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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Improving Word Similarity Using PPMIC with Estimates of Word Polysemy

Nagajothi P[1], Hemalatha L[2], Kumari K[3], Jeevarathinam S[4]

Assistant Professor, Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, Namakkal – 637215, Tamilnadu, India[1]

Final Year Student, Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, Namakkal – 637215, Tamilnadu, India[2, 3, 4]

ABSTRACT— Measuring the semantic similarity between words is an important component in various tasks on the web such as relation extraction, community mining, document clustering, and automatic metadata extraction. But accurately measuring semantic similarity between two words or entities remains a challenging task. Point wise mutual information (PMI) is a widely used word similarity measure and it generates single sense for given word, but it lacks a clear explanation of how it works. PMI differs from distributional similarity, a novel metric is introduced PMImax, that augments PMI with information about a word's number of senses. PMImax estimates the maximum correlation between two words, i.e., the correlation between their closest senses. The existence system found out the PMImax and also produced an empirical method to estimate semantic similarity using page counts and text snippets retrieved from a web search engine for two words. PMImax can only find synonymous concepts and "siblings" concepts (e.g., "train" and "truck") but miss the "cousin" concepts. So the proposed system PPMIC (Positive Pointwise Mutual Information Cousins) concept can implement the cousin concept and also generates the top 50 most similar words for the noun. PPMIC has an amazing ability to improve the word similarity with word polysemy.

KEYWORDS — Semantic similarity, pointwise mutual information, Positive Pointwise Mutual Information Cousins, Distributional Similarity

I. INTRODUCTION

Word similarity is a measure of how semantically similar a pair of words is, with synonyms having the highest value. It is widely used for applications in natural language processing (NLP), information retrieval (IR), and artificial intelligence, including tasks like word sense disambiguation, malapropism detection, and paraphrase recognition, image and document retrieval and predicting hyperlink-following behavior. There are two prevailing approaches to computing word similarity, based on either using of a thesaurus (e.g., WordNet) or statistics from a large corpus. There are also hybrid approaches combining the two methods. Many well known word similarity measures have been based on Word Net and most of semantic applications rely on these taxonomy-based measures. Organizing all words in a well-defined taxonomy and linking them together with different relations is a labor intensive task that requires significant maintenance as new words and word senses are formed. Furthermore, existing Word Net-based similarity measures typically depend heavily on "IS-A" information, which is available for nouns but incomplete for verbs and completely lacking for adjectives and adverbs.

II. EXISTING SYSTEM

2.1 POINTWISE MUTUAL INFORMATION

PMI (Pointwise Mutual Information) can serve as a semantic similarity measure. PMI can generate a word with single sense. The PMI of a pair of outcomes x and y belonging to discrete random variables X and Y quantifies the discrepancy between the probability of their coincidence given their joint distribution and their individual

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

distributions, assuming independence PMI concept has the largest context overlap with itself, a concept has the largest chance to co-occur with itself. In other words, a concept has the strongest correlation with itself (auto correlation). Auto correlation is closely related to word burstiness, a phenomenon that words tend to appear in bursts. If the "one sense per discourse" hypothesis is applied. Thus, word burstiness is a reflection of the autocorrelation of concepts.

2.1.1 Process of Pointwise Mutual Information

There are two type approach:

a) PMI as a Semantic Similarity Measure

b)Augmenting PMI to Account for Polysemy

2.1.2 PMI as a semantic similarity measure

The semantic similarity between two concepts 1can be defined as how much commonality the system share. Since there are different ways to define commonality, semantic similarity tends to be a fuzzy concept. If commonality is defined as purely involving IS-A relations in a taxonomy such as WordNet, Two concepts are more likely to co-occur in a common, shared context and less likely in an unshared one and in a shared context. But if the base commonality on aptness to a domain, then soldier would be more similar to gun. People naturally do both types of reasoning, and to evaluate computational semantic similarity measures, the standard practice is to rely on subjective human judgments.

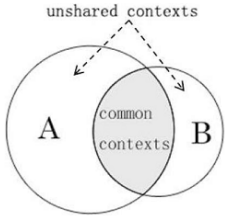


Figure 3.1 PMI as semantic similarity measure

Two concepts are more likely to co-occur in a common, shared context and less likely in an unshared one and in a shared con text.

$$PMI(c_1, c_2) \approx \log\left(\frac{f_d(c_1, c_2).N}{fc_1.fc_2}\right)$$

The number of co-occurrences also depends on the sizes of the two concepts. Therefore, the system need a normalized measure of co-occurrences to represent their similarity. PMI fits this role well to compute PMI for concepts in a sense-annotated text corpus, where fc1 and fc2 are the individual frequencies (counts) of the two concepts c1 and c2 in the corpus is the co-occurrence frequency of c1 and c2 measured by the context window of d words and N is the total number of words in the corpus. In this paper, log always stands for natural logarithm.

PMI use contexts differently which results in different behaviors. First, the PMI similarity between two concepts is determined by how much their contexts overlap, while distributional similarity depends on the extent that two concepts have similar context distributions. For example,garage has a very high PMI similarity with car because garage rarely occurs in contexts which are not subsumed by the contexts of car .However, distributional similarity typically does not consider garage to be very similar to car because the context distributions of garage and car vary considerably.While one might expect all words related by a PART-OF relation to have high PMI similarity, this is not the case and is not PMI-similar to leg , because there are many other contexts related to leg but not , and vice versa.

2.1.3 Augmenting PMI to account for polysemy

PMI has a well-known problem that it tends to overemphasize the association of low frequency words. The conjecture that the fact that more frequent content words tend to have more senses is an important cause for PMI's frequency bias. However, when PMI is applied to measure correlation between words, it has a problem because it assumes that words only have a single sense. Consider "make" and "earn" as an example. "Make" has many senses,

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

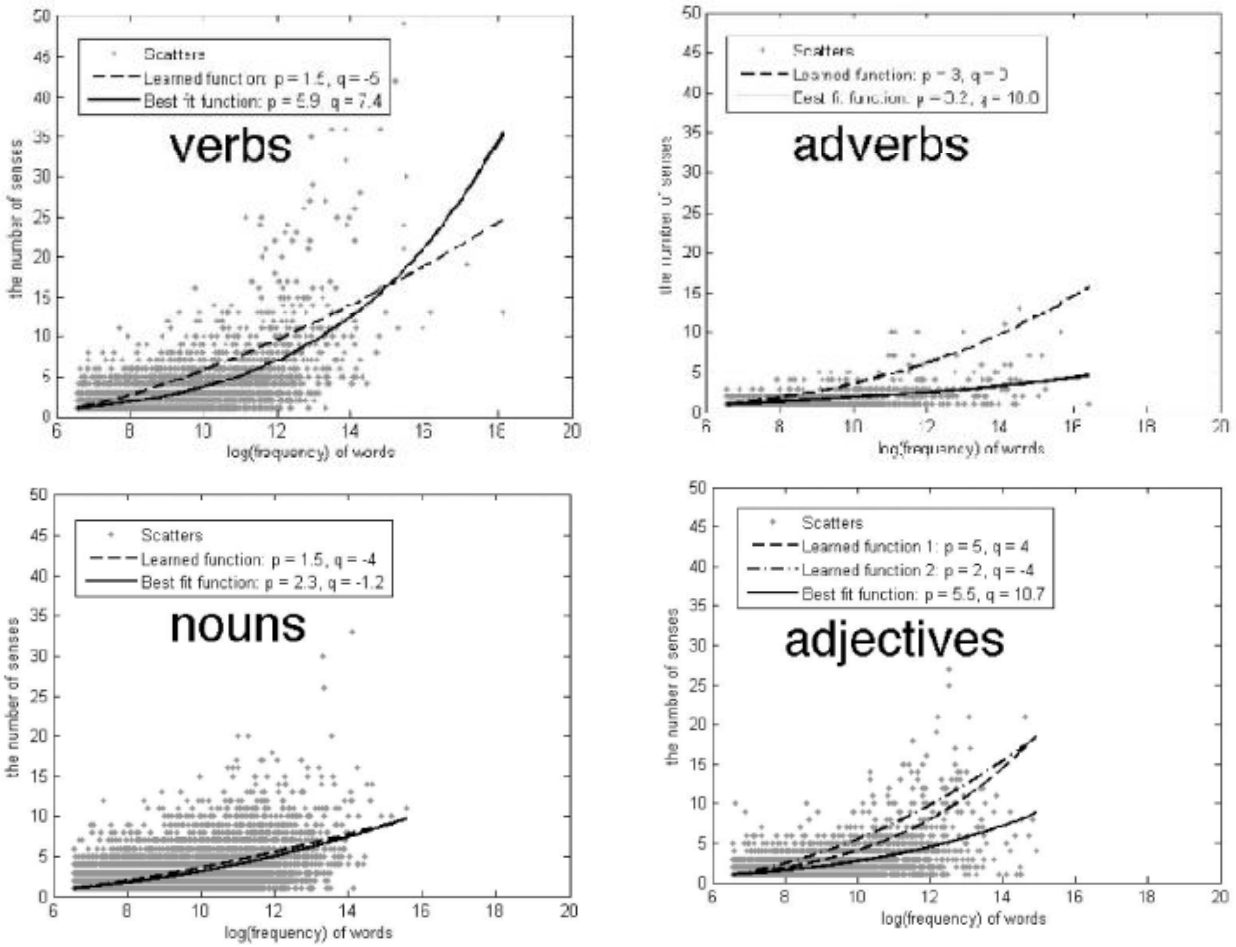
Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

only one of which is synonymous with "earn," and so it is inappropriate to divide by t he whole frequency of "make" in computing the PMI correlation similarity between "make" and "earn," since only a fraction of "make" occurrences have the same meaning of "earn." Although it can be difficult to determine which sense of a word is being used, this does not prevent from making a more accurate assumption than the "single sense" assumption. More specifically,

$$y_w = a(\log(f_w) + q)^p$$



The next estimate a word pair's PMI value between their closest senses using two assumptions. Therefore , t he c o-occurrence frequency between the two particular senses of w1 and w2 can b e estimated by subtracting the co-occurrence frequency contributed by other combinations of senses, denoted by x, from the total co-occurrence frequency between the two words. Given a word pair, it is hard to know the proportions at which the closest senses are engaged in their own words. Since it can be either a major or minor sense, simply assume the average proportion 1/fw. Consequently, the frequency of a word used as the sense most correlated with a sense in the other word is estimated as fw/yw. More specifically,

$$f_d(w_1, w_2) - x,$$

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

$$\log\left(\frac{x.N}{f_{w1}.f_{w2}-\frac{f_{w1}}{y_{w1}}.\frac{f_{w2}}{y_{w2}}}\right)=k$$

Above Equation amounts to asking the question with a correlation degree of k. Finally, the modified PMI, called PMImax, between the two words w1 and w2 is given in

$$PMI_{max}(w_1,w_2)=\log\left(\frac{(f_d(w_1,w_2)-x).N}{\frac{f_{w1}}{y_{w1}}.\frac{f_{w2}}{y_{w2}}}\right)$$

$$=\log\left(\frac{\left(f_d(w_1,w_2)-\frac{e^k}{N}\left(f_{w1}.f_{w2}-\frac{f_{w1}}{y_{w1}}.\frac{f_{w2}}{y_{w2}}\right)\right).N}{\frac{f_{w1}}{y_{w1}}.\frac{f_{w2}}{y_{w2}}}\right)$$

PMImax estimates the maximum correlation between two words, i.e., the correlation between their closest senses. In circumstances where the system cannot know the particular senses used, it is reasonable to take the maximum similarity among all possible sense pairs as a measure of word similarity.

2.2 Drawback

- PMImax can only generate the sibling concept.
- It does not support the cousin concept.
- It lacks the clear explanation about the word

III. PROPOSED SYSTEM

3.1 PPMIC

PPMIC (Positive Pointwise Mutual information Cousins) can generate the top 50 most similar words for the noun and also can generate the word with synonyms, siblings and cousins concept. PPMIC can be used in various applications that require word similarity measures. However, the system are more interested in combining PPMIC with distributional similarity in the area of semantic acquisition from text because this direction is not yet explored.

Example train, automobile, boat, wagon, carriage, engine, vehicle, motor, truck, coach, cab, wheel, ship, machine, cart, locomotive, chariot, canoe, vessel, craft, horse, bus, auto, driver, sleigh, gun, launch, taxi, buggy, barge, yacht ,ambulance, passenger, freight, box, round, plane, trolley, station, team, street, track, window, rider, chair, mule, elevator, bicycle, door, shaft The example shows that, as a state-of-the-art distributional similarity, PPMIC has an amazing ability to find the concepts that are functionality or utility similar to "car." These concepts, such as "train," "boat," "carriage," "vehicle," are typically neighboring concepts of "car" in a taxonomy structure such as WordNet.

In contrast, as illustrated in the "car" can only find synonymous concepts and "siblings" concepts (e.g., "train" and "truck") but miss the "cousin" concepts (e.g., "boat" and "carriage"). The distributional similarity can find neighboring concepts of the target word in a taxonomy. A subsequent question on the course is how the system could classify these concepts into different clusters corresponding to the "sibling," "parent," and "cousin" sets in a taxonomy. This is an largely unsolved problem in ontology learning from text that the combination of distributional similarity can help address. For example, of the 50 most distributional similar words of "car," which are most likely to be classified together with "boat". A simple approach is to take the intersection of two top 50 candidate lists generated by PPMIC for "car" and "boat," respectively. The results for "boat" and several other examples obtained this way. The words that can be classified with "boat" include its "sibling" concepts (conveyances on water) and "parent" concepts (vessel, craft) but no "cousin" concepts (conveyances on land). Similarly, the intersection list for "carriage" includes archaic vehicles and rejects modern ones. Although in the example of "chariot," "car" appears in the list, it is due to a different sense of "car." And identifies the different patterns that describe the multiple semantic relations.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Regarding to identifying the "parent" set, common words can be good cues. For example, "boat" and "ship" have the common word "vessel" whereas "carriage," "chariot," and "truck" share the words "vehicle" and "wheel." This suggests that "boat" and "ship" may have parent "vessel" while "carriage," "chariot," and "truck" may have parent "vehicle" or "wheel." This also implies that in the time of Gutenberg corpus, about 80 years ago, "vehicle" cannot be used to describe "boat" or "ship."

PPMIC implementation differs from Bullinaria and Levy's in using lemmatized and POS-tagged words (noun, verb, adjectives, and adverbs) rather than unprocessed words as vector dimension. This variation is simply due to convenience of reusing the systems already have the system tested the implementation of PPMIC on TOEFL synonym test.

3.1.1 Mean Average Precision (MAP) Evaluation

Mean Average Precision is a common measure used to evaluate systems in information retrieval tasks. Automatic thesaurus generation can be evaluated as an IR task if the system make an analogy between an IR query and the need to identify synonyms for a target word (in this analogy, correct synonyms are the relevant documents). The system compares PMI, PMImax, and PPMIC using MAP

	Noun	Verb	Adj.	Adv.
PMI	0.120	0.160	0.163	0.103
PMImax	0.168	0.256	0.261	0.179
PPMIC	0.433	0.442	0.436	0.487

Table Precision of PMI, PMImax and PPMIC

3.1.2 Comparison to Distributional Similarity

To demonstrate the efficacy of PMImax in automatic thesaurus generation, compare it with a state-of-the-art distributional similarity measure proposed by Bullinaria and Levy. Their method achieved the best performance after a series of work on distributional similarity from their group. The method is named Positive PMI components and Cosine distances (PPMIC) because it uses positive pointwise mutual information to weight the components in the context vectors and standard cosine to measure similarity between vectors. Bullinaria and Levy demonstrated that PPMIC was remarkably effective on arrange of semantic and syntactic tasks, achieving, for example, an accuracy of 85 percent on TOEFL synonym test using the BNC corpus.

3.2 Advantages

- The new system integrates different similarity measures using a PPMIC approach
- Identifies the different patterns that describe the multiple semantic relations
- Supports cousin concept and generates multiple sense for a given word
- Improves word similarity with estimates of word polysemy in the system

IV. SYSTEM MODELS

4.1 SEMANTIC SIMILARITY

4.1.1 Page count based co-occurrence measures

Word1 and word2 are keyed in and the words are combined and displayed as word pair. The 'webdocument' folder located in root folder of the application contains HTML pages. The pages are searched with these words. Three list boxes are provided. The first listbox is populated with the page names containing the 'word1'. The second listbox is populated with the page names containing the 'word2'. The third listbox is populated with the page names contain both the words. The counts of word1 pages, word2 pages and both words are also displayed in label controls. The values are stored in 'GlobalClass' class and used in successive modules.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

4.1.2 PMI and PMImax calculation

PMI value is calculated as follows,

$$PMI(c_1, c_2) \approx \log\left(\frac{f_d(c_1, c_2).N}{fc_1.fc_2}\right)$$

PMI is explained as the logarithmic ratio of the actual joint probability of two events to the expected joint probability if the two events were independent. Here, the module interprets it from a slightly different perspective and this interpretation is used in deriving the novel PMI metric.

The term fc1. fc2 can be interpreted as the number of all co-occurrence possibilities or combinations between c1 and c2. The term f_d(c1, c2) gives the number of co-occurrences actually fulfilled. Thus, the ratio f_d(c1, c2) / fc1. fc2 measures the extent to which two concepts tend to co-occur. The modified PMI, called PMImax, between the two words w1 and w2 is given in,

$$PMI_{max}(w_1, w_2) = \log\left(\frac{(f_d(w_1,w_2)-x).N}{\frac{f_{w1}}{y_{w1}}\frac{f_{w2}}{y_{w2}}}\right)$$

$$= \log\left(\frac{\left(f_d(w_1, w_2) - \frac{e^k}{N}\left(f_{w1}.f_{w2} - \frac{f_{w1}}{y_{w1}}.\frac{f_{w2}}{y_{w2}}\right)\right).N}{\frac{f_{w1}}{y_{w1}}.\frac{f_{w2}}{y_{w2}}}\right)$$

In circumstances where the system cannot know the particular senses used, it is reasonable to take the maximum similarity among all possible sense pairs as a measure of word similarity.

4.1.3 Co-occurrence measures

4.1.3.1 Web jaccard

The H(P) page count with word1, H(Q) page count with word2, H(P^Q) page count with word pair are displayed in label controls and Web Jaccard Value is calculated and displayed in a label control.

4.1.3.2 Web overlap

The H(P) page count with word1, H(Q) page count with word2, H(P^Q) page count with word pair, minimum of H(P) and H(Q) are displayed in label controls and Web Overlap Value is calculated and displayed in a label control.

4.1.3.3 Web dice

The H(P) page count with word1, H(Q) page count with word2, H(P^Q) page count with word pair, 2 * H(P^Q) are displayed in label controls and Web Dice Value is calculated and displayed in a label control.

4.1.3.4 Web PMI (Point wise Mutual Information)

The H(P) page count with word1, H(Q) page count with word2, H(P^Q) page count with word pair, H(P)/N, H(Q)/N, H(P^Q)/N are displayed in label controls and Web PMI Value is calculated and displayed in a label control. In the sample values, 'N' is taken as 10. In real time the 'N' will be 10 to the power of 10 or more.

4.1.4 Lexical pattern extraction and clustering

Search pattern input with multiple words

The search pattern is entered in which the first word and last word are taken. In the web pages, the phrase is checked such that the pattern is first word, any number of words and the last word. During the pattern extraction, the skip count number of words can be discarded in the phrase found in the web pages.

4.1.5 PPMIC

PPMIC (Positive Pointwise Mutual information Cousins) can generate the top 50 most similar words for the noun and also can generate the word with synonyms, siblings and cousins concept. PPMIC can be used in various

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

applications that require word similarity measures. However, the system are more interested in combining PPMIC with distributional similarity in the area of semantic acquisition from text because this direction is not yet explored.

V. PROPOSED RESULT AND DISCUSSION

The word similarity in the data mining can be improved by augumenting PPMIC and it can generate the word with multiple senses that is synonyms, siblings, cousins and almost top 50 related words can be generated. The system anticipates that PMI and PMImax and PPMIC will play an important role in lexical semantic applications in the future. Identifies the different patterns that describe the multiple semantic relations. Supports cousin concept and generates multiple sense for a given word. Improves word similarity with estimates of word polysemy in the system.

VI. CONCLUSION

To improve the word similarity in the data mining, the PPMIC can generate the word with multiple senses that is synonyms, siblings, cousins and almost top 50 related words can be generated. The PPMIC outperforms PMImax in automatic thesaurus generation and on benchmark data sets for human similarity ratings and TOEFL synonym questions. PMImax also gives, among corpus-based approaches, the highest correlation coefficient for the Miller-Charles data set based on 27 pairs. PMImax need not rely on web search engine data or an information retrieval index to be effective in a range of semantic tasks.

Compared with distributional similarity, PMImax is a lightweight measure, though it requires a larger corpus to be effective. With the vast amount of data available today, data sparseness, becomes a much less severe issue than 20 years ago when Church and Hanks popularized the use of PMI in computational linguistics. The system anticipates that PMI and PMImax and PPMIC will play an important role in lexical semantic applications in the future.

REFERENCES

[1] Andrea Rodrguez and Egenhofer M.J, (2003) "Determining Semantic Similarity among Entity Classes from Different Ontologies", VOL. 15, No.2, pp.442-456.

[2] Andrew Skabar and Khaled Abdalgader, (2013) "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm", VOL. 25, NO. 1, pp.62-75.

[3] Alexandros Potamianos, Elias Iosif and Shrikanth Narayanan, (2013) "Distributional Semantic Models for Affective Text Analysis", VOL. 21, NO. 11, pp.2379-2392.

[4] Bandar ZA, David McLean and Yuhua L.I, (2003) "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", VOL.15, NO.4, pp.871-882.

[5] Bin Fang and Taiping Zhang, (2012) "Document Clustering in Correlation Similarity Measure Space", VOL. 24, NO. 6, pp.1002-1013.

[6] Bin Jiang and Jian Pei, (2013) "Clustering Uncertain Data Based on Probability Distribution Similarity", VOL. 25, NO. 4, pp.751-763.

[7] Capitaine H.L, (2012) "A Relevance-Based Learning Model of Fuzzy Similarity Measures", VOL. 20, NO. 1, pp.57-68.

[8] Chandresh , Debasish Dutta and Lalit Patil, (2012) "An Information-Based Approach to Compute Similarity Between Engineering Changes" VOL. 9, NO. 2, pp.330-341.

[9] Chi-Huang Chen ,Wen-Yung Chang and Yung-Ching Weng, (2013) "Semantic Similarity Measures in the Biomedical Domain by Leveraging a Web Search Engine", VOL. 17, NO. 4, pp.853-861.

[10] Dan Lin, Prathima Rao and Rodolfo Ferrini, (2013) "A Similarity Measure for Comparing XACML Policies", VOL. 25, NO. 9, pp.1946-1959.