Efficient Information Retrieval Using Measures of Semantic Similarity

Krishna Sapkota Laxman Thapa Shailesh Bdr. Pandey

Abstract—**The semantic information retrieval (IR) is pervading most of the search related vicinity due to relatively low degree of recall or precision obtained from conventional keyword matching techniques. Such techniques miss to retrieve semantically or lexically related terms that are not explicit in the query. In this paper, we present a search engine framework using Google API that expands the user query based on similarity scores of each term of user's query. We calculated the semantic similarity of noun words to obtain the related concepts described by the search query using WordNet as knowledge source. Users query was replaced with concepts discovered from the similarity measures and fed to the Google search API that resulted in efficient document retrieval.**

***Index Terms*—Search, Semantic Similarity, Query, WordNet**

I. 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 are 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 Jiang & Conrath [2], Lin [3], and Resnik [4] approach to calculate similarity between two concepts in the taxonomy to discover the related concepts, which are not implicit in the query. Jiang & Conrath gives the best result overall as evaluated by Budanitsky & Hirst [5, 6]. For example a search 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. Results obtained for search data are given in section 5 of this paper.

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 was computed. The pair of concept that has higher similarity value was chosen as the concept described by the words. This discovered concept was used to supplement users query with its synonyms and hypernyms class based on relatedness score. In section II various approaches to calculate semantic similarity are given. In section III overall system architecture is presented with brief description to each module. In section IV various result conducted during the development of system are presented. Finally, in section V and VI further works and conclusion are discussed.

Manuscript received July 31, 2006

Krishna Sapkota is a final year student of Bachelor of Engineering in Information technology at Nepal Engineering College, Changunarayan, Bhaktapur, Nepal. He is also a research student to the Language Technology Group, Center for Research in Social Defense Technology, Nepal Engineering College. (Corresponding author, phone: 00977-9803046628, e-mail: krishnasatch@yahoo.com).

Laxman Thapa is a final year student of Bachelor of Computer Engineering at Nepal Engineering College (Phone: 00977-9803041708, e-mail: lxnthapa@hotmail.com)

Shailesh Bahadur Pandey is Teaching Assistant in Department of Computer Engineering, Nepal Engineering College. Er. Pandey is also a research affiliate to the Language Technology Group, Center for Research in Social Defense Technology, Nepal Engineering College. (Phone: 00977-1-6611744 fax: 00977-1-6611681; e-mail: shaileshpandeynec@yahoo.com).

I. Semantic Similarity

A. WordNet

We used WordNet [1] 2.0 as the taxonomy to calculate semantic similarity between words. WordNet[1] is an online English lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets.

Various kinds of relationship exist in the WordNet taxonomy, which can be categorized as semantic and lexical relationship. The former relation holds between the concepts where as later one between word forms. Examples of semantic relations are hypernym/hyponym (IS-A), meronym (HAS-A) and lexical relations are synonymy, antonymy. Since usually the search keywords are Nouns, we are going to cover only

[1] WordNet is an electronic lexical database created by WordNet was developed at Princeton University by Miller *et al.* [Mil95]; it is available on line at http://wordnet.princeton.edu

the noun network, which accounts for close to 80% of the relations.

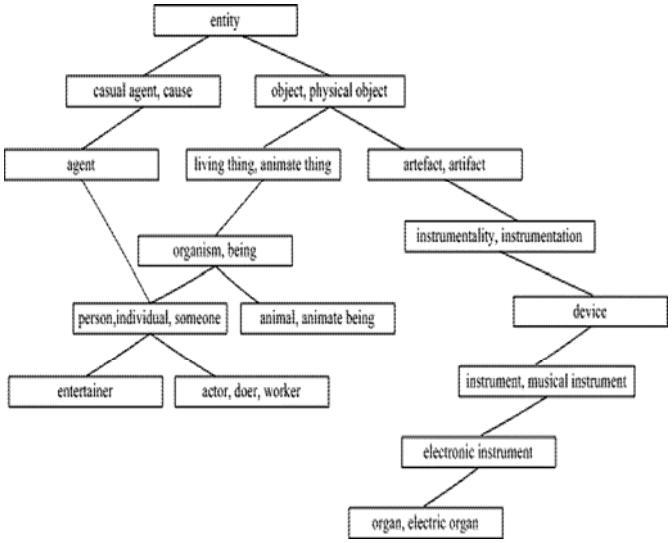

Fig. 1. WordNet Hypernyms

The hypernym, hyponym relations are particularly interesting. These relations form IS-taxonomies with the noun synsets. These taxonomies have a tree-like structure. All noun synsets belong to at least one taxonomy. Because multiple inheritances are allowed, some synsets belong to more than one taxonomy. There is no unique root node that links all noun synsets together. Instead, there are multiple taxonomies. In WordNet 2.0, there are nine noun taxonomies.

B. Similarities and Relatedness

Semantic similarity applies to the words of same syntactic category where as semantic relatedness applies to words of all parts of speech. Only noun-noun pairs and verb-verb pair can be easily classified into is-a hierarchy. Semantic relatedness can be computed in other than is-a hierarchy such as meronym (part-of) and holonym. However both similarity and relatedness signifies the notion of closeness between meaning of different words, for example human and plant are more similar than are human and chair, since the former one share the common class 'living thing' in a semantic network IS-A hierarchy. Thus the similarity information can be extracted from any lexical database such as WordNet that organizes concepts in hypernyms/hyponyms relation. Various approaches have been proposed for computing the semantic similarity. We discuss the integrated approach of semantic network and information content method to compute similarity of words.

C. Similarity Using Path Length

Semantic relatedness can be computed by simply counting the length of the path or node between the concepts. Resnik (1995), found that "the shorter the path from one node to another, the more similar they are"

In an is-a taxonomy such as WordNet, a simple approach to measuring similarity is to treat the taxonomy as an undirected graph and use the distance in terms of path length between the two synsets as a measure of similarity. The greater the distance between two synsets, the less similar they are. In figure 1, for example, the synset {actor, doer, worker} is closer to {person, individual, someone} than it is to {organism, being} so it is considered to be more similar to {person, individual, someone} than {organism, being}. The distance may be calculated by counting node or edge. Here the similarity between {actor, doer, worker} and {organism, being} counting the node is 3 and counting the edge the similarity value is 2.

D. Information Content Approach

The key idea underlying Resnik's (1995) approach is the intuition that one criterion of similarity between two concepts is "the extent to which they share information in common", which in an IS-A taxonomy can be determined by inspecting the relative position of the most-specific concept that subsumes them both. Information content is a measure of specificity. The information content of a concept is inversely proportional to the frequency with which the concept is expected to occur. A concept that rarely occurs would have high information content, and a concept that frequently occurs would have low information content. Information content of a concept can be mathematically expressed by:

$$IC(c) = -\log P(c) \tag{1}$$

Where c is concept represented by a synset in WordNet and P(c) the probability of encountering an instance of concept c in the corpus. The probability of a concept is the frequency of the concept divided by the number of concepts occurring in a corpus[2]:

$$P(c) = frequency(c) / N \tag{2}$$

Here, N is the number of concepts in the corpus from which the frequency counts were extracted. Frequency counts are propagated up the hierarchies so that the count of a concept is equal to the sum of the counts of its hyponyms plus the count of the concept itself.

1) Resnik Resnik [4] proposed a simple information content approach to calculate the semantic similarity as the information content of Lowest Common Subsumer (LCS) of two concepts as expressed by equation 3.

$$Sim_{res}(c_1, c_2) = IC(LCS(c_1, c_2)) \tag{3}$$

In cases where there is more than one subsumer of c_1 and c_2, the LCS is defined as the common subsumer with the

[2] We used the Semantic Concordance, the sense-tagged subset of brown corpus for our experiment. All of 186 files of SemCor were converted to MySQL database for use in our system. The SemCor files can be downloaded at http://www.cs.unt.edu/~rada/download

greatest information content. One drawback that Resnik's measure has is the same information content value for all concepts, which have the same LCS. For example, synset {object, physical object} is the LCS of many synsets such as {living thing, animate thing} and {artifact, artifact}, {organism, being} and {device}, even {animal, animate being} and {organ, electric organ. All of these pairs similarity would be same though every pair does not seem to be similar as evidence from simple path length would suggest different value for these pairs.

2) Lin: Another similarity measure by Lin [3] takes an information-content approach based on three assumptions. Firstly, the more similar two concepts are, the more they will have in common. Secondly, the less two concepts have in common, the less similar they are. Thirdly, maximum similarity occurs when two concepts are identical. The measure of similarity expressed by equation 4 meets these assumptions.

$$Sim_{in}(c_1, c_2) = 2 * IC(LCS(c_1, c_2)) / IC(c_1) + IC(c_2) \quad (4)$$

The information content of the LCS will always be less-than or equal-to the information content of both c_1 and c_2; therefore, the similarity score can be at most one. The score is zero only if the information content of the LCS is zero. The score is undefined if the information contents of c_1 and c_2 are zero.

3) Jiang-Conrath: Jiang and Conrath's [2] idea was to synthesize edge and node based techniques by restoring network edges to their dominant role in similarity computations, and using corpus statistics as a secondary, corrective factor. This approach 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.

$$Dist_{jcn}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 * IC(LCS(c_1, c_2)) \quad (5)$$

This distance measure can be converted to a similarity measure by taking the multiplicative inverse of it:

$$Sim_{jcn}(c_1, c_2) = 1 / Dist_{jcn}(c_1, c_2) \quad (6)$$

Budanitksy and Hirst found that Jiang and Conrath measure gives the best result overall when compared their result to human similarity judgment.

II. Search Engine Framework

The system architecture proposed in this paper is shown in Fig. 2. This framework utilizes the existing search engine to process the user's query. Each module was implemented using PHP object oriented features. Various processing are carried out utilizing different resources to expand the query [7]. As shown in Fig. 2, the system consists of:

- Query input module
- Similarity computation module
- Query expansion module
- Search engine

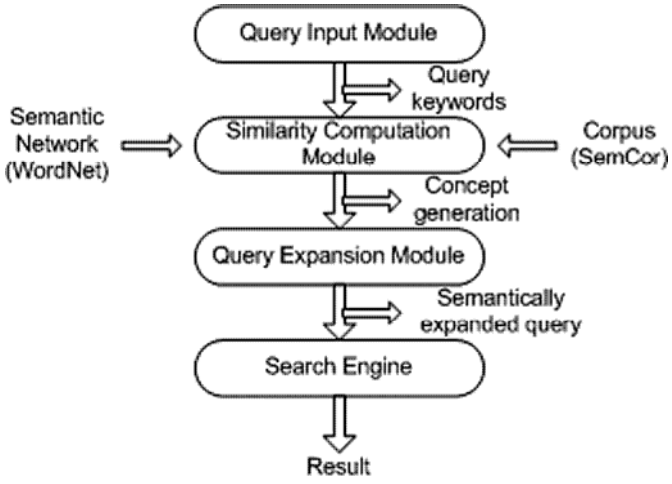

Fig. 2. System Architecture

A. Query Input Module

The system proposed in this paper requires specifying some nouns as search keyword. This is because that we only calculated the similarity in IS-A hierarchy of noun network and that search queries are more likely to be noun. This module tokenizes the input query as required by the similarity computation module. It may be that some words fall in more than one syntactic category. For example, interest can be both noun and verb, the sense of interest in noun has a meaning "a sense of concern with and curiosity about someone or something" and verb interest is "excite the curiosity" .In such case only the noun syntactic category is considered for similarity calculation.

B. Similarity Computation Module

This module computes the semantic similarity with the information content approach. All of three Resnik [4], Lin [3] and Jiang-Conrath [2] information theoretic approach are used for calculating the similarity between words. Any similarity value of these measures may be used to generate concepts from the query keywords. The similarity scores between any given two word is computed by following expression

$$Sim(w1, w2) = \underset{c1 \sum s(w1), c2 \sum s(w2)}{MAX} [Sim(c1, c2)] \quad (7)$$

Where s (w_i) is "the set of concepts in the taxonomy that are senses of word (Resnik). That is, the similarity of two

words is equal to that of the most-related pair of concepts that they denote.

When the search query "bank interest" is presented this module computes the similarity score of concepts represented by each term bank and interest. For the sake of simplicity let us consider that the word bank has two senses bank1-"A financial institution that accepts deposits and channels the money into lending activities and bank2-"Sloping land (especially the slope beside a body of water)" and the word interest has two senses interst1-"a sense of concern with and curiosity about someone or something" and interest2-"a fixed charge for borrowing money". Similarity between each possible pair (bank1-interest1, bank1-interest2, bank2-interest1, and bank2-interest2) of concepts of bank and interest is computed. The pair of concept which has the highest score (in this case bank1-interest2) is input to the next query expansion module.

C. *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. 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 synonyms and hypernyms for the concept bank1 are {depository financial institution, bank, banking concern and banking company} and {financial institution, financial organization, financial organization} respectively. So here we replace the 'bank' with sets of both synonym excluding the term bank and its Hypernyms. Similarly 'interest' is replaced with its respective synonym and hypernym.

D. *Search Engine*

In the system a WWW search engine accepts the noun, which is generated by the query expansion module, as an extra query keyword in addition to the ones specified by the user. We used google search engine to supplement the query with our analyzed terms. Google[3] API was accessed using the SOAP architecture, which allows invocation of the remote object. We used the NuSOAP: the implementation of SOAP in PHP developed by NuSphere1[4] Corporation.

III. EXPERIMENTS

A. *Similarity Score*

Three of all information content measures were implemented to get the similarity score of word pairs. Among these, the Jiang-Conrath measure was found to be most promising when compared to human judgment [3]. The human rating for the similarity judgment was performed by Rubenstein–Goodenough for about 51 word pairs. They were asked to rate them from scale 0 to 4, according to their "similarity of meaning". Some rating of these pairs is given below in Table I [6].

TABLE I
HUMAN AND COMPUTER RATINGS OF THE RUBENSTEIN–GOODENOUGH SET OF WORD PAIRS

Word Pairs	Humans	Sim_{res}	Sim_{lin}	$Sim`_{jcn}$
fruit-furnace	0.05	1.85	0.14	0.05
monk-slave	0.57	2.53	0.21	0.05
coast-hill	1.26	6.19	0.53	0.09
magician-oracle	1.82	13.58	0.96	1.00
brother-lad	2.41	2.53	0.23	0.06
food-fruit	2.69	1.50	0.22	0.09
furnace-stove	3.11	1.85	0.13	0.04
boy-lad	3.82	8.29	0.72	0.18
automobile-car	3.92	8.62	1.00	maximum

B. *Search Results*

TABLE II
FIRST 10 RESULTS FROM GOOGLE SEARCH FOR 'BANK INTEREST'

Topic Retrieved	WWW Address
The National Neopian Bank	http://www.neopets.com/bank.phtml
Bankrate.com	http://www.bankrate.com
Certificate of deposit of interest rates: Compare the best rate	http://www.bankrate.com/brm/rate/deposits
Bank Interest Calculator	http://www.digita.com/tisali/calculators/bankinterestcalculate/
Personal Banking System, savings, bank interest rate, tax	http://www.thisismoney.co.uk/saving
Indian Bank-Interest rates	www.indian-bank.com/interest.htm
National Australian Bank	www.national.com.au/business-solution/02253300.htm
Bank Interest	http://www.ato.gov.au/content/48327.htm
Infochoice Banking	http://www.infochoice.com.au/banking/default.asp
National Bank Interest Rate Graph	http://www.nbnz.co.nz/economics/interest/

The tables given in experiments section shows the result obtained by the system for a search query bank interest. Table II shows the result with the user's original query and Table II shows the result with optimized query. We see that in Table II the results have same words as described in query and in Table III the results show are of implicitly hidden concept of bank interest.

[3] Information about accessing Google API can be found at http://www.api.google.com

[4] This SOAP library is available for download at http://www.nusphere.com/

TABLE III
FIRST 10 RESULTS FROM OUR EXTENDED GOOGLE FOR 'BANK INTEREST'

Topic Retrieved	WWW Address
Scholarly articles (for all expanded term)	http://www.google.com
Bloomberg.com: Financial Glossary	http://www.bloomberg.com/invest/glossary/bfglost.hym
FCAC-Glossary	www.fcac-acfc.gc.ca/eng/glossary.asp
Guide to organize a new state bank in Florida	http://www.flofr.com/banking/howtoorg.htm
PSI- Performance Solution International	http://www.goto-psi.com/glossary.htm
[pdf] How should financial institution and market should be structured	http://www.iadb.org/res/publications/pubfiles/
Women's wallstreet.com – glossary	http://www.wimenswallstreet.com/tools-resources/glossary/f.htm
Operational risk poses challenges to financial institution	http://knowledge.wharton.upenn.edu/article.cfm?articleid=582
FDIC:FDIC Banking Review	www.fdic.gov/bank/analytical/banking/2004nov/article1/index
BKD ,LLP- Financial Services	http://www.bkd.com/industry/financial-services

IV. POSSIBLE ENHANCEMENT AND FURTHER WORK

The system 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 [8] for the generation of the concepts required by the query expansion module. This method works better when the words in the query are more to sufficiently describe the context used. As different module was developed for this system using object-oriented features of PHP, incorporating such method is easy. Furthermore the system can be extended to make it work on any part of speech, this requires part of speech (POS) tagger to assign syntactic category to the words and computing relatedness from other than in is-a relationship such as meronymic (part-of) relations.

More over the system can be mapped to work on semantic similarity of Nepali words building a Nepali language search engine interface as soon as Nepali corpus and semantic network are developed utilizing the same existing module.

V. CONCLUSION

In the development of the system various resources such as WordNet, Semantic Concordance (SemCor) was utilized. We used the WordNet 2.0 prolog MySQL version for similarity calculation as mentioned. Since the SGML tagged files of SemCor were not consistent to use with MySQL version of WordNet, we converted the semantic concordance files to MySQL format. This conversion has potential benefit for the use in various statistical analyses that can be done just by a SQL query.

We have found that replacing query with terms 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. We showed the process to overcome problem when users are not intelligent to describe the query efficiently. We did this by expansion of the original query with semantically related terms and by omitting user's original query. The results obtained were promising in information retrieval.

REFERENCES.

[1] Miller G.A., Beckwith R., Fellbaum C., Gross D. & Miller K. *Five Papers on WordNet*. CSL Report 43, Cognitive Science Laboratory, Princeton University, pp 1-25 July 1990.

[2] Jay J. Jiang and David W. Conrath. 1997. Semantics and Similatity Based on Corpus Statistics and Lexical Taxonomy. *In Proceedings of International Conference on Research in Computational Linguistics*, Taiwan.

[3] Lin. 1998. An information-theoritic definition of similarity. In *Proceedings of the international Conference on Machine learning*, Madison, August.

[4] Philip Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal.

[5] Budanitsky A., Graeme H.: Semantic distance in WordNet: an Experimental, Application oriented Evaluation of Five Measures. *Proceedings Workshop WordNet and Other Lexical Resources.* The North American Chapter of the Association for Computation Linguistics (NAACL), Pittsburgh, PA, (2001)

[6] Budanitsky A., Graeme H.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness *Association for Computational Linguistic*, pp 4-14, 2006

[7] Complement Keywords for Query toward Efficient Information Retrieval. In *1999 IEEE International Conference on Systems, Man, and Cybernetics* (SMC'99), 1999

[8] Pederson T., Banerjee S., Patwardhan S. Maximizing Semantic Relatedness to Perform Word Sense Disambiguation, Technical Report UMSI 2005/25, University of Minnesota supercomputing institute, March 2005.

[9] Miller, G., C. Leacock, R. Tengi, and R.T. Bunker, 1993, "A Semantic Concordance", *Proceedings of ARPA Workshop on Human Language Technology*, 303-308, March 1993