A Semantic Search Engine using Semantic Similarity Measure between Words

M.Karthiga, M.Priyadharsini

Abstract— Measuring semantic similarity between words is very useful in information retrieval. Semantic similarity measure is so useful in many applications, and in the proposed work it is used to create a model Semantic Search Engine. The Semantic Search Engine uses in one hand a Technical Database for computer technology and a Semantic Similarity database to retrieve the resultant Web page for the query word. When a query word is given in the user interface the search engine first searches for the word in the technical database if the word is present the respective Webpage is displayed. If the word is not present in the technical database then the query word is searched in the semantic similarity database. If there are any similar words for the query word those words are displayed as recommendations to the user. The user has to select one of the similar words from the recommendation and accordingly the result page is retrieved. The semantic similarity measure between the words is evaluated using both Pearson correlation coefficient and Spearman correlation coefficient. The time taken to retrieve the relevant Webpage in semantic search engine is compared with normal search engine. The Precision and Recall is calculated for semantic search engine and the results are compared with normal search engine.

Index Terms— - Information retrieval, Precision, Recall, Search engine, user generated content

1 Introduction

Web has a lot of hidden information that are interconnected by various semantic relations. Semantic similarity measure helps to identify the semantic relations within the web data which in turn help to extract the useful information from the data. Providing a semantic wise search engine is a challenging task in information retrieval.

WordNet could not provide an efficient method to find the semantic similarity because the semantic similarity between the words usually changes with time and domain.

More efficient way to measure the semantic similarity is automatic method through search engines [1]. Page counts, dictionary based metric and snippets are some types of useful information provided by a search engine. Page count for a query is the number of web pages returned as a result to the user by the search engine. Page counts for two words provide the global co-occurrences of the two words on the result web pages. If two words have more page count then they are more similar.

But page counts alone as a similarity measure has lot of drawbacks. First, page count for a query word ignores the position of that word in a page. Second, a page count for polysemous word (a word with multiple senses) contains a combination of all its meanings. Moreover, due to the presence of scale and noise on the web, unrelated words might co-occur on same pages. So, page counts measure alone could not be used for measuring the semantic similarity.

Snippet is the window of text provided by the search engine which reveals the information in brief about the query terms. Snippets avoid the need to download the exact Webpage. Snippets saves time to the users. Snippets of the two words represent the local context occurrence of the query word. Consider a snippet from Google for the query *Glass* AND *Magician*.

*"Wednesday started for me with a lecture by Tim Star, a Swedish **magician** who has and **glass** steal from the table and finished by producing both his shoes!"*

Since only top ranked snippets are processed well by the search engine considering snippets alone as a metric to measure semantic similarity poses many drawbacks. Also it is not guaranteed that the top ranked snippets provide the complete needed information about the query to the users.

In this paper, a dynamic search engine is proposed which considers the technical dictionary and similarity measure using page counts and text snippets for effectively retrieving the information for the user query. This semantic search engine overcomes the problems of normal search engine.

The rest of the paper is organized as follows: Section 2 introduces the literature survey on semantic similarity methods. In section 3, detailed representation about the proposed work is provided. Section 4 reveals the system implementation. In section 5, conclusion and some future perspectives is presented.

2 Preliminary Works

The semantic similarity between the words is measured using a metric called distance when the knowledge base is like a graph [16]. If the two words are represented as two nodes then their conceptual distance is the minimum number of edges separating the nodes. The drawback with this approach is that it considers that all links in the knowledge base has a uniform distance.

Besides evaluating the semantic similarity by considering distance, Resnik et al [17] measured the similarity using the information content. The similarity between two concepts is based on how the two concepts share the common information. If the two concepts have more common information then they are considered as a highly specific content. In case of multiple inheritances, similarity among words is taken into account. Word sense disambiguation problem is not considered in this approach. So this approach provides the similarity measure based on irrelevant word senses.

Li et al [13] proposed a non linear model which uses the combination of structural semantic information from a lexical taxonomy and information content from a corpus. The experiment reported a high Pearson correlation coefficient of 0.8914 on the Miller and Charles benchmark data set. But the proposed work did not measure the similarities among named entities.

Lin et al [11] defined the similarity between two concepts as the information that is in common to both concepts and the information contained in each individual concept. A universal definition of similarity in terms of information theory was presented. A definition for similarity is provided by Lin that achieves two goals: Universality and Theoretical Justification. Universality means definition of similarity is applied to many different domains. Theoretical Justification means similarity measure is not defined directly by a formula. Rather, derived from a set of assumptions about similarity.

Cilibrasi and Vitanyi [3] proposed a similarity metric using only page counts retrieved from a web search engine which is called as Normalized Google Distance (NGD). NGD is based on normalized information distance. The proposed methodology does not take into account the context in which the words co-occur.

Sahami [19] used snippets to measure the semantic similarity between two queries. For each individual query, snippets are collected and each snippets are represented as a TF-IDF weighted term vector. Then normalize each vector and centroid of the set of vectors is calculated. The similarity measure is then defined as the inner product between the corresponding centroid vectors. But the similarity measure in the proposed work was not compared with the taxonomy-based similarity measure.

Chen et al [2] proposed a double-checking model using text snippets returned by a web search engine to compute semantic similarity between words. For two words P and Q, snippets are collected for each word from a web search engine. Then, the occurrences of word P in the snippets for word Q and the occurrences of word Q in the snippets for word P are counted. These values are combined nonlinearly to compute the similarity between P and Q. But this method depends heavily on the search engine's ranking algorithm. Though two words P and Q might be very similar, one cannot assume that the word Q could be found in the snippets for P or P in snippet for Q, because a search engine considers many other factors besides semantic similarity, such as publication date (novelty) and link structure (authority) when ranking the result set for a query.

Imen Akermi [8] introduced a new similarity measure between words using an online English dictionary provided by the Semantic Atlas project of the French National Centre for Scientific Research and page counts returned by the social website Digg.com whose content is generated by the users. In the proposed work, polysemy and semantic disambiguation problem has been dealt.

Bollegala et al [1] proposed a web search engine based approach to measure the semantic similarity which is used in query expansion, word sense disambiguation. The proposed idea of measuring the semantic similarity is using page counts and text snippets. Support vector machine is used for classification.

3 Proposed Work

For measuring the semantic similarity between the words page counts and text snippets based metric collected from Web Resources is used. The results are evaluated by comparing the semantic similarity measure with human ratings in three benchmark datasets: Miller-Charles (MC), Rubenstein-Goodenough (RG) and WordSimilarity-353. In the proposed work both Spearman correlation coefficient and Pearson Correlation efficient have been used as evaluation measures on semantic similarity. The semantic similarity score obtained are collected in a database.

In most of the syntactic based search systems, if the information related to the query word is not found in the database, it will not provide recommendations about the related query terms. The Semantic Search Engine helps the user by providing the similar words related to the query terms as recom-

- *Karthiga.M is currently pursuing masters degree program in computer science and engineering in Kongu Engineering College, Erode,E-mail:mkarthiga22@gmail.com*
- *Priyadharsini.M is currently pursuing masters degree program in computer science and engineering in Kongu Engineering College, Erode,E-mail:priyamoorthy@gmail.com*

mendations. The Semantic based search systems is more efficient when compared to most of the syntactic based search systems by providing the most relevant web pages to the user. In the proposed system a model Semantic Search Engine has been created which uses the technique of Semantic similarity between the words. The proposed Search Engine uses technical database related to computer terms as well as semantically similar words database. The semantic similarity between the words is calculated using the existing system and the results are updated in the database. Manually a technical database has been created which contains the technical terms with their meanings related to computer technology. A framework of the proposed system is depicted in the figure 1.

In the proposed system when an user query is given in the user interface the search module searches and provide the resultant web page.

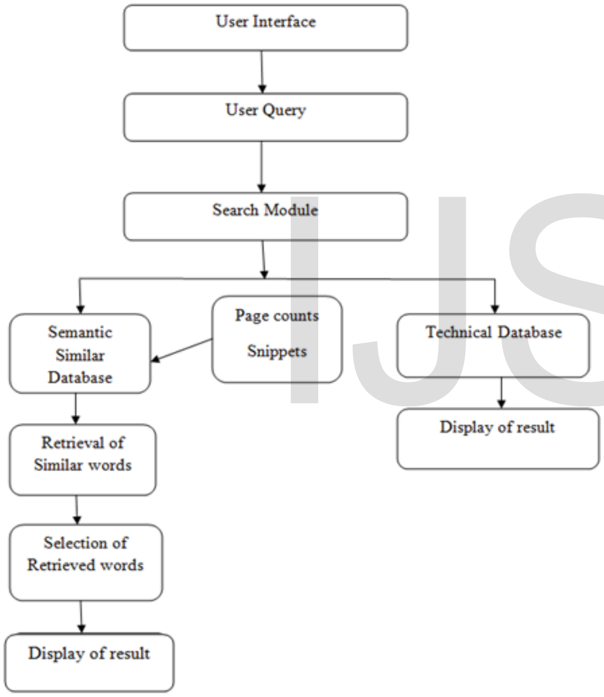


Fig. 1. Framework of the Proposed System

The steps involved in the search module for retrieving the resultant web page are as follows:

1. In a semantic search engine when a query term is given first the query term is searched in a Technical Database where the Technical Synonyms for each word is collected. If the query term matches with the Technical Database then the respective result is retrieved as a result to the user.

2. If the query term does not match with the Technical Database then it is searched in Semantic Similarity Database where the semantically similar words ae collected using page counts and text snippets collected from the web resources in existing system.

3. If the query term matches, then the semantically similar words are categorized to the user as recommendations. From the list of category the user have to select the one which the user intended to search and accordingly the result web page is returned to the user.

After measuring the semantic similarity between the words using the existing system the similar words are collected in a database. The technical words related to computer technology are collected in another database. A Semantic Search engine is created which is designed to do both normal as well as semantic search. In Normal search systems the normal way of searching and retrieving the relevant documents for the user query is performed. In Semantic search the user query keyword is searched in technical database as well as in semantic similar words database.

The time taken to retrieve the resultant pages in Semantic search is compared with Normal search. The accuracy of the resultant pages retrieved in Semantic search is high when compared to Normal search. Also the precision and recall is calculated for both normal as well as semantic search and the results are compared with both searches.

4 Experimental Works

For measuring the semantic similarity three benchmark datasets such as Miller-Charles (MC), Rubenstein-Goodenough (RG) and WordSimilarity (WS) datasets have been taken. MC dataset contains 28 pairs of words with 38 annotators and RG dataset contains 65 pairs of words with 36 annotators and WS dataset contains 353 pair of words with 13 annotators. For each pair of words page counts and text snippets are collected. For training the Support Vector Machine 3000 synonymous and 3000 nonsynonymous word pairs are collected from the WordNet.

The semantic similarity score along with the semantically similar words are collected in a semantic similarity database. The technical words related to Computer technology is collected in a Technical database.

The parameters used for evaluating the semantic similarity measure is

1. Pearson correlation coefficient
2. Spearman correlation coefficient

Pearson correlation coefficient is a measure of the correlation (linear dependence) between the variables X and Y, giving a value between +1 and -1 inclusive. Spearman correlation coefficient helps to identify the strength of correlation within a dataset of two variables and whether the correlation is positive or negative.

Pearson correlation coefficient is

$$\frac{\sum XY - \frac{\sum X \sum Y}{N}}{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)} \qquad (1)$$

Spearman correlation coefficient r = $1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$ (2)

For computing the effectiveness of the Semantic search engine the parameters that have been used for evaluation are:

1. Time
2. Precision
3. Recall

Time is the time taken by the semantic search engine to retrieve a Web page according to the user query term. Precision as per the Equation (3) is an important measure of search effectiveness. It is the ability to filter out irrelevant hits and focus on potentially useful information. In other words, Precision is the fraction of retrieved documents that are relevant to the search. Recall as per the Equation (4) measures how well a search finds every possible document that could be of interest to the searcher. In other words, Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved. The formula for calculating precision and recall are as follows:

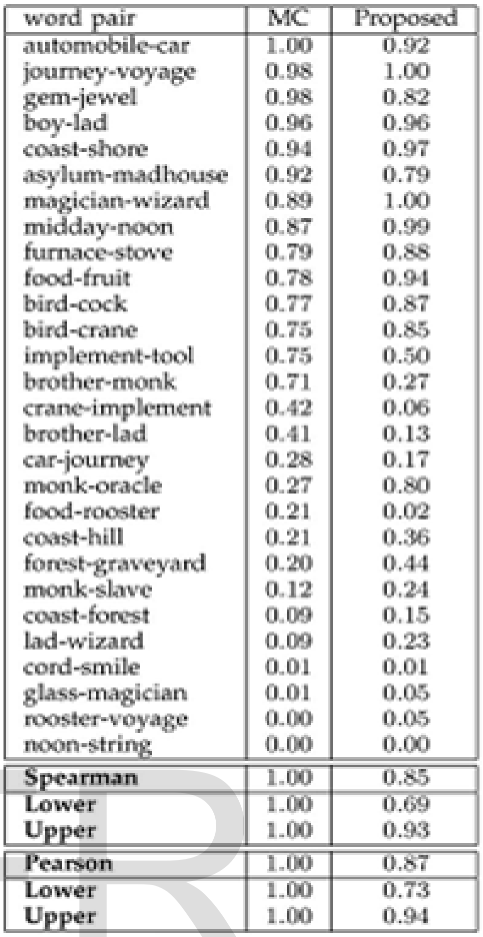
Precision is

$$\frac{\left|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}\right|}{\left|\{\text{retrieved documents}\}\right|} \qquad (3)$$

Recall is

$$\frac{\left|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}\right|}{\left|\{\text{relevant documents}\}\right|} \qquad (4)$$

The Semantic similarity score on Miller-Charles dataset using Pearson correlation coefficient and Spearman correlation coefficient is shown in the table 1.



word pair	MC	Proposed
automobile-car	1.00	0.92
journey-voyage	0.98	1.00
gem-jewel	0.98	0.82
boy-lad	0.96	0.96
coast-shore	0.94	0.97
asylum-madhouse	0.92	0.79
magician-wizard	0.89	1.00
midday-noon	0.87	0.99
furnace-stove	0.79	0.88
food-fruit	0.78	0.94
bird-cock	0.77	0.87
bird-crane	0.75	0.85
implement-tool	0.75	0.50
brother-monk	0.71	0.27
crane-implement	0.42	0.06
brother-lad	0.41	0.13
car-journey	0.28	0.17
monk-oracle	0.27	0.80
food-rooster	0.21	0.02
coast-hill	0.21	0.36
forest-graveyard	0.20	0.44
monk-slave	0.12	0.24
coast-forest	0.09	0.15
lad-wizard	0.09	0.23
cord-smile	0.01	0.01
glass-magician	0.01	0.05
rooster-voyage	0.00	0.05
noon-string	0.00	0.00
Spearman	1.00	0.85
Lower	1.00	0.69
Upper	1.00	0.93
Pearson	1.00	0.87
Lower	1.00	0.73
Upper	1.00	0.94

Table I: The semantic similarity scores on MC data set

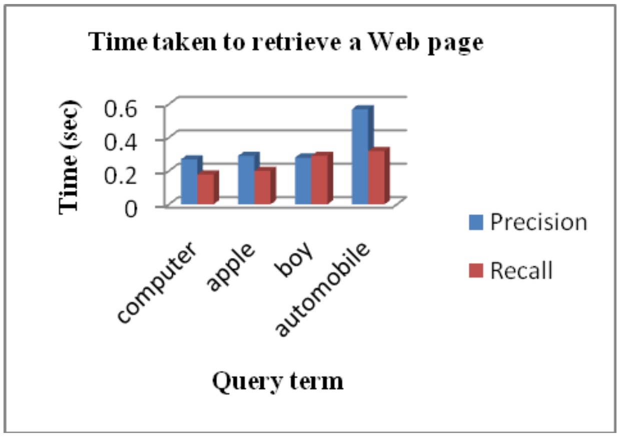


Fig. 3. The time taken by the Semantic Search Engine and Normal Search Engine to retrieve a Web page for a Query term

From the figure 3 it is clearly understood that the Semantic Search Engine will retrieve the relevant documents faster compared to the normal search engine. Since the normal search engine retrieve all the documents that contain the query term it takes time for the user to choose the relevant document from the retrieved documents whereas the Semantic Search engine will retrieve only the relevant documents using the semantic similarity measure.

The precision measure for the Semantic Search Engine and Normal Search engine is depicted in the figure 4. It is clearly understood from the figure 4 the Semantic Search Engine retrieves more relevant documents than the normal search systems. The normal search systems do not consider the semantic similarity measure while retrieving the documents. So the precision is low compared to Semantic Search Engine.

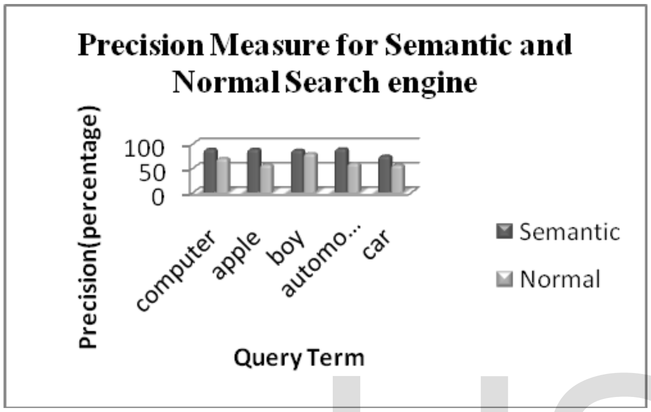


Fig. 4. Comparison between Semantic and Normal Search Engine using Precision Measure

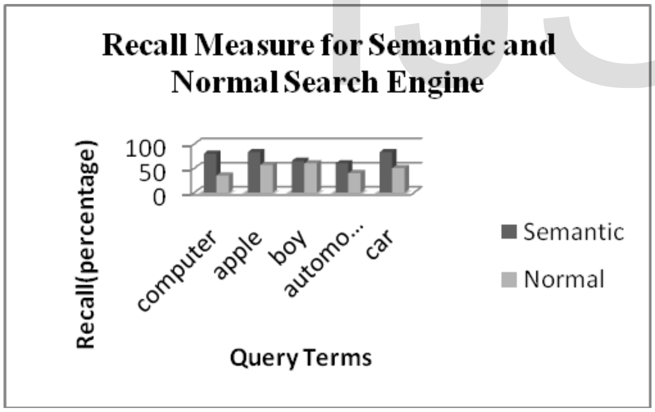


Fig. 5. Comparison between Semantic and Normal Search Engine using Recall Measure

The recall measure for the Semantic Search Engine and Normal Search engine is depicted in the figure 5. The Semantic Search Engine refines the searching according to the user selection from the recommendations listed by the search systems. So the number of documents retrieved relevant to the query are high compared to normal search systems. In normal search systems when apple is given as a query term it will retrieve all the documents related to apple as a fruit and apple as a computer. But the user intended to search as apple as a fruit so the user have to select the relevant documents from the retrieved result. So the recall measure is low for normal search systems compared to semantic search systems.

6 CONCLUSION AND FUTURE WORK

In this paper, a model Semantic search engine is created which provides the semantically similar words for the query words as recommendations. The semantic similar words are calculated by using both page counts and text snippets. The proposed work uses both Pearson correlation coefficient and Spearman correlation coefficient for evaluating the semantic similarity measure. A Technical Database for Computer Technology is created which provides the technical meanings for the query words. The semantic similarity database contains the semantically similar words. The semantic search engine evaluates and retrieves the resultant webpage by using technical as well as semantic similarity database. Precision, Recall and time taken to retrieve the webpage is compared with normal search engine and the results are obtained.

Further the work can be extended by internally modifying the user given query by the semantically similar query by the search engine to modify the original query. The Semantic Search engine has also been further used in the process of Query expansion that is a user query is modified using synonymous words to improve the relevancy of the search.

REFERENCES

[1] Bollegala D, Matsuo Y, and Ishizuka M (2011),"*Measuring semantic similarity between words using web search engines*", IEEE Transactions on Knowledge and Data Engineering, vol.23, Issue 7, pp.977-990.

[2] Chen H, Lin M, and Wei Y (2006), "*Novel Association Measures Using Web Search with Double Checking*", Proceedings of the 21st International Conference on Computational Linguistics, pp. 1009-1016.

[3] Cilibrasi R and Vitanyi P (2007), "*The Google Similarity Distance*," IEEE Transactions on Knowledge and Data Engineering, vol. 19, Issue 3, pp. 370-383.

[4] Church K and Hanks P (1991)," *Word Association Norms, Mutual Information and Lexicography*," Computational Linguistics, vol. 16, pp. 22-29.

[5] Hearst M (1992), "*Automatic Acquisition of Hyponyms from Large Text Corpora*," Proceedings of the 14th Conference on Computational Linguistics (COLING), pp. 539-545.

[6] Hirst G and St-Onge D (1998), "*Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms*," WordNet: An Electronic Lexical Database, pp. 305-332, MIT Press.

[7] Hughes T and Ramage D (2007), *"Lexical Semantic Relatedness with Random Graph Walks*," Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07), pp. 581-589.

[8] Imen Akermi and Rim Faiz (2012), "*Semantic similarity measure based on multiple resources*", Proceedings of the International Conference on Information Technology and e-Services, pp.546-550.

[9] Kilgarriff A, "*Googleology Is Bad Science (2007*)," Computational Linguistics, vol. 33, pp. 147-151.

[10] Lapata M and Keller F (2005), "*Web-Based Models for Natural Language Processing*," ACM Transaction Speech and Language Processing, vol. 2, no. 1, pp. 1-3.

[11] Lin D (1998), "*An Information-Theoretic Definition of Similarity*," Proceedings of the 15th International Conference on Machine Learning (ICML), pp. 296-304.

[12] Matsuo Y, Sakaki T, Uchiyama K and Ishizuka M (2006)," *Graph-based word clustering using web search engine*", Proceedings of EMNLP, pp. 523-530.

[13] Mclean D, Li Y, and Bandar Z. A (2003), "*An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources,*" IEEE Transactions on Knowledge and Data Engineering, vol. 15, Issue 4, pp. 871-882.

[14] Pei T, Han J, Mortazavi-Asi B, Wang J, Pinto H, Chen Q, Dayal U, and Hsu M (2004), "*Mining Sequential Patterns by Pattern- Growth: The Prefixspan Approach*," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 11, pp. 1424-1440.

[15] Pasca M, Lin D, Bigham J, Lifchits A, and Jain A (2006), "*Organizing and Searching the World Wide Web of Facts - Step One: The One-Million Fact Extraction Challenge*," Proceedings of National Conference on Artificial Intelligence (AAAI '06).

[16] Rada R, Mili H, Bichnell E, and Blettner M (1989), "*Development and Application of a Metric on Semantic Nets*", IEEE Transaction Systems, Man and Cybernetics, vol. 19, Issue 1, pp. 17-30.

[17] Resnik P (1995), "*Using Information Content to Evaluate Semantic Similarity in a Taxonomy*", Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp.448-453.

[18] Rosenfield R (1996), "*A Maximum Entropy Approach to Adaptive Statistical Modelling,*" Proceedings on Computer Speech and Language, vol. 10, pp. 187-228.

[19] Sahami M and Heilman T (2006), "*A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets*", Proceedings of the 15th International World Wide Web Conference, pp.326-331.

[20] Schickel-Zuber V and Faltings B (2007), "*OSS: A Semantic Similarity Function Based on Hierarchical Ontologies*," Proceedings of International Joint Conference on Artificial Intelligence (IJCAI '07), pp. 551-556.

[21] Siddharth P, Banerjee S and Pedersen T (2003),"*Using measures of semantic relatedness for word sense disambiguation*", Proceedings of the Fourth International Conference on Intelligent on Text Processing and Computational Linguistics, Mexico City, Mexico, pages 241-257.

[22] Snow R, Jurafsky D, and Ng A (2005), "*Learning Syntactic Patterns for Automatic Hypernym Discovery*," Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 1297-1304.

[23] Strube M and Ponzetto S.P (2006), "*Wikirelate! Computing Semantic Relatedness Using Wikipedia*," Proceedings of National Conference on Artificial Intelligence(AAAI '06), pp. 1419-1424.

[24] Turney P.D (2001), "*Mining the web for synonyms: Pmi-ir versus lsa on toefl*", proceedings of ECML, pp. 491–502.

[25] Wu Z and Palmer M (1994), "*Verb Semantics and Lexical Selection*," Proceedings of Ann. Meeting on Assoc. for Computational Linguistics (ACL '94), pp. 133-138.