil,educate are called as synonyms of sense 1 of the word “student”. The word “worker” has four senses:

- sense 1 :(29)S: (n) **worker** (a person who works at a specific occupation) *"he is a good worker"*
- sense 2: (4)S: (n) proletarian, prole, **worker** (a member of the working class (not necessarily employed)) *"workers of the world--unite!"*
- sense 3:(4)S: (n) **worker** (sterile member of a colony of social insects that forages for food and cares for the larvae)
- sense 4::S: (n) actor, doer, **worker** (a person who acts and gets things done) *"he's a principal actor in this affair"; "when you want something done get a doer"; "he's a miracle worker"*

The word “interest” has seven senses:

- sense 1: (62)S: (n) **interest**, involvement (a sense of concern with and curiosity about someone or something) *"an interest in music"*
- sense 2: (32)S: (n) sake, **interest** (a reason for wanting something done) *"for your sake"; "died for the sake of his country"; "in the interest of safety"; "in the common interest"*
- sense 3: (21)S: (n) **interest**, interestingness (the power of attracting or holding one's attention (because it is unusual or exciting etc.)) *"they said nothing of great interest"; "primary colors can add interest to a room"*
- sense 4: (14)S: (n) **interest** (a fixed charge for borrowing money; usually a percentage of the amount borrowed) *"how much interest do you pay on your mortgage?"*
- sense 5: (7)S: (n) **interest**, stake ((law) a right or legal share of something; a financial involvement with something) *"they have interests all over the world"; "a stake in the company's future"*

- sense 6: (5)S: (n) **interest**, interest group ((usually plural) a s ocial group whose members control some field of activity and who have common aims) *"the iron interests stepped up production"*
- sense 7: (3)S: (n) pastime, **interest**, pursuit (a diversion that occupies one's time and thoughts (usually pleasantly)) *"sailing is her favorite pastime"; "his main pastime is gambling"; "he counts reading among his interests"; "they criticized the boy for his limited pursuits"*

The word "subject" has eight senses:

- sense 1:(20)S: (n) **subject**, topic, theme (the subject matter of a co nversation or discussion) *"he didn't want to discuss that subject"; "it was a very sensitive topic"; "his letters were always on the theme of love"*
- sense 2: (14)S: (n) **subject**, content, depicted object (something (a person or object or scene) selected by an artist or photographer for graphic representation) *"a moving picture of a t rain is more dramatic than a s till picture of the same subject"*
- sense 3: (11)S: (n) discipline, **subject**, subject area, subject field, field, field of study, study, bailiwick (a branch of knowledge) *"in what discipline is his doctorate?"; "teachers should be well trained in their subject"; "anthropology is the study of human beings"*
- sense 4: (9)S: (n) topic, **subject**, issue, matter (some situation or event that is thought about) *"he kept drifting off the topic"; "he had been thinking about the subject for several years"; "it is a matter for the police"*
- sense 5: (4)S: (n) **subject** ((grammar) one of the two main constituents of a sentence; the grammatical constituent about which something is predicated)
- sense 6: (2)S: (n) **subject**, case, guinea pig (a person who is subjected to experimental or other observational procedures; someone who is an object of investigation) *"the subjects for this investigation were selected randomly"; "the cases that we studied were drawn from two different communities"*
- sense 7: (2)S: (n) national, **subject** (a person who owes allegiance to that nation) *"a monarch has a duty to his subjects"*
- sense 8:S: (n) **subject** ((logic) the first term of a proposition)

The sense 1 of word "interest" and sense 3 of word "subject" are semantically similar. To measure the semantic similarity between two words, we use hyponym/hypernym (or is-a relations). Due to the limitation of is-a hierarchy, we only work with "noun-noun". A simple way to measure the semantic similarity between two synsets is to treat taxonomy as an undirected graph and measure the distance between them in WordNet. The length of the path between two members of the same synset is 1 (synonym relations).

3.3 Similarity Measurement Using Path Length

Semantic similarity can be measured by simply counting the length of the path or node between the concepts. Resnik (1995), said that "the shorter the path from one node to another, the more similar they are". This figure shows an example of the hyponym taxonomy in WordNet used for path length similarity measurement:

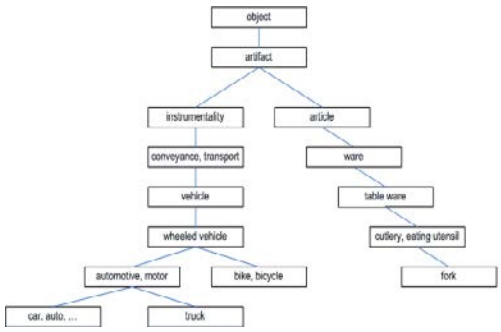

Fig.1. Taxonomy in WordNet

In the above figure, we observe that the length between car and auto is 1, car and truck is 3, car and bicycle is 4, car and fork is 12.A shared parent of two synsets is known as a sub-sumer. The least common sub-sumer (LCS) of two synsets is the sumer that does not have any children that are also the sub-sumer of two synsets. In other words, the LCS of two synsets is the most specific sub-sumer of the two synsets. Back to the above example, the LCS of {car, auto..} and {truck..} is {automotive, motor vehicle}, since the {automotive, motor vehicle} is more specific than the common sub-sumer {wheeled vehicle}.

3.4 Semantic Similarity using Information Content

WordNet connects concepts or senses, but most words have more than one sense. Word similarity can be determined by the best conceptual similarity value among all the concept (sense) pairs. It can be defined as follows:

$$sim(w_1, w_2) = \underset{c_1 \in sen(w_1), c_2 \in sen(w_2)}{Max} [sim(c_1, c_2)]$$

Where *sen* (*w*) denotes the set of possible senses for word *w*.
Traditionally, in order to evaluate the semantic similarity of hierarchically related concepts, the information content approach is adopted. It is based on the association of probabilities with the concepts of the hierarchy. In particular, the *probability* of a concept *c* is defined as:

$$p(c) = \frac{freq(c)}{M}$$

Where *freq*(*c*) is the *frequency* of the concept *c* estimated using noun frequencies from large text corpora [3] and *M* is the total number of observed instances of nouns in the corpus. In this example, probabilities have been assigned according to the *SemCor* project, which labels subsections of the *Brown Corpus* to senses in the *WordNet* lexicon. Lin [4] takes information content approach for computing the semantic similarity between two words. The *information content similarity* (sim) of two concepts c_1, c_2 as follows:

$$sim(c_1, c_2) = \frac{2 \log p(c)}{\log p(c_1) + \log p(c_2)}$$

Where *c* is the concept providing the maximum information content shared by c_1 and c_2 in the taxonomy, i.e., the more information two concepts share, the more similar they are. Note that *c is* the upper bound of c_1, c_2 in the taxonomy whose information content is maximum, i.e., when defined, the least upper bound.

Another method which is proposed by Jiang and Conrath based on a combination of using edge counts in the WordNet IS-A hierarchy and using the information content values of the concepts. This approach [3][14][15] takes both of the concept and their common ancestor in the calculation of similarity. Jiang-Conrath measure gives semantic distance rather than similarity or relatedness.

$$\text{Dist}(c_1, c_2) = \text{IC}(c_1) + \text{IC}(c_2) - 2*\text{IC}(c)$$

Where *c* is the concept providing the maximum information content shared by *c*1 and *c*2 in the taxonomy. This distance measure can be converted to a s imilarity measure by taking the multiplicative inverse of it:

$$\text{Sim}_{JC}(c_1, c_2) = 1/\text{Dist}(c_1, c_2)$$

Thus $\text{Sim}_{JC}(c_1, c_2)$ gives the similarity between concept c_1 and concept c_2.

4. Search Engine

Search engine uses the following algorithm for retrieving information.
Algorithm

1. Enter two words as query input for finding semantic similarity.
2. Compute the semantic similarity between two words using information content approach. Given two words, the measurement determines how similar the meaning of two words is. The higher the score, the more similar the meaning of the two words.
3. Query expansion module-To represent the semantically similar terms the user query is not sufficient for semantic information retrieval task. The concept that the words represent in the search query is used for the expansion of the query [10][11]. The expansion takes all the synonyms of the concept and its one or more hypernyms and hyponyms. Hypernyms may be included based on the similarity score or hypernym up to one level is included in every words of query.

For example the word "person" and "worker" for computing semantic similarity. Information Content similarity of Words is
person#n#1, worker#n#1=0.5959
person#n#1, worker#n#2=0.3099
person#n#1, worker#n#3=0.3069
person#n#2, worker#n#1=0.1685
person#n#2, worker#n#3=0.1259
person#n#2, worker#n#2=0.1236
The word person sense 1 and worker sense 1 have highest similarity score. So here we replace the word "person" with set of synonyms of person sense 1 and replace the word "worker" with set of synonyms of worker sense1.

4. In the system, a WWW search engine accepts the set of synonyms, which is generated by the query expansion module, as an extra query keywords in addition to the ones specified by the user. We used Google search engine to supplement the query with our analyzed terms.

5. Evaluation

In this section, we compare the results of semantic similarity based on information content with semantic similarity using Jiang and Conrath method. Table 2 lists the results of each similarity measure for the pairs of words [3][4][7] using information content and Jiang-Conrath method. It is used to evaluate semantic similarity of hierarchically organized concepts. Table 2 shows results of semantic similarity based on information content are good.

After calculating a semantic similarity, replace the concept c_1 with set of synonyms of concept c_1 sense which has highest similarity score and replace the concept c_2 with set of synonyms of concept c_2 sense which has highest similarity score. A search engine accepts the set of synonyms.

Table 2. Word Semantic Similarity Measurement

Word Pair		Sim	Sim_{JC}
Boy	lad	0.7979	0.2929
Coast	shore	0.9632	1.6154
automobile	car	1.0000	1.2876
implement	tool	0.9146	0.8484
Food	rooster	0.0762	0.0671
Brother	monk	0.2097	0.0689
Brother	lad	0.2400	0.0829
Car	journey	0.0000	0.0707
data	mining	0.0861	0.0604
Monk	slave	0.2011	0.0661
Coast	forest	0.1181	0.0627
Lad	wizard	0.2241	0.0758
Chord	smile	0.3269	0.0790
Person	worker	0.5959	0.3875
Person	student	0.4375	0.2043
computer	java	0.1119	0.0605
Pen	pencil	0.9162	0.6202
Baby	child	1.0000	1.2876
Array	list	0.3249	0.0783
Segment	page	0.0742	0.0653
engineering	technology	1.0000	1.2876
knowledge	data	0.1526	0.1155
Array	vector	0.0946	0.0670
Attribute	record	0.2122	0.1302
File	buffer	0.0842	0.0590
Journey	Voyage	0.8277	0.3533
Magician	Wizard	1.0000	0.0000
Midday	Noon	1.0000	0.0000
Furnace	Stove	0.2294	0.0597
Food	Fruit	0.1559	0.0861
Bird	Cock	0.7881	0.2680
Crane	Implement	0.3327	0.0784
Coast	Hill	0.7286	0.2187
Glass	Magician	0.1421	0.0604
Noon	String	0.0923	0.0653

6. Conclusion and Future Work

In this paper, we present a co ncept similarity matching method based on information content using the hierarchy of WordNet. The results give the similarity measures of words. We have found that replacing query with set of synonyms based on the similarity score can indeed enhance the information retrieval (IR) task. Users frequently fail to describe the information they want to retrieve in the search query.

In future work, we are extending the semantic matching approach by computing semantic similarity among different ontologies. The algorithm presented here can be further enhanced with incorporating Word Sense Disambiguation (WSD). With the computed similarity, in the Similarity computation module, WSD can be performed by maximizing relatedness for the generation of the concepts required by the query expansion module.

References

[1] C.Fellbaum; *A Semantic Network of English: the Mother of all WordNets.* Computers and the Humanities 32, 209-220, 1998.

[2] A.Formica, M.Missikoff Concept Similarity in SymOntos: an Enterprise Ontology management Tool; The Computer Journal, 583-594, 2002.

[3] J. J.Jiang, D. W.Conrath *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*; The Computing Research Repository , 1997.

[4] D.Lin *An Information-Theoretic Definition of Similarity.* Proc. of the Int. Conference on Machine Learning (ICML), Morgan Kaufmann, 296-304, 1998.

[5] P.Resnik *Using Information Content to Evaluate Semantic Similarity in a Taxonomy.* Proc. of the Int. Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann, 448-453, 1995.

[6] *WordNet 2.1: A lexical database for the English language*; http://www.cogsci.princeton.edu/cgi-bin/webwn, 2005.

[7] Krishna Sapkota,Laxman Thapa,Shailesh Pandey Efficient Information Retrieval using measures of Semantic Similarity,2006.

[8] Anna Formica Concept similarity by evaluating information contents and feature vectors: A combined approach. Communications of the ACM, Vol.52, 2009.

[9] Leacock C.,Chodorow M.,"Combining local context and WordNet similarity for word sense identification",In Fellbaum 1998,pp.133-138.

[10] Jiuling Zhang, Beixing Deng, Xing Li, Concept Based Query Expansion Using WordNet. International e-Conference on Advanced Science and Technology,2009.

[11] George Tsatsaronis and Vicky Panagiotopoulou, A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness,2009.

[12] Mark D. Wood, Alexander Loui, Stacie Hibino, Searching Consumer Image Collections Using Web-based Concept Expansion,2010.

[13] Christof M uller and Iryna Gurevych, A Study on t he Semantic Relatedness of Query and Document Terms in Information Retrieval. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1338–1347,Singapore, 6-7 August 2009.

[14] Zhongcheng Zhao, Jianzhuo Yan, Liying Fang, Pu Wang, Measuring Semantic Similarity Based On WordNet. Sixth Web Information Systems and Applications Conference.IEEE, 2009.

[15] P. Nithiya, V. Vidhya, Dr. L. Ganesan, Development of semantic based information retrieval using word-net approach. Second International Conference on C omputer and Network Technology, 2010.

[16] Wu Z.,Plamer M.,"Verb semantics and lexical selection",Proceeding of the 32nd annual meeting of the association for computational linguistics,New Mexico,1994.

Kishor Wagh Diploma in Computer Engineering from Government Polytechnic Jalgaon, BE (CSE) from Government College of Engineering Aurangabad, ME (CSE) from Walchand College of Engineering Sangli.Presently I am working as Assistant Professor in Information Technology Department at Government College of Engineering, Amravati.

Satish Kolhe received his B.E. degree from the Amaravati University, Amaravati, in Computer Engineering in 1991; M.Tech. degree from the Dayalbagh Educational Institute, Dayalbagh, Agra, in Engineering Systems in 1994 and Ph.D. in Computer Engineering from North Maharashtra University, Jalgaon in 2007. He is presently Associate Professor in the Department of Computer Science, North Maharashtra University, Jalgaon where he has been involved teaching and research in the field of artificial intelligence and neural networks for over a decade.
His current research interests are in the areas of artificial intelligence, soft computing, image processing and w eb mining. He has authored and c o-authored over 35 p apers in referred academic journals and nat ional/international conference proceedings. He is a Fellow life member of Institute of Electronics and Telecommunication Engineers (FIETE), India, Life Member of Computer Society of India (CSI), India, Linguistic Society of India (LSI), and Indian Science Congress Association (ISCA), Calcutta. He is also a member of Special Interest Group in Artificial Intelligence (SIGAI), India.