ISSN:2348-2079

Volume-5 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Improving word similarity by using ppmic with estimates of word polysemy

[1]Mr C. Mani MCA., M.Phil., M.E., Associate Professor,
[2]Mrs S. Sakthi, Final MCA,
Department of MCA, Nandha Engineering College (Autonomous), Erode-52
Email Id: cmanimca@gmail.com, ssakthi28@gmail.com

Abstract - Semantic similarity is a metric defined over a set of documents or terms, where the idea of distance between them is based on the likeness of their meaning or semantic content which can be estimated regarding their syntactical representation. Its applications are Biomedical informatics, GeoInformatics, Computational linguistics, Natural language processing. There are two approaches to compute word similarity, on either using of thesaurus (e.g., Word Net) or statistic from large corpus.PMI means point wise mutual information. It is used to measure semantic similarity. In the existing system PMI and PMImax is used to measure semantic similarity. PMImax is used to find out the maximum correlation between the words. PMImax only find siblings concept but it fails to find out cousin concept. So proposed system uses PPMIC to find and improve word similarity.

1.INTRODUCTION

Word similarity is used to measure semantic similarity between two words. It is applied in natural language processing task, retrieval of information, and AI, disambiguation of word sense, detection of malapropism, recognition of paraphrase, retrieval of image, and retrieval of document and finding out its behaviour.[1] Semantic similarity can be estimated by defining a topological similarity, by using ontologies to define the distance between terms/concepts. Semantic similarity measure plays a key role in web related to clustering of document, extraction of metadata.

2.EXISTING SYSTEM

2.1 PMI

PMI stands for point wise mutual information. It is used to measure semantic similarity. It can generate word with single sense.point wise mutual information is a correlation measure[1][2], for two events, *x* and *y*; mutual information measures the point wise mutual information over all possible events: that is, MI is the expectation (average) of PMI over all possible outcomes. The point wise aspect of PMI indicates that we are considering specific events. Consider the following formula.

$$\operatorname{pmi}(x;y) \equiv \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}.$$

2.1.1. PMI Process

The approaches are:

a) PMI used as semantic similarity measure.
b) PMI augmenting is to accounting polysemy

2.1.2 PMI used as semantic similarity measure

Semantic similarity between the two words is defined as the commonality the system they share. there are a different ways to define the commonality .The commonality can be defined as IS-A relation such word net, these concepts is likely to occur in common and likely in unshared one. [3][4]If the base commonality apted to domain, for eg, soldier is similar to a gun. Naturally, the people do both types of reasoning. It is done to evaluate semantic similarity measures, the practice can be used to rely on human judgements.

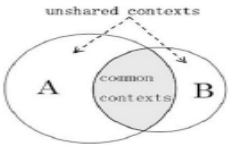


Figure: PMI used as semantic similarity measure

The concepts were likely co-occur in common text and sharing context[4], unshared context and in shared context.

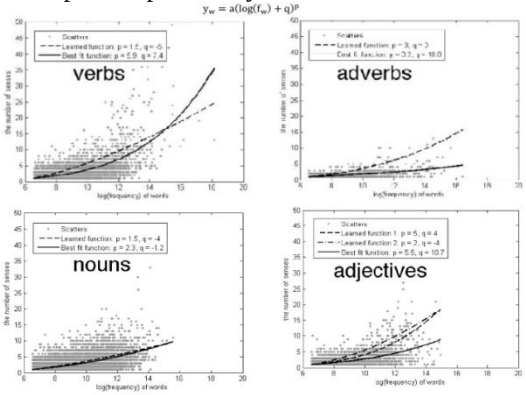
$$\mathrm{PMI}(c_1, c_2) \approx \log\left(\frac{f_d(c_1, c_2).N}{fc_1.fc_2}\right)$$

The co-occurrences depends on the sizes of the these concepts. Hence, the system requires a normalize measuring of co-occurrences. It represents their similarity. It fits the role fairly for

computing PMI concepts in a sense of annotated text corpus, where fc1 and fc2 are the individual counts of the concepts c1 and c2 in the corpus. It is the co-occurrence frequency of c1, c2 measured by context window represented by d words .N is the total number of words in the corpus. In this log represent natural logarithm. PMI uses the contexts in various ways .It results in different behaviour.The PMI similarity between the concepts is found by how much its contexts overlap, whereas distributional similarity depend ending upon the extent that two concepts having the similar context distributions. For example, crane has a very high PMI similarity with bird because crane rarely occurs in contexts which are not subsumed by the contexts of bird[4][5] .However, distributional similarity typically does not consider crane to be very similar to bird because the context distributions of crane and bird 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 PMI augmenting is to accounting polysemy

PMI has a problem which intends to overemphasize association of less frequency words. Its fact is to more frequently content words intend to have more than one senses. It is most important reason that it leads to PMI's frequency bias. Applying PMI to measure the correlation between words. There is a problem occurs as it also take words which has single sense. Consider "make" and "earn" as an example. "Resign" has many senses only one of which is synonymous with "to sign up again" and is not appropriate to divide by whole frequency of "Resign" in computing the PMI correlation similarity between "Resign" and "to sign up again," since only a fraction of "Resign" occurrences have the same meaning of "to sign up again."It is too tedious to determine the sense of a word is being used, which do not prevent to make more accurate assumption than "single sense" assumption. Specifically,



The next thing is to estimate a word pair's PMI value that occurs between their closest senses using the two assumption. Hence, the co occurrence frequencies that occur between particular senses of w1 and w2.It can be estimated by subtracting co occurrence frequencies[5], given by other combinations of senses, denoted by x, from total co-occurrence frequencies among the two words. We take word pair; it is difficult to know the proportions in which the closest senses are used in their own words. It may be either a major or minor sense assuming the average proportion 1/fw. Often, the frequency of a word can be used as the sense in most correlated sense. In the other word, it is estimated as
fw/yw. Expressing specifically,

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

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

The Equation relates to correlation degree of k. The modified PMI, is called PMImax, between the two words w1 and w2 is given as,

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

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

PMImax is used to estimate the maximum correlation between their closest senses. In circumstances which the system cannot know the particular senses where it is used[5][6], taking maximum similarity among sense pairs which is a measure of word similarity.

2.1.4 Drawback

- Efficient estimating of semantic similarity between words is very critical for various natural language processing tasks.
- If page counts alone is used for measuring co-occurrence of two words. It presents several drawbacks. Analysis of page count ignores the position of a word in a page.
- Although two words appear in a page, might not be actually related. The page count of a polysemous word (a word with multiple senses) might also contain a combination of all its senses.
- Similarlily Search patterns (lexical patterns) were not clustered.

3. PROPOSED SYSTEM

3.1 Positive point wise mutual information

It can generate top 50 most similar words for noun it can also can generate the word with synonyms, siblings and cousins concept. PPMIC can be used in applications which require word similarity measures.example:plant,shrub,seed,garden,flower, soil,water,animal,genus,flowering,tree,species,vegetable,apple,orange,lemon,tomato,pineapple,corn, bananaetc. The example shows that, as a state-of-the-art distributional similarity, PPMIC has amazing ability to find utility similar to "plant." These concepts, such as "seed","garden"are typically neighbouring concepts of "plant" in a taxonomy structure such as WordNet.

In contrast, as illustrated in the "plant" can only find synonymous concepts and "siblings" concepts(e.g., "seed" and "shrub") but miss the "cousin" concepts (e.g., "species" and "flowering"). The distributional similarity can find neighbouring concepts of the word. Question arises on the course is how the system could classify these concepts into different groups. These groups corresponding to the "sibling," "parent," and "cousin" sets in a taxonomy. This is a largely unsolved problem.

For example, of the 50 most distributional similar words of "plant," which are most likely to be classified together with "seed". A simple approach is to take the intersection of two top 50 candidate lists generated by PPMIC for "plant" and "seed," respectively. The results for "seed" and several other examples obtained this way. The words that can be classified with "seed" include its "sibling" concepts (conveyances on species) and "parent" concepts but no "cousin" concepts.

Regarding to identifying the "parent" set, common words can be good cues. For example, "plant" and "flower" have the common word "species" . Thus in this way it identifies the different patterns that describe the multiple semantic relations.

Mean Average Precision is a measure used to evaluate systems in retrieval of information tasks. PMI, PMImax, and PPMIC can be compared using mean average precision evaluation.

	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

3.1.1Comparison to Distributional Similarity

In automatic thesaurus generation, the efficiency of PMImax is demonstrated by comparing it with the state-of the- art distributional similarity measure which is proposed by Bullinaria and Levy. The proposed method's performance achieved best due to series of work on distributional similarity. As it used positive pointwise mutual information, inorder to weigh components in the context vectors and the standard cosine for measuring the similarity between the vectors.This method is called Positive PMI components and Cosine distances (PPMIC). They demonstrated that it was remarkably effective on arrangement of semantic task achievement. for eg, there was an accuracy of 85 percent on TOEFL synonym test using BNC corpus. The TOEFL task was first used by Tom Landauer & Susan Dumais, a solution to Plato's problem. The latent semantic analysis theory of acquisition, induction and representation of knowledge. It consists of 80 multiple-choice synonym questions.

3.2Advantages

- using a machine learning approach,the new system integrates different web-based similarity measures
- Identifies the different patterns which describe the same semantic relation.
- Lexical pattern extraction algorithm considers the word subsequences in the text snippets.

4. SYSTEM MODELS

4.1 SEMANTIC SIMILARITY

4.1.1 Page count based co-occurrence measures

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

4.1.2 calculation of PMI and PMImax

In this module, PMI value is calculated as follows.

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

PMI can be explained as the logarithmic ratio of the actual joint probability of two events to the expected joint probability, unless the two events are independent. if the module interprets it from slightly different perspective, this interpretation is used in derivation of novel PMI metric.

The term *fc1. fc2* can be described 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 fulfilled actually. The ratio f_d (c1, c2) / *fc1. fc2* measures extent to which two concepts intend to co-occur.

By analogy to the correlation in the statistics which measures degree that two random variables tend to co-increase/decrease, PMI measures the likelihood that two concepts intend to co-occur versus occurring alone. In this sense, we say that the PMI computes correlation between the concepts c1 and c2. It is also referred to the semantic similarity that PMI represents as correlation similarity.

The modified PMI, called PMImax, between the two words w1 and w2 is given in

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

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

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

4.1.3 Co-occurrence measures

4.1.3.1 Web jaccard

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

4.1.3.2 Web overlap

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

4.1.3.3 Web dice

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

4.1.3.4 Web PMI (Point wise Mutual Information)

The H(P) page count populated with word1, H(Q) page count populated with word2, H(P^Q) page count populated 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 the label control. In 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.In this the search pattern is found out from the web pages and the pages names are added in a list.The patterns can be clustered using the lexical pattern clustering algorithm. The patterns are clustered and then the count and co-occurrence of the word can be considered. Based on this the word can be extracted. The cluster can be grouped based on the threshold value entered in textbox control, the words are clustered and the n the results are produced in the listbox control.

4.1.5 Positive Pointwise Mutual information Cousins

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

5. PROPOSED RESULT AND DISCUSSION

The new system will be implemented that the application will further improve so that the project can work with the full efficiency. To ensure reliability in ever growing client count. The content can be searched from more than one search engine. Multitasking can be performed .If developed as web service, it can be accessed from anywhere.The project can be further developed by working in different operating system independently.

6. CONCLUSION

In data mining, the word similarity can be improved, so that 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. PMImax need not rely on web search engine data or an information retrieval index to be effective in a range of semantic tasks. When Comparison is made with distributional similarity, PMImax is a lightweight measure, though it requires a larger corpus to be effective. The system anticipates that PMI and PMImax and PPMIC play an important role in lexical semantic applications.

REFERENCES

[1] Lushan Han," Improving Word Similarity by Augmenting PMI with Estimates of Word Polysemy", ieee transactions on knowledge and data engineering, vol. 25, no. 6, june 2013,pp:1307-1322

[2] Danushka Bollegala,"Minimally Supervised Novel Relation Extraction
Using a Latent Relational Mapping" ieee transactions on knowledge and data engineering, vol. 25, no. 2, february 2013,pp:419-432.

[3] D. Hindle, "Noun Classification from Predicate-Argument Structures,"Proc. Ann. Meeting ACL, pp. 268-275, 1990.

[4] G. Grefenstette, Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, 1994.

[5] D. Lin, "Automatic Retrieval and Clustering of Similar Words," Proc. 17th Int'l Conf. Computational Linguistics, pp. 768-774, 1998.

[6] J.R. Curran and M. Moens, "Improvements in Automatic Thesaurus Extraction," Proc. Workshop Unsupervised Lexical Acquisition,
pp. 59-66, 2002.