ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Combined Information Content and Path Length Based Semantic Similarity Measurement

Dr. Kishor Wagh

Assistant Professor, Department of Information Technology, Government College of Engineering, Amravati, 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, Natural Language Processing (NLP), web link mining, databases and information retrieval. Web contains very large amount of information, which are scattered and dynamic as well as diverse in terms of contentand 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. The system uses the Lin, Resnik and new approach to calculate semantic similarity between two concepts in the taxonomy to discover the related concepts, which are not implicit in the query by using WordNet ontology. 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. In this paper, a search engine framework using Google API that expands the user query based on similarity scores of each term of user's query. The semantic similarities of word pairs and correlation coefficient between human judgment and computational measures are calculated. The experimental result shows Combined Information Content and Path Length Based Semantic Similarity (CICPLBSS) is better than other existing computational models. The correlation coefficient value of CICPLBSS is 0.8449. Two datasets are used for semantic similarity measurement. Search engine framework retrieves the web documents using semantic similarity information.

KEYWORDS: Information theory, Semantic, WordNet

I. INTRODUCTION

The standard argumentation of information theory [4] [10], the information content of a concept C can be quantified as negative the log likelihood, -log p(C). Quantifying information content in this way makes intuitive sense in this setting: as probability increases, informativeness decreases; so the more abstract a concept, the lower its information content. Moreover, if there is a unique top concept, its information content is 0. This quantitative characterization of information provides a new way to measure semantic similarity. The more information two concepts share, the more similar they are, and the information shared by two concepts is indicated by the information content of the concepts that subsume them in the taxonomy.

A natural, time-honored way to evaluate semantic similarity in a taxonomy is to measure the distance between the nodes corresponding to the items being compared - the shorter the path from one node to another, the more similar they are. Given multiple paths, one takes the length of the shortest one. A taxonomy is often represented as a hierarchical structure, which can be seen as a special case of network structure, evaluating semantic similarity between nodes in the network can make use of the structural information embedded in the network. There are several ways to determine the conceptual similarity of two words in a hierarchical semantic network. Here, one method is employed for semantic similarity measurement such as Combined Information Content and Path Length Based Semantic Similarity.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

II. RELATED WORK

Resnik [6] [9] proposed a simple information content approach to calculate the semantic similarity[7] as the information content of Lowest Common Subsumer (LCS) of two concepts as expressed by equation 1.

$$sim_R(C_1, C_2) = IC(LCS(C_1, C_2)) \quad (1)$$

Lin [8] defined measure of similarity between two concepts in a taxonomy is as :

$$sim_L(C_1, C_2) = \frac{2 \log p(C)}{\log p(C_1) + \log p(C_2)} \quad (2)$$

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.

III. COMBINED INFORMATION CONTENT AND PATH LENGTH BASED SEMANTIC SIMILARITY

Three of all measures are implemented to find the similarity score of word pairs. Among these, proposed new approach (CICPLBSS) is found to be most promising when compared to human judgment. Two methods (Lin [8] [11], Resnik [6]) are compared with new method to evaluate the performance of the proposed new approach. Here, the semantic similarity of a set of word pair is measured and examined the correlation between human judgement and machine calculations.

Word similarity [9] [11] [12] [14] can be determined by the best conceptual similarity value among all the concept (sense) pairs. It can be defined as follows:

$$sim(W_1, W_2) = \max_{C_1 \in sen(W_1), C_2 \in sen(W_2)} \left[sim(C_1, C_2) \right] \quad (3)$$

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:

$$p(C) = \frac{freq(C)}{M} \quad (4)$$

Where *freq*(C) is the *frequency* of the concept *C* estimated using noun frequencies from large text corpora [33] and *M* is the total number of observed instances of nouns in the corpus. The information content [3] [1] [2] of a concept C is defined as:

$$IC(C) = -\log p(C) \quad (5)$$

This part computes the semantic similarity using the Combined Information Content and Path Length Based Semantic Similarity (CICPLBSS). The Combined Information Content and Path Length Based Semantic Similarity(CICPLBSS) calculate the semantic similarity as addition of the negative ratio between length of the shortest path , depth and the ratio between the amount of information needed to state their commonality , the information needed to fully describe them as expressed by following equation .

$$sim_{CICPLBSS}(C_1, C_2) = \frac{2 \log p(C)}{\log p(C_1) + \log p(C_2)} - \log \frac{len(C_1, C_2)}{2D} \quad (6)$$

Where 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. Any similarity value of these measures may be used to generate concepts from the query keywords. Here Rubenstein-Goodenough [3] set of word pairs is considered for similarity measurement.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

The Combined Information Content and Path Length Based Semantic Similarity (CICPLBSS) is implemented. Evaluation is carried out on two datasets. The flow chart of system process using Combined Information Content and Path Length Based Semantic Similarity (CICPLBSS) is given in Figure 1. Search engine framework retrieves the web documents using synsets which generated by CICPLBSS measure.

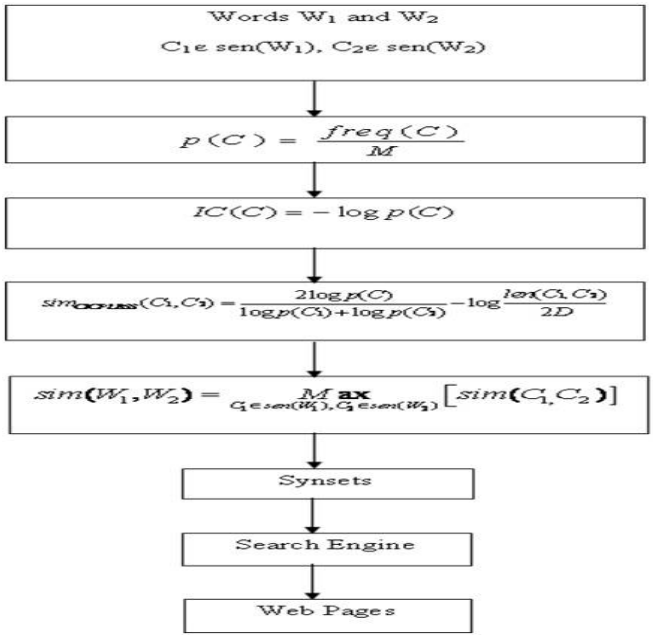

Figure 1: Flow Chart of System Process Using CICPLBSS.

The mathematical model of Combined Information Content and Path Length Based Semantic Similarity (CICPLBSS) is given in Figure 2.

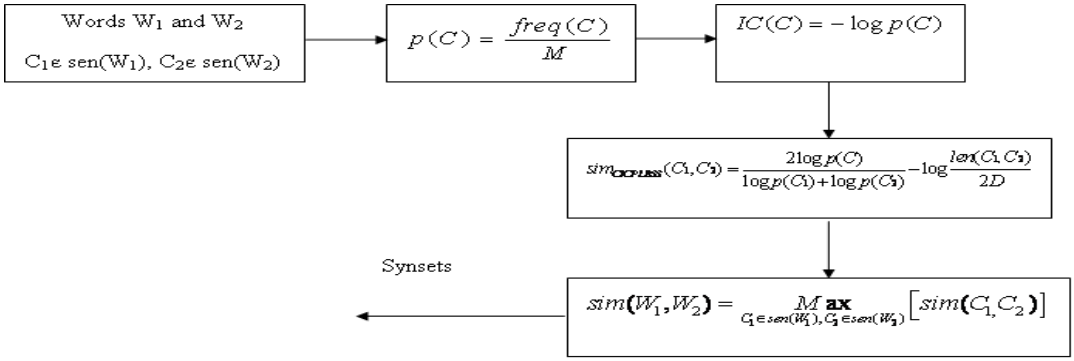

Figure 2: Mathematical Model of Combined Information Content and Path Length Based Semantic Similarity.

IV. Results and Discussion

The mean ratings from Rubenstein and Goodenough's and Miller and Charles's [3] [13] original experiments and the ratings of the Rubenstein–Goodenough and Miller–Charles word pairs produced by the Combined Information Content and Path Length Based Semantic Similarity (CICPLBSS) measure of similarity are given in following Tables.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Table 1: Word Pair Semantic Similarity Measurement for CICPLBSS (65 Pairs).

Word Pair		RG(X)	CICPLBSS (Y)
Cord	Smile	0.02	1.0185
Rooster	Voyage	0.04	0.5877
Noon	String	0.04	1.0986
Fruit	Furnace	0.05	1.8524
Autograph	Shore	0.06	1.0185
Automobile	Wizard	0.11	1.1875
Mound	Stove	0.14	1.8905
Grin	Implement	0.18	0.9444
Asylum	Fruit	0.19	1.6376
Asylum	Monk	0.39	1.1856
Graveyard	Madhouse	0.42	0.9513
Glass	Magician	0.44	1.5854
Boy	Rooster	0.44	1.3565
Cushion	Jewel	0.45	1.8821
Monk	Slave	0.57	2.2203
Asylum	Cemetery	0.79	1.0986
Coast	Forest	0.85	1.7722
Grin	Lad	0.88	0.9444
Shore	Woodland	0.9	1.9306
Monk	Oracle	0.91	1.7295
Boy	Sage	0.96	2.0431
Automobile	Cushion	0.97	1.8014
Mound	Shore	0.97	2.6335
Lad	Wizard	0.99	2.2457
Forest	Graveyard	1	1.3606
Food	Rooster	1.09	1.0456
Cemetery	Woodland	1.18	1.3606
Shore	Voyage	1.22	1.0185
Bird	Woodland	1.24	1.5341
Coast	Hill	1.26	2.6837
Furnace	Implement	1.37	2.0515
Crane	Rooster	1.41	1.504
Hill	Woodland	1.48	1.9404

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Car	Journey	1.55	0.9444
Cemetery	Mound	1.69	1.2792
Glass	Jewel	1.78	1.7555
Magician	Oracle	1.82	1.8542
Crane	Implement	2.37	1.974
Brother	Lad	2.41	2.2399
Sage	Wizard	2.46	2.015
Oracle	Sage	2.61	2.4054
Bird	Crane	2.63	2.1972
Bird	Cock	2.63	3.6513
Food	Fruit	2.69	1.607
Brother	Monk	2.74	2.8903
Asylum	Madhouse	3.04	3.873
Furnace	Stove	3.11	1.7071
Magician	Wizard	3.21	4.5835
Hill	Mound	3.29	4.5835
Cord	String	3.41	3.8268
Glass	Tumbler	3.45	3.7667
Grin	Smile	3.46	4.5835
Serf	Slave	3.46	2.1972
Journey	Voyage	3.58	3.5469
Autograph	Signature	3.59	2.8903
Coast	Shore	3.6	3.8568
Forest	Woodland	3.65	4.5835
Implement	Tool	3.66	3.8109
Cock	Rooster	3.68	4.5835
Boy	Lad	3.82	3.7064
Cushion	Pillow	3.84	3.8903
Cemetery	Graveyard	3.88	4.5835
Automobile	Car	3.92	4.5835
Midday	Noon	3.94	4.5835
Gem	Jewel	3.94	4.5835

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Table 2: Terms and Calculated Values Used for CICPLBSS (65 Pairs).

Terms	Calculated Values
Mean(X)	1.8756
Mean(Y)	2.3731
$Sum((X-X')^2$	116.3130
$Sum((Y-Y')^2$	98.4339
Numerator=Sum((X-X')*(Y-Y'))	90.4074
Denominator=$sqrt(Sum((X-X')^2)*Sum((Y-Y')^2))$	107.0006
Correlation=Numerator/Denominator	0.8449

Table 3: Word Pair Semantic Similarity Measurement for CICPLBSS (30 Pairs).

Word Pair		MC(X)	CICPLBSS (Y)
Automobile	Car	3.92	4.5835
Gem	Jewel	3.84	4.5835
Journey	Voyage	3.84	3.5469
Boy	Lad	3.76	3.7064
Coast	Shore	3.7	3.8568
Asylum	Madhouse	3.61	3.873
Magician	Wizard	3.5	4.5835
Midday	Noon	3.42	4.5835
Furnace	Stove	3.11	1.7071
Food	Fruit	3.08	1.607
Bird	Cock	3.05	3.6513
Bird	Crane	2.97	2.1972
Implement	Tool	2.95	3.8109
Brother	Monk	2.82	2.8903
Crane	Implement	1.68	1.974
Brother	Lad	1.66	2.2399
Car	Journey	1.16	0.9444
Monk	Oracle	1.1	1.7295
Cemetery	Woodland	0.95	1.3606
Food	Rooster	0.89	1.0456
Coast	Hill	0.87	2.6837
Forest	Graveyard	0.84	1.3606
Shore	Woodland	0.63	1.9306
Monk	Slave	0.55	2.2203
Coast	Forest	0.42	1.7722

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Lad	Wizard	0.42	2.2457
Chord	Smile	0.13	1.4591
Glass	Magician	0.11	1.5854
Rooster	Voyage	0.08	0.5877
Noon	String	0.08	1.0986

Table 4: Terms and Calculated Values Used for CICPLBSS (30 Pairs).

Terms	**Calculated Values**
Mean(X)	1.9713
Mean(Y)	2.5139
$Sum((X-X')^2$	59.4005
$Sum((Y-Y')^2$	44.9905
Numerator=Sum((X-X')*(Y-Y'))	42.0889
Denominator=sqrt(Sum$((X-X')^2$)*Sum$((Y-Y')^2$))	51.6958
Correlation=Numerator/Denominator	0.8141

The magician-wizard, hill-mound, grin-smile, forest-woodland,cock-rooster,cushion-pillow,cemetery-graveyard,automobile-car,midday-noon and gem-jewel pairs the value of one by Lin [8]. For instance, for the pair *rooster–voyage* (MC, RG dataset), the synsets rooster and voyage have different "unique beginners", and hence their lso— in fact their sole common subsumer — is the (fake) global root, which is the only concept whose probability is 1. The value of sim=0 by Lin [8], Resnik [9]. Analogously, although perhaps somewhat more surprisingly for a human reader, the same is true of the pair *grin–lad* (RG dataset).

Although, already for measures the details of their medium similarity regions differ, there appears to be an interesting commonality at the level of general structure: in the vicinity of sim = 2, the plots of human similarity ratings for both the Miller–Charles and the Rubenstein–Goodenough [3]word pairs display a very clear horizontal band that contains no points. For the Miller–Charles data (Figure 4), the band separates the pair *crane–implement* from *brother–monk* and for the Rubenstein-Goodenough [3]set (Figure 3), it separates *magician–oracle* from *crane– implement*.

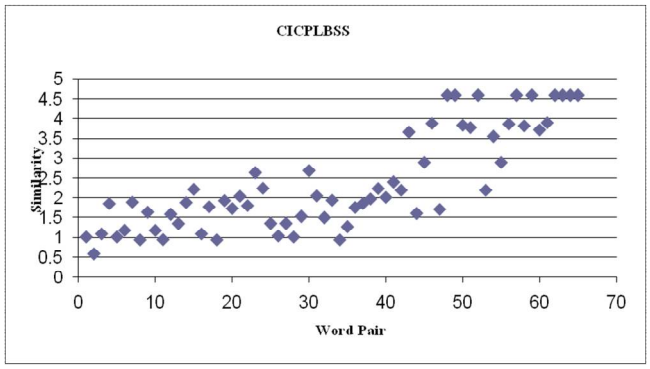

Figure 3: Human and Computer Ratings of the Rubenstein-Goodenough Set of Word Pairs, TheWord Pair Rated by CICPLBSS Measure.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

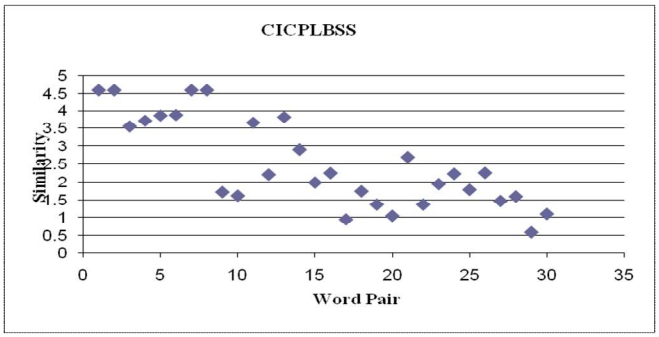

Figure 4: Human and Computer Ratings of the Miller-Charles Set of Word Pairs, TheWord Pair Rated by CICPLBSS Measure.

The correlation coefficient value of Combined Information Content and Path Length Based Semantic Similarity is 0.8449 for RG dataset and 0.8141 for MC dataset as shown in above graphs. The semantic similarity of noun words is calculated to obtain the related concepts described by the search query using ontology. Users query is replaced with concepts discovered from the similarity measures and fed to the Google search API.

V. CONCLUSION

Semantic similarity measurement using Combined Information Content and Path Length Based Semantic Similarity (CICPLBSS) method employed along with evaluation of this method is presented. Furthermore, the evaluation is carried out by implementing Combined Information Content and Path Length Based Semantic Similarity (CICPLBSS) approach. Three of all measures are implemented to find the similarity score of word pairs. Among these, new approach (CICPLBSS) is found to be most promising when compared to human judgment. Two methods (Lin [8], Resnik [6]) are compared with new method to evaluate the performance of new approach.

The correlation coefficient value of Combined Information Content and Path Length Based Semantic Similarity (CICPLBSS) is 0.8449 for RG dataset and 0.8141 for MC dataset. In graphs, all measure's behave quite similarly to each other in the low-similarity region.

REFERENCES

[1] Krishna Sapkota,Laxman Thapa,Shailesh Pandey Efficient Information Retrieval using measures of Semantic Similarity,2006.
[2] Anna Formica Concept similarity by evaluating information contents and feature vectors: A combined approach. Communications of the ACM, Vol.52, 2009.
[3] Leacock C.,Chodorow M.,"Combining local context and WordNet similarity for word sense identification",In Fellbaum 1998,pp.133-138.
[4] George Tsatsaronis and Vicky Panagiotopoulou, A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness,2009.
[5] Yung-Shen Lin, Jung-Yi Jiang and Shie-Jue Lee. A Similarity Measure for Text Classification and Clustering. IEEE Transactions On Knowledge And Data Engineering, 2013.
[6] P. Resnik. Using Information Content to Evaluate Semantic Similarity in Taxonomy. Proc. 14th International Joint Conf. Artificial Intelligence, 1995.
[7] G.A. Miller. WordNet: A Lexical Database for English. Comm. ACM, vol. 38, no. 11, PP. 39-41, 1995.
[8] D.Lin. An Information-Theoretic Definition of Similarity. Proc. of the Int. Conference on Machine Learning (ICML), Morgan Kaufmann, PP.296-304,1998.
[9] P. Resnik. Semantic Similarity in a Taxonomy: An Information- Based Measure and Its Application to Problems of Ambiguity in Natural Language. J. Artificial Intelligence Research, vol. 11, PP. 95-130, 1999.
[10] Rajendra Kumar Roul, Omanwar Rohit Devanand, Sanjay Kumar Sahay. Web Document Clustering and Ranking using Tf-Idf based Apriori approach. IJCA Proceedings on International Conference on Advances in Computer Engineering and Applications ICACEA (2):34-39, March 2014.
[11] Kishor Wagh, Satish Kolhe, Improving Web Link Mining using Semantic Similarity Measurement, International Journal of Applied Engineering Research (IJAER), Volume 9, Number 19 (2014),2014, ISSN 0973-4562 , PP. 5663-5677.
[12] Kishor Wagh, Satish Kolhe, Evaluate Semantic Similarity of Words Using Semantic Distance, VNSGU Journal of Science and Technology Vol. 3, Issue 2, March 2012, ISSN: 0975-5446, PP. 22-29.
[13] Kishor Wagh, Satish Kolhe, Information Retrieval Based on Semantic Similarity Using Information Content, IJCSI International Journal of Computer Science Issues, Vol. 8 ,Issue 3,May 2011, ISSN(Online): 1694-0814,PP. 364-370 .
[14] Kishor Wagh, Satish Kolhe, Semantic Similarity Based on Information Content, International Journal of Computer Science and Application, Issue 2010, ISSN: 0974-0767, PP. 82-85.