Semantic Similarity Measure Using Information Content Approach With Depth For Similarity Calculation

Atul Gupta, Dharamveer kr. Yadav

Abstract: Similarity is criteria of measuring nearness or proximity between two concepts. Several algorithmic approaches for computing similarity have been proposed. Among the existing Similarity measure, majority of them utilize WordNet as an underlying ontology for calculating semantic similarity. WordNet is a lexical database for English Language which was created and maintained by Congnitive Science Laboratory at Princeton University under the supervision of Professor George A. Miller. It is organized as a network which consists of concepts or terms called Synsets (list of synonyms terms) and the relationship between them. There are different type of relationship exists in WordNet such as is-a, part-of, synonym and antonym. It has thdatabases, one for noun, one for verb and one for adverb and adjective. This project work proposes a metric for semantic relatedness calculation between pair of concepts which uses Tversky's feature based approach which takes into account the common and distinct feature of the two terms or concepts. If commonality is more as compared to differences the similarity between concepts is high otherwise similarity is low. Tversky's theory is quantified by information content of two concepts and the Information content of most specific common ancestor of two concepts. As we move down in the WordNet hierarchy, more specific and more Informative concept are there, where as when we move up in the hierarchy more Generalized and less Informative concepts are there. So depth of a concept in the WordNet hierarchy is a critical factor in similarity calculation. We take into consideration the depth of the specific concept in the WordNet hierarchy which is the deciding factor for determining the relevance of distinct feature specific to a concept in similarity calculation. Introduction of depth reduces the impact of the less relevant dissimilarity indulge in similarity calculation thereby increase precision. We carried out our experiment of 28 word–pair common to Rubenstein-Goodenough and Millers-Charles set. These word-pair range from low similarity, intermediate similarity and finally to high similarity pairs. Evaluation is done by calculating our similarity values calculated using the proposed measure with the human rating. We utilize Pure Java Wordnet Similarity Library for implementing our proposed metric. Experimental results shows that the proposed metrics is at par with the existing similarity measure and superior to some of the traditional ones.

I. INTRODUCTION

A. Semantic Similarity

Semantic Similarity or semantic relatedness is a concept of measuring closeness between set of terms or document in context of their meaning. We have two different methodologies for calculating semantic similarity, one by defining a topological similarity, using ontology to define a distance between words or using statistical means such as vector space model to correlate words and textual contexts from a suitable text corpus. We focus on the former approach using WordNet ontology for semantic similarity calculation. Similarity calculation in this approach relies on the fact that similarity is dependent on both common and distinct features of the objects.

- *Atul Gupta: Department of Computer Science & Engineering, G. L. Bajaj Institute of Technology and Management, Greater Noida, UP, India, atul.gupta@glbitm.org, atul.cdacnoida@gmail.com*
- *Dharamveer kr. Yadav: Department of Computer Science & Engineering, G. L. Bajaj Institute of Technology and Management, Greater Noida, UP, India, dharamveer.yadav@glbitm.org*

Another approach introduced by Resnik, Information content based approach capture the Informative part of the concept; high information content implies more relevance and specific to the subject than the lower one. Path length based approach measure similarity as a function of distance between concepts in the ontology. Hybrid approach is a combination of the similarity measure mentioned above. Parameters Length, depth and local density forms a part of nonlinear function which measures similarity between concepts.

B. Word Net

WordNet is lexical ontology for English language. It models the lexical Knowledge into a taxonomic hierarchy. WordNet contains three databases one for nouns, one for verbs and one for adverb and adjectives. Terms and concepts are organized into synsets (list of synonyms terms or concepts). We have considered only **is-a** relationship and **noun** concepts in the wordnet hierarchy for similarity calculation. We use **WordNet 2.0** which contains nine separate noun hierarchies containing, path between two concepts may not exists in the wordnet. So we create a root node that subsumes all the nine given hierarchies in the WordNet. The measure of semantic relatedness given in this thesis focus on is-a relationship between noun concepts in WordNet.

C. Motivation

Semantic Similarity calculation is useful in several emerging research areas such as Semantic Information retrieval, artificial intelligence, biomedicine and psychology. There are various approaches for estimating semantic similarity and based on those approaches there are number of

semantic similarity measures. Since semantic similarity plays critical role in application like improving accuracy of information retrieval, to perform word sense disambiguation, to discover mapping between ontology's and in various application of artificial intelligence. It is a challenging task to find out a measure close to human similarity ratings and highly accurate. Existing similarity measure, which uses WordNet ontology for determining semantic similarity, are accurate to some extend but sometimes fails on highly similar wordpair. Wordpair which are highly similar, are crucial than those who are partially similar or completely dissimilar. For example in semantic information retrieval task, we assign score to available pages based on similarity between user query and the content present in the pages available on the Web. If similarity measure indulges in similarity calculation, is inaccurate on high similar wordpair then the retrieved result is out of the user's context. The problem of highly accurate similarity measure need to be tackled and hence new similarity measure in required.

II. Type Background and Related Work

Semantic relatedness measure for calculating semantic similarity between terms represent in WordNet ontology. Semantic measures used for performing tasks such as term disambiguation (e.g. a user needs the explanation or definition of a term) as well as retrieving Information to user queries. A huge volume of early literature is present in the area of similarity calculation with numerous number of similarity measure, but we classify these approaches into four major approaches; distance based approach, Information content based approach, feature based approach and hybrid approaches. Edge counting based semantic similarity measure considers the path length between the two concepts in the ontology hierarchy for estimating semantic similarity score. Wu and Palmer and Leacock and Chodorow are the similarity measure which uses path length between concepts for semantic similarity calculation. We discuss in details in next section. Information content based semantic similarity measure compute the information content specific to a concept and the information content of the most specific common ancestor of the two concepts, which represents the information shared by the two concepts. Feature based approach semantic similarity takes into account the features of the concept in the ontology. It focuses on the common and distinct feature specific to a concept for similarity calculation. Tvesky's, P &S and FAITH uses feature based approach in similarity calculation. Hybrid based measures utilize the combination of more than one approaches discussed above for determining similarity between the concepts. Lin approach uses Multiple Information sources to calculate similarity.

A. Edge Counting Based Similarity Measure

Leacock-Chodorow

Leacock and Chodorow [10] measure is based on the shortest path between the noun concepts in a WordNet is-a hierarchy, scaled by maximum depth.

$$Sim(c1,c2) = -log\ \frac{Shortestpath(c1,c2)}{2*D}$$

Shortest length (c1, c2) is the shortest path length (include minimum number of Intermediate concepts and D is the maximum depth in the hierarchy).Since the measure takes into account the depth of the hierarchy, behavior of the measure is deeply affected by the presence and absence of the unique root node. If root node has been used then it is possible for a synset to belong to more than one than one taxonomy. As there are 9 different taxonomies in absence of root node.

Wu and Palmer Measure

Wu and Palmer [12] similarity measure calculate the most specific common ancestor of the two Concepts, with minimum number of is-a Link in the path of the common subsume.

$$Sim = \frac{2*h}{h1+h2+h}$$

Here h is depth of the subsume from the root of the hierarchy, h1 and h2 is the minimum number of is-a link from concept c1 and c2 to the most specific common subsumer. It scores between 1 and 0.

Shortest Path Measure

Shortest path measure [13] focuses on the closeness of two concepts in the hierarchy.

$$Sim = 2*MAX - l$$

Where MAX is the maximum path length between two concepts in the taxonomy and L is the minimum number of is-a link between concepts c1 and c2.

B. Information Content Based Similarity Measure Resnik similarity

Resnik [1] measure is based on the fact that semantic relatedness between two concepts is directly related to the amount of information they shared in common. More information they shared in common, more is the similarity between them. The shared information is determined by Information content of the most specific subsume of the two concepts in the hierarchy.

$$Sim = IC(lcs(c1,c2))$$

Information content of a concept is calculated by counting the frequency of that concept in the corpus and thus determining the probability of encountering an instance of that concept.

$$IC = -\log(p(c))$$

Frequency of a concept includes the frequency of its entire subordinate concept as the count of specific concepts is added to its subsuming concepts as well.

Jiang and Conrath Similarity Measure

Jiang and conrath[14] approach captures the Information content of the two concepts along with the Information content of most specific common subsumer. It basically calculates the distance between two concepts.

$$Distance = IC(c1) + IC(c2) - 2 * IC(lcs(c1, c2))$$

$Distance_{JCn}$ measure give the measure of un-relatedness between the two concepts, high score indicate low similarity and low score indicate high similarities.

Lin Similarity Measure

Lin[2] measure capture semantic relatedness between two concepts as the ratio of amount of information shared between two to the total amount of Information possessed by the two concepts.It uses both the amount of Information needed to state the commonality between two concepts and the information needed to fully described them.

$$Sim = \frac{2 * IC(lcs(c1, c2))}{IC(c1) + IC(c2)}$$

Commonality of the concepts is determined by the information content of most specific common and the Information content of the two concepts. The value of this similarity measure varies between 0 and 1.In this measure a term compared with itself always scores 1, hiding the information revealed by the resniks measures.

C. Feature Based Similarity Measure Tversky's similarity Measure

Tversky [4] measure is based on the description set of terms. We suppose that each term is described by the set of words indicating its properties and features. Feature in common to both the concepts increase similarity and feature unique to the specific concept decrease similarity between two terms. According to him, similarity between two concepts c1 and c2 depends on the common feature to c1 and c2, those are in c1 and not in c2, those are in c2 and not in c1.If a function ψ(c) denotes all features of concept c. Tversky model is represented by the following equation.

$$Sim = \alpha.F(\psi(c1) \cap \psi(c2)) - \beta.F\left(\frac{\psi(c1)}{\psi(c2)}\right) - ¥.F\left(\frac{\psi(c2)}{\psi(c1)}\right)$$

α, β, ϒ are the sets of parameter to focus on different components.

P&S Similarity Measure

P&S [6] similarity measure exploits the tversky approach and maps into Information theoretical domain. In the above equation, F quantifies the salient set of features and quantification is done in the form of information content which implies that F(ψ(c1)∩ψ(c2) is equivalent to IC(lcs(c1,c2)),ψ(c1)/ψ(c2) to IC(c1)-IC(lcs(c1,c2)) and ψ(c2)/ψ(c1) to IC(c2)-IC(lcs(c1,c2)).

$$Sim = 3 * IC(lcs(c1,c2)) - IC(c1) - IC(c2) \;\; if(c1 \neq c2)$$

1 Otherwise

Here IC is the information content between concept c1 and c2.

III. Proposed Approach

Depth of a concept has deep concern with similarity between concepts in the hierarchy. As we move down in the WordNet hierarchy, concepts are more specific and as we upward concept are more generalized. So concept at higher depth is more significant as compared to the concept at lower depth. According to Tversky's[4], both common and distinct feature, are significant in estimating similarity between concepts. Focusing on the distinct feature of the two concepts, one at higher depth is more significant as compared to concept at lower depth. Impact of the concept toward differences at higher depth will be more as compared to concept at lower depth. We utilized Resnik's information content and Tversky's feature based approaches for similarity calculation. We introduce new parameters α and β which takes into account the relevance of distinct feature specific to each concept on the basis of the depth.

$$Sim(new) = 3.Ic(lcs(c1,c2) - \alpha.IC(c1) - \beta.IC(c2) when\ c1 \neq c2$$

$$\alpha = \frac{depth(c1)}{depth(c1) + depth(c2)} \quad \beta = \frac{depth(c2)}{depth(c1) + depth(c2)}$$

If both the concept are at the same level then value of α=β which implies that distinct feature of both the concepts are equally significant for calculating similarity. If the difference between the depth of the concepts is high, α>β or β>α then concept are at higher depth will be more significant towards contributing differences to similarity. In our similarity measure, we introduce relevance of distinct feature, specific to the concept, on the basis of its depth in the hierarchy. With the introduction of parameter α and β, proposed semantic similarity measure shows improvement in term of accuracy especially for highly similar wordpairs in the data set. As we reduce the relevance of distinct feature specific to concept on the basis of their depth in the taxonomy, the most specific common ancestor between two concept that represents the commonality between two concepts find the weightage over the distinct feature in case of highly similar values, as values of msca is high. Unlike measure like Lin and Jiang and Conrath which focuses either on commonality or differences of the two concepts. The relevance of distinct feature specific to a concept, is decided by the basis of depth of that concept.

IV. Evaluation and Results

This section summarize the outcome of our approach for estimating similarity and proved the effectiveness by comparing our approach with the existing similarity measures. We evaluate our proposed similarity by setting up an experiment on the test set consist of wordpairs. Similarity calculation is performed on the test set and compared with human rating.

A. Implementation

We carried out our experiment with 28 common word pairs out of 65 present in Rubenstein and Goodenough set and 30 in Miller Charles set. We use Pure Java WordNet Similarity Library developed by Mark A. Greenwood for implementing our proposed semantic measure. Similarity library include semantic measure given by Lin and Jiang and Conrath. Table 1: explains the detailed experimental results and list the similarity values corresponding to each wordpairs computes separately for all the three similarity measure. Jiang and Conrath, Lin and our proposed

measure exploits information content based approach. Correlation coefficient is calculated using the formula:

$$r(x,y) = \frac{n\sum(xi * yi) - \sum xi * \sum yi}{depth\sqrt{n\sum xi2 - (\sum xi)2} \quad \sqrt{n\sum yi2 - (\sum yi)2(c1)}}$$

Where,
n=Total no of word pairs in data set.
x_i=Human rating of i^{th} word pair.
y_i=Rating of similarity measure of i^{th} word pair

B. Results

Correlation coefficient between computed similarity values and human rating of Rubenstein Goodenough is calculated to judge the suitability of the proposed measure to the existing similarity measure. Table 2 summarizes the correlation between human rating in RG experiment and the computed semantic similarity. Correlation coefficient of all the three similarity measure is calculated with Human judgement, which brings out the comparison among the similarities measures. The result shows that our proposed measure with r=0.8069, closely resembles the Human judgment in comparison to the other considered measure.

TABLE I. SIMILARITY RESULTS FROM DIFFERENT MEASURE

Word Pair	R&G Ratings	Sim_{NEW}	Sim_{JCn}	Sim_{Lin}
Cord-Smile	0.02	0.0	0.058	0.0
Rooster-Voyage	0.04	0.0	0.0453	0.0
Noon-String	0.04	0.0	0.04547	0.0
Glass-Magician	0.44	0.0	0.04795	0.062
Monk-Slave	0.57	0.0	0.07112	0.248
Coast-Forest	0.85	0.0	0.05219	0.0
Crane-Rooster	1.41	9.009	0.0494	0.179
Lad-Wizard	0.99	0.0	0.07243	0.2514
Forest-Graveyard	1.00	0.0	0.05137	0.0
Mound-Shore	0.94	7.892	0.0614	0.137
Coast-Hill	1.26	8.927	0.140	0.6365
Car-Journey	1.55	0.0	0.0661	0.0
Crane-implement	2.37	0.9031	0.0619	0.1385
Hill-Mound	3.29	20.6107	0.0582	0.1311
Bird-Crane	2.63	11.103	0.0658	0.2252
Bird-Cock	2.63	11.250	0.0693	0.0884
Food-Fruit	2.69	0.0	0.0901	0.1119
Brother-Monk	2.74	18.3747	0.0686	0.2413
Asylum-Madhouse	3.04	17.758	0.0698	0.3708
Furnace-Stove	3.11	0.0	0.0652	0.2517
Magician-Wizard	3.21	22.240	0.0558	0.2144
Journey-Voyage	3.58	12.759	0.2363	0.7553
Coast-Shore	3.60	18.724	1.5948	0.9681
Implement-tool	3.66	12.047	1.4029	0.9457
Boy-Lad	3.82	16.516	0.1545	0.6363
Automobile-Car	3.92	15.501	5.1472	1.0
Midday-Noon	3.94	21.236	5.1472	1.0
Gem-Jewel	3.94	23.080	0.05	0.07163

TABLE II. CORRELATION OF DIFFERENT MEASURE AGAINST HUMAN JUDGMENT

SEMANTIC SIMILARITY MEASURE	CORRELATION(r)
Jiang and Conrath	0.4561
Lin	0.7138
Proposed Measure	0.8069

V. AND FUTURE WORK

A. Conclusion

We present an approach for calculating semantic similarity with inclusion of depth in the feature based approach mapped with information content. Among the existing measure, this approach is promising in term of accuracy with the compared measure and outperforms with r=0.8069. We implemented our similarity measure in Pure Java WordNet Similarity Library and carried out our experiment with 28 common words pairs out of 65 present in Rubenstein-Goodenough and 30 in Miller-Charles set. We observed that our similarity measure gives accurate similarity results on highly similar and highly dissimilar word pairs.

B. Future Work

We have only considered hypernym, is-a relationship between noun concepts in WordNet hierarchy, we can generalize the experiment by including all type of relationship present in WordNet. We have calculated similarity between wordpair in single ontology. We can extend our work on cross ontology. Determining similarity between wordpair that belong to different ontologies.

REFERENCES

[1] P.Resnik,Information content to evaluate semantic similarity in taxanomy.In proceeding of IJCAI,pp.448-453,1995

[2] Y.Li,Bandar,and D.McClean,An approach for measuring Semantic Similarity between Words Using Multiple Information Sources.IEEE Transaction on Knowledge and Data Engineering,15(4):871-882,2003.

[3] Guisepee Pirr,Jrne Euzant A feature and Information Theoritic Framework for Semantic Similarity and Relatedness.In proceedings of International Semantic Web Conference(1)2010.pp615-630.

[4] A.Tversky, Feature of Similarity, Psychological Review, 84(2):327-352, 1997.

[5] E.L Rissland, AI and Similarity, IEEE Intelligent Systems, 21:39-49, 2006.

[6] Piarro, G.A: Semantic Similarity Metric Combining Features and Intrinsic Information Content .Data Knowledge engg, 68(11),pp.1289-1308, 2009.

[7] Miller, G.A Wordnet an on-Line Lexical Database International Journal of lexicography,3(4),pp.235-312, 1990.

[8] H.Rubenstein and J.B Goodenough, Contextual Correlates of Synonymy, ACM, vol.8, pp.627-633, 1965.

[9] G.A Miller and W.G.Charles, Contextual Correlates of Semantic Similarity, Languages andCongnitive Processes, vol6, no. 1, pp1-28, 1991.

[10] Claudia Leacock and Martin Chodorow.Combining Local Context and WordNet Similarity for word Sense identification. In [10], 1998.

[11] Mark A.Greenwood, The University of Sheffield http://nlp.shef.ac.uk/result/software.html.

[12] Z.Wu and M.Palmer.Verb semantic and Lexical Selection .In Proceddings of 32 Annual Meeting of the association of computer Linguistics(ACL 994),Las Cruces,New Mexico,1994.

[13] R.Rada.Development and Application of a metric on Semantic Nets.IEEE Transaction on Systems, Man and Cybernetics, 19(1):17, 30 January.

[14] J.J.Jiang and D.W Conrath, Semantic Similarity Based on Corpus Statistics and Lexical Taxanomy, Procs.ROCKLING X, 1997.

[15] M. Young, the Technical Writer's Handbook. Mill Valley, CA: University Science, 1989