A Web Search Engine based approach to Measure the Semantic Similarity between Words using Page Count and Snippets Method (PCSM)

Ms. Vaishali Nirgude, Dr. Rekha Sharma, Dr. R.R. Sedamkar

Abstract— **Semantic similarity measures play an important role in Information Retrieval, Natural Language Processing and Web Mining applications such as community mining, relation detection, entity disambiguation and document clustering etc. This paper proposes Page Count and Snippets Method (PCSM) to estimate semantic similarity between any two words (or entities) based on page counts and text snippets retrieved from a web search engine. It defines five page count based concurrence measures and integrates them with lexical patterns extracted from text snippets. A lexical pattern extraction algorithm is proposed to identify the semantic relations that exist between any query word pair. Similarity score of both methods are integrated by using Support Vector Machine (SVM) to get optimal results. The proposed method is compared with Miller and Charles (MC) benchmark data sets and the performance is measured by using Pearson correlation value. The correlation value of proposed method is 0.8960% which is higher than existing methods. The PCSM also evaluates semantic relations between named entities to improve Precision, Recall and F-score.**

Index Terms— **Community Mining, Information Retrieval, Lexical Patterns, Page Counts, Text Snippets, Correlation.**

I. INTRODUCTION

Information retrieval (IR) and Natural Language Processing (NLP) are two important aspects involved in all web mining applications. Search engines have become the most helpful tool for obtaining useful information from the Internet. The search results returned by even the most popular search engines are not satisfactory. It surprises users because they do input the right keywords and search engines do return pages involving these keywords, and the majority of the results are irrelevant. To evaluate the effectiveness of a Web search mechanism in finding and ranking results, measures of semantic similarity are needed. The semantic similarity between words can be resolved using dictionaries. But accurately measuring the semantic similarity between two words (or entities) on WWW is a challenging task because Semantic similarity between entities changes over time and across domains. For example, blackberry is frequently associated with phones on the Web. However, this sense of blackberry is not listed in thesauri or dictionaries.

The Proposed Page Count and Snippets (PCSM) method is an automated method to measure semantic similarity between words or entities using Web search engines. Page counts and Snippets are two useful information sources provided by most Web search engines. Page count of a query is the number of pages that contain the query words. Page count for the query W1 AND W2 can be considered as a global measure of co-occurrence of words W1 and W2. e.g. the page count of the query “blackberry" AND “phone" in Google is 605,000,000. Whereas the same for “strawberry" AND “phone" is only 58,600,000. Page counts for “blackberry" AND “phone" is 10 times more than the page counts for “strawberry" AND “phone”. It indicates that blackberry is more semantically similar to phone than the strawberry.

Though page count is simple method to measure semantic similarity between words but it has several drawbacks. Page count ignores the position of a word within a page. Therefore, even though two words appear in a page, they might not be actually related. Page count of a polysemous word (a word with multiple senses) might contain a combination of all its senses. For example, page counts for *apple* contain page counts for apple as a fruit and apple as a company. Hence, page count method alone is unreliable when measuring semantic similarity.

Snippets, a brief window of text extracted by a search engine around the query term in a document, provide useful information regarding the local context of the query term. Downloading large size documents can be avoided by using snippets. However, main drawback of using snippets is that, only those snippets for the top-ranking results for a query can be processed efficiently.

Manuscript received July, 2013.

Ms. Vaishali Nirgude, M..E. Student, Department of Computer Engineering, Thakur College of Engineering and Technology, Mumbai, India.

Dr. Rekha Sharma, Associate Professor, Department of Computer Engineering, Thakur College of Engineering and Technology, Mumbai ,India.

Dr.R.R.Sedamkar, Professor, H.O.D., *Department of Computer Engineering, Thakur College of Engineering and Technology, Mumbai, India.*

The Proposed PCSM use both page counts and lexical patterns extracted from snippets to overcome the problems described above. For example, let us consider the following snippet from Google for the query “Apple And Computer”.

> **Apple Inc.,** formerly **Apple Computer, Inc.,** is an American multinational corporation headquartered in California that designs, develops, and sells computer software and personal computers.

Fig.1: Snippet retrieved for the query “Apple and Computer”

Here, the phrase “is an” indicates a relationship between Apple and Computer. Phrases such as ‘also known as’, ‘is a’, ‘part of’, ‘is an example of’ all indicate various semantic relations.

***Contribution*:**

The Proposed PCSM is based on a Lexical Pattern Extraction and a Pattern Clustering algorithm to find semantic similarity measure between words. It has defined five page count concurrence measures such as Web Dice, Web Overlap, Web Jaccard, WebPMI and NGD to find semantic similarity between words. Support Vector Machine [1] (SVM) integrates semantic similarity score of Page Count and Snippet Methods to find optimal semantic similarity score. The proposed PCSM is compared with Miller and Charles (MC) benchmark data sets [2] and the performance is measured by using Pearson correlation value. The Proposed method is also evaluated for named entities to improve the Precision, Recall, and F-score.

II. LITERATURE SURVEY

Information Resources are very important factor for measuring the semantic similarity between words. Wordnet and web search engines are two important information resources to measure semantic similarity between words.

A. Traditional Ontology based methods

Ontology-based semantic similarity measurement methods are the ones which uses ontology source as the primary information source. They can be roughly classified into three groups as follows:

1. Distance based method

Find Shortest Path between words: Given taxonomy of words, a straightforward method to calculate similarity between two words is to find the length of the shortest path connecting the two words in the taxonomy [8]. It estimates the distance (e.g. edge length) between nodes which corresponds to the concepts being compared.
Drawback: The problem with this approach is that it considers a uniform distance for all links in taxonomy.

2. Information content based method

Resnik [3] proposed a similarity measure using information content. Similarity of two concepts is based on the extent to which they share common information. The similarity between two concepts C1 and C2 in the taxonomy is checked based on maximum information from the given set of concept, C. WordNet is being used as the taxonomy.
Lin [4] calculates semantic similarity using a formula derived from information theory.
Drawback: Word Sense Disambiguation (WSD) problem is not discussed.

3. Distance and Information Content based method

Li et al. [5] combined structural semantic information from a lexical taxonomy and information content from a corpus in a nonlinear model. They proposed a similarity measure that uses shortest path length, depth, and local density in taxonomy.
Jiang and Conrath [6] presented an approach for measuring semantic similarity between words and concepts. This method is a combined approach which inherits the edge-based approach of the edge counting scheme, which is then enhanced by the node-based approach of the information content measurement.
Drawback: They did not evaluate their method in terms of similarities among named entities.

B. Web Search Engines based methods

By using the existing ontology for measuring semantic similarity between words there is a limitation of new words. So to overcome this limitation many researchers have worked on web.

1. Snippets based Method

Sahami and Heilman [7] measured semantic similarity between two queries using snippets returned by a search engine. Snippets from a search engine are collected and represented in the form of TF-IDF (Term Frequency – Inverse Document Frequency) weighted term vector. High TF gives more semantic similarity and IDF gives less semantic similarity between words.
Drawback: Only top ranking results for a query can be processed efficiently.

Double Checking model: Chen et al. [8] proposed a double-checking model using text snippets returned by a web search engine. Two objects are considered to be associated if one can be found out from the other one using web search engines, e.g. the Co-occurrence Double-Checking (CODC) measure is defined as:

$$\mathrm{CODC}(W_1, W_2)$$

$$= \begin{cases} 0 & (W_1@\,W_2) = 0, \\ \exp\left(\log\left[\frac{f(W_1@\,W_2}{H(W_1)} \mathrm{x} \frac{f(W_2@\,W_1)}{H(W_2)}\right]^{\alpha}\right), & \text{otherwise} \end{cases} \quad (1)$$

Drawback: The major problem with this approach is that we cannot assure the occurrence of one word in the snippet for the other even though they are related.

2. Page counts based Method

Normalized Google Distance (NGD): Cilibrasi and Vitanyi [9] proposed a distance metric between words using page counts retrieved from a web search engine. The proposed metric is named Normalized Google Distance (NGD) and is given by:

$$\mathrm{NGD}(W_1, W_2) = \left\{ \frac{\max\{\mathrm{iog}H(W_1), \log H(W_2)\} - \log H(W_1, W_2)}{\log N - \min\{\log H(W_1), \log H(W_2\}} \right. \qquad (2)$$

Here, W1 and W2 are the two words between which distance NGD (W1, W2) is to be computed, H (W1) *denotes* the page count for the word W1, H (W1, W2) is the page count for the query W1 AND W2. NGD is fully based on normalized information distance, which is defined using Kolmogorov complexity.
Drawback: This method based on Page count hence considers only global occurrences of words and ignores position of a word within a page.

3. *Snippets and Page Counts Based Method*

D.Bollegala.et.al: Bollegala, Matsuo and Ishizuka [10] have proposed an automatic method to estimate the semantic similarity between words or entities using web search engines. D.Bollegala measured semantic similarity between two words using snippets and page counts returned by a search engine [10][11].They defined four different Page co-occurrence metrics (WebJaccard, WebDice, WebOverlap, WebPMI) to calculate semantic similarity and integrate those with the lexical patterns extracted from text snippets.

C. Comparisons of Semantic Similarity Methods:

On the basis of the results calculated according to the approaches by different researchers, fig. 2 and fig.3 shows the similarity methods and their respective correlation result.

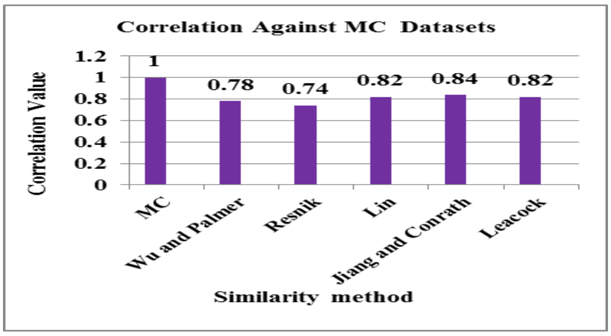


Fig. 2: Correlation of Semantic Similarity method against MC Datasets based on Wordnet

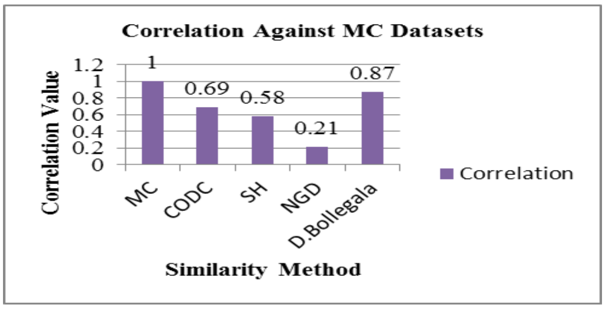


Fig. 3: Correlation of Semantic Similarity method against MC Dataset based on web search engines

Semantic similarity measures are important in many applications such as IR, NLP and various web mining tasks. Hence, we required more efficient semantic similarity measures techniques.

III. PROPOSED PCSM SYSTEM

The Proposed PCSM system based on Page Count and text Snippets retrieved by a web search engine.

A. Problem Definition:

We measure the semantic similarity between given two words W1 and W2 by assigning the weight in the range of [0, 1]. If W1 and W2 are highly similar (e.g., synonyms), we expect SemSim (W1, W2) to be closer to 1. On the other hand, if W1 and W2 are not semantically similar, then we expect SemSim(W1, W2) to be closer to 0.We measure similarity between W1 and W2 using page counts and snippets retrieved from a web search engine for the two words.
To get better results, page counts-based co-occurrence measures is integrated with lexical pattern clusters using support vector machine. Performance of these methods is evaluated using MC benchmark data sets. The main objective is to find the semantic similarity between two words and improving the correlation value with the MC benchmark data sets.

B. System Architecture:

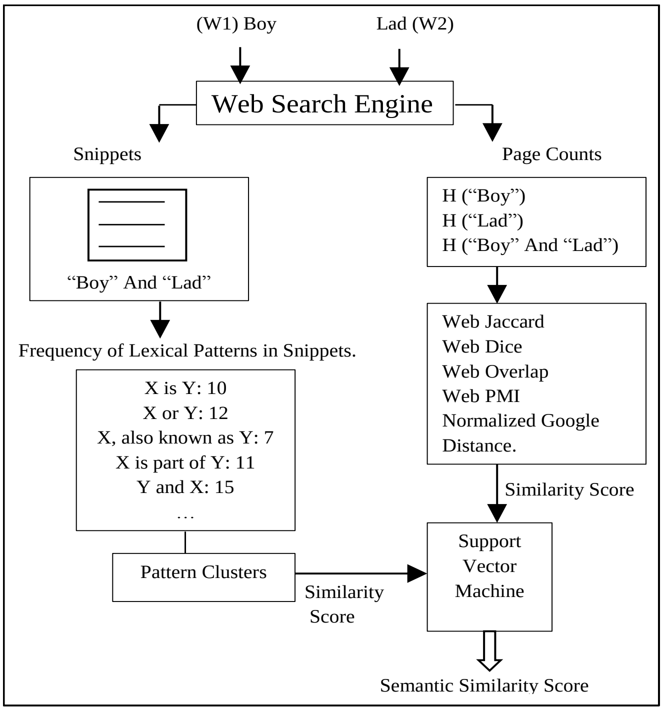


Fig. 4: Outline of the proposed PCSM method.

Fig. 4 illustrates an example using the proposed PCSM to compute the semantic similarity between two query words (i.e. Boy and Lad).

C. Page Count Co-occurrence Measure

There are five popular co-occurrence measures; Jaccard, Overlap (Simpson), Dice, and Pointwise mutual information (PMI), Normalized Google Distance (NGD), to compute semantic similarity using page counts. The WebJaccard coefficient between words (or multiword phrases) W1 and W2 is defined as:

$$Webjaccard(W_1, W_2) = \begin{cases} 0, & H(W_1 \cap W_2) \leq c, \\ \dfrac{H(W_1 \cap W_2)}{H(W_1) + H(W_2) - H(W_1 \cap W_2)}, & \text{otherwise} \end{cases} \quad (3)$$

Here, W1 ∩ W2 denotes the conjunction query W1 AND W2. H (W1) is page count for W1, H (W2) is page count for W2 and H (W1 ∩ W2) is page count for both W1 and W2. We set c = 5 experimentally.

Similarly, we define Web Overlap (W1, W2):

$$\text{WebOverlap}(W_1, W_2) = \begin{cases} 0, & H(W_1 \cap W_2) \leq c, \\ \dfrac{H(W_1 \cap W_2)}{\min\,[H(W_1), H(W_2)]}, & \text{Otherwise} \end{cases} \quad (4)$$

We define the Web Dice coefficient as a variant of the Dice coefficient. Web Dice (W1, W2) is defined as:

$$\text{WebDice}(W_1, W_2) = \begin{cases} 0, & H(W_1 \cap W_2) \leq c, \\ \dfrac{H(W_1 \cap W_2)}{H(W_1) + H(W_2)}, & \text{Otherwise} \end{cases} \quad (5)$$

Point wise mutual information [12] is a measure that is motivated by information theory. We define WebPMI as a variant form of point wise mutual information using page counts as:

$$\text{WebPMI}(W_1, W_2) = \begin{cases} \log\left(\dfrac{\frac{H(W_1 \cap W_2)}{N}}{\frac{H(W_1)}{N}\frac{H(W_2)}{N}}\right) & 0, \quad H(W_1 \cap W_2) \leq c, \\ , & \text{Otherwise} \end{cases} \quad (6)$$

Here, N is the number of documents indexed by the search engine. We set N= 10^{10} according to the number of indexed pages reported by Google.

Normalized Google Distance (NGD) is a distance metric between words and is defined as:

$$NGD(W_1, W_2) = \begin{cases} \dfrac{\max\{\log H(W_1), \log H(W_2)\} - \log H(W_1, W_2)}{\log N - \min\{\log H(W_1), \log H(W_2)\}} \end{cases} \quad (7)$$

NGD is based on normalized information distance [10] which is defined using kolmogorov complexity.

D. Lexical Pattern Extraction Algorithm

Page counts actually represent the global co-occurrence of two words on the web and do not consider the local context in which those words co-occur. This can be problematic if one or both words are polysemous. On the other hand, the snippets returned by a search engine for the conjunctive query of two words provide useful clues related to the semantic relations that exist between two words. A snippet contains a window of text selected from a document that includes the queried words, e.g. consider the snippet in Fig. 5. Here, the phrases ‘also known as’ and ‘is a. indicates a semantic relationship between Apple and Computer.

> The original *Apple Computer*, also known retroactively as the Apple I, or Apple-1, is a personal computer released by the *Apple Computer* Company, released by the *Apple Computer* Company

Fig. 5: Snippet retrieved for the query “Apple and Computer”.

For a snippet S, retrieved for a word pair (W1, W2), First, we replace the two words W1 and W2, respectively, with two variables X and Y.

```
Algorithm1: Lexical Pattern Extraction Algorithm
Input: Snippets S returned by the web Search engine for
query words W1 and W2.
Output: Lexical patterns (P) & Frequency (F).

Step 1: Read each snippet S, store it in database & perform
data cleaning operation.
Step 2: for each snippet S do
            if word is same as W1 then
            Replace W1 by X
            end if
            if word is same as W2 then
            Replace W2 by Y
            end if
        end for
Step 3: for each snippet S do
        if X € S or Y€ S then extract all patterns from S
        end if
        end for
Step 4: for each snippet S do
        if X And Y € S or Y And X € S then extract all
            patterns from S
        end if
        end for
Step 5: for each pattern do
        Perform stemming operation.
    end for
Step 6: Find the frequency of all repeated extracted
        Patterns.
Step 7: Return Patterns (P) & Frequency (F).
```

A semantic relation can be expressed using more than one pattern. By identifying such patterns that convey the similar relation helps to represent the relation between words accurately. A lexical pattern clustering algorithm is used to cluster together such patterns.

IV. COMMUNITY MINING

Measuring the semantic similarity between named entities is essential in many applications such as query expansion, entity disambiguation (e.g., namesake disambiguation) and community mining. The proposed semantic similarity measure is evaluated for these applications because it does not require precompiled taxonomies. In order to evaluate the performance of the proposed measure in capturing the semantic similarity between named entities, we set up a community mining task. By clustering semantically related lexical patterns, we see that both precision as well as recall can be improved in a community mining task.

Evaluation of PCSM for Community Mining:

Step1. On sample basis we have formed three clusters i.e. three communities: Sports, Politician and Actors (Initially all clusters are empty).
Step2. For each pair of names, their semantic similarity measured using the PCSM.
Step3. Based on semantic similarity score, add name word pair into specific cluster and also adds in database.
Step4. If the word pair is not matching with the specific cluster (i.e. other than three communities), display no specific cluster found.
Step5. Repeat step 2, 3 and 4 for 15 different name word pairs.
Step6. Finally we will get three clusters which include specific name word pairs.

We compute precision, recall and *F*-Measure for each name in the data set and average the results over the dataset.

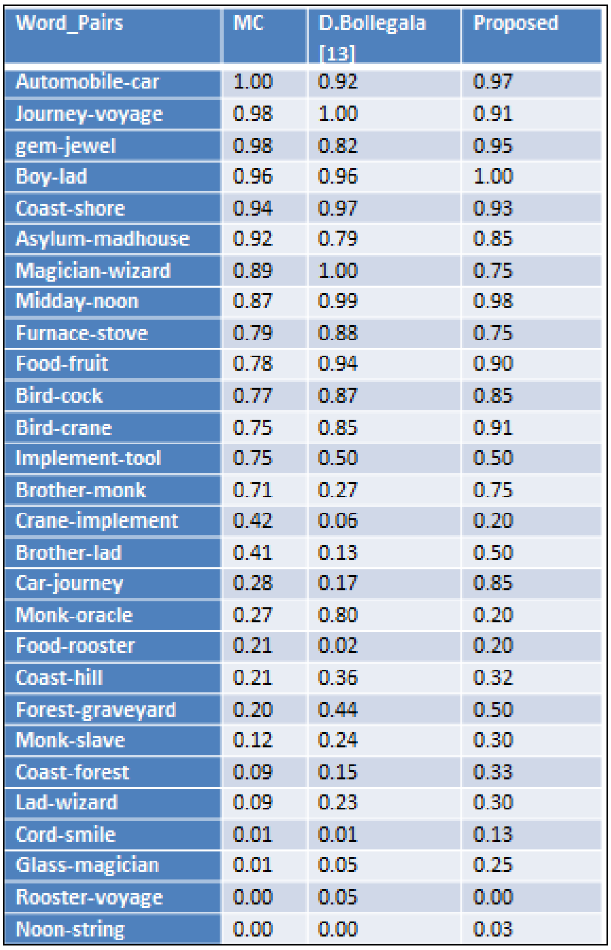
V. RESULTS AND DISCUSSION

The performance of the PCSM semantic similarity measure is evaluated by conducting two sets of experiments. Firstly, compare the similarity scores produced by the proposed measure against the Miller –Charles (MC) benchmark datasets [2] of 28 word-pairs rated by a group of 38 human subjects. The word pairs are rated on a scale from 0 (no similarity) to 1 (perfect synonymy). Then analyze the performance of the proposed measure by using Pearson Correlation coefficient. Secondly, apply the proposed measure in a real-world named entity clustering task and measure its performance.

A. *Experiment 1*

1. Comparison of semantic similarity score of proposed PCSM system against MC Datasets and existing method.

Table 1: Semantic Similarity Score on MC-data sets



Word_Pairs	MC	D.Bollegala [13]	Proposed
Automobile-car	1.00	0.92	0.97
Journey-voyage	0.98	1.00	0.91
gem-jewel	0.98	0.82	0.95
Boy-lad	0.96	0.96	1.00
Coast-shore	0.94	0.97	0.93
Asylum-madhouse	0.92	0.79	0.85
Magician-wizard	0.89	1.00	0.75
Midday-noon	0.87	0.99	0.98
Furnace-stove	0.79	0.88	0.75
Food-fruit	0.78	0.94	0.90
Bird-cock	0.77	0.87	0.85
Bird-crane	0.75	0.85	0.91
Implement-tool	0.75	0.50	0.50
Brother-monk	0.71	0.27	0.75
Crane-implement	0.42	0.06	0.20
Brother-lad	0.41	0.13	0.50
Car-journey	0.28	0.17	0.85
Monk-oracle	0.27	0.80	0.20
Food-rooster	0.21	0.02	0.20
Coast-hill	0.21	0.36	0.32
Forest-graveyard	0.20	0.44	0.50
Monk-slave	0.12	0.24	0.30
Coast-forest	0.09	0.15	0.33
Lad-wizard	0.09	0.23	0.30
Cord-smile	0.01	0.01	0.13
Glass-magician	0.01	0.05	0.25
Rooster-voyage	0.00	0.05	0.00
Noon-string	0.00	0.00	0.03

2. Comparison of Pearson Correlation value of proposed PCSM system with MC data set & existing method. Our Proposed system has achieved 0.8960 correlation value as shown in fig.6.

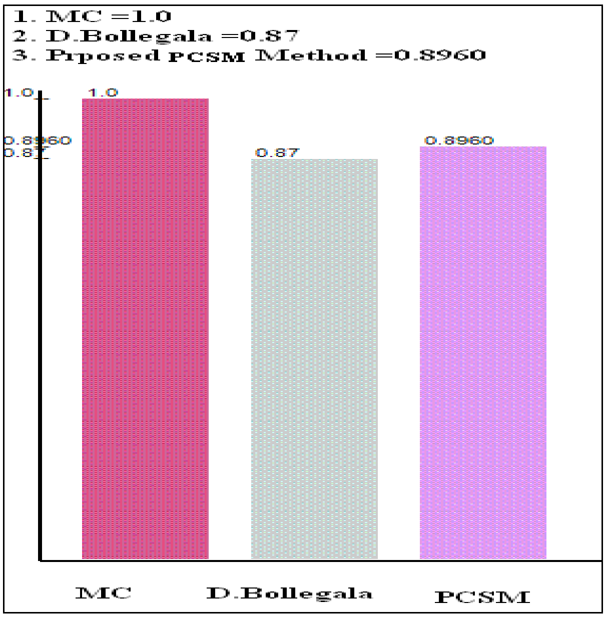
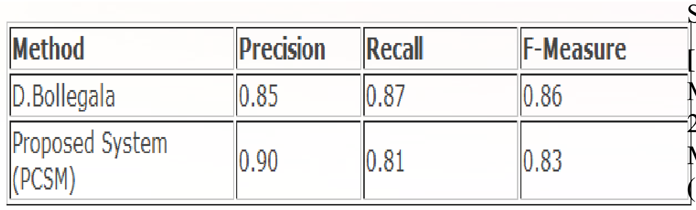


Fig.6: Comparison of Pearson correlation value

B. *Experiment 2*

We have evaluated the proposed system for community mining (named entities) to improve precision, recall and F-score. High Recall means that an algorithm returned most of the relevant results. High precision means that an algorithm returned more relevant results than irrelevant. F-measure is computed based on the precision and recall evaluation metrics. We have used dynamic approach to find specific community cluster .Table 2 shows that precision. Recall and f-measure of PCSM are closer to existing D.Bollegala method.

Table 2: Comparison of Precision, Recall and F-Measure value of PCSM method with existing method



Method	Precision	Recall	F-Measure
D.Bollegala	0.85	0.87	0.86
Proposed System (PCSM)	0.90	0.81	0.83

VI. CONCLUSION AND FUTURE SCOPE

The proposed PCSM method is based on both page counts and snippets to calculate semantic similarity between two given words. The PCSM method consists of five page count co-occurrence measures and lexical pattern extraction algorithm. Semantic similarity score of proposed method is compared with Miller- Charles (MC) benchmark data sets and the performance is evaluated by using Pearson correlation value. The correlation value of PCSM is 0.8960 which is higher than existing D. Bollegala method and closer to MC benchmark data sets. The proposed method is also applied on real world named entities to improve precision, recall and F-measure.

The existing results can be improved by taking into account more information resources or combination of them for measuring semantic similarity between words.

In future research, the proposed methods can be enhanced to measure semantic similarity in automatic synonym extraction, query suggestion and name alias recognition applications.

REFERENCES

[1] Thorsten Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features".

[2] G.Miller and W.Charles, "Contextual Correlates of Semantic similarity," Language and Cognitive Processes, vol.6,no,l,pp.1-28,1998.

[3] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," Proc. 14th Int'l Joint Conf. Artificial Intelligence, 1995.

[4] R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and Application of a Metric on Semantic Nets," IEEE Trans. Systems, Man and Cybernetics, vol. 19, no. 1, pp. 17-30, Jan./Feb. 1989.

[5] D.Mclean, Li.Y and Z.A.Bandar, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", IEEE Trans. Knowledge and Data Eng., 2003, vol. 15, no. 4, pp. 871-882

[6] J. Jiang & D. Conrath, (1997) "Semantic similarity based on corpus statistics and lexical taxonomy", Proceeding of International Conference on Research in Computational Linguistics (ROCLING X).

[7]M. Sahami and T. Heilman, "A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets," Proc. 15th Int'l World Wide Web Conf., 2006.

[8] H. Chen, M. Lin, and Y. Wei, "Novel Association Measures Using Web Search with Double Checking," Proc. 21st Int'l Conf. Computational Linguistics and 44th Ann. Meeting of the Assoc. for Computational Linguistics (COLING/ACL '06), pp. 1009-1016,2006.

[9] R. Cilibrasi and P. Vitanyi, "The Google Similarity Distance," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 370-383, Mar. 2007.

[10] Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka, "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words" IEEE , 2011.

[11] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring Semantic Similarity between Words Using Web Search Engines," Proc. Int'l Conf. World Wide Web (WWW '07), pp. 757-766, 2007

[12] K.Church and P.Hanks," Work association Norms, Mutual Information and Lexicography," Computational Linguistics, vol.16,pp .22-29,1991.

Ms. Vaishali Nirgude is presently working as Assistant Professor in Computer Engineering Department, Thakur College of Engineering & Technology, Mumbai. She has completed her B.E. in Computer Engineering from A.V.C.O.E. Sangamner (Pune University). She is currently pursuing M.E in Computer Engineering from Thakur College of Engineering & Technology. She has published 1 paper in National conference and presented 2 papers in International Conference.

Dr. Rekha Sharma is presently working as Associate Professor and Dy. HOD in Computer Engineering Department, Thakur College of Engineering & Technology, Mumbai. She is having 13 years of experience in teaching & research. Her area of interest includes S/W Localization, Natural Language Processing and Computer Graphics. She has guided more than 20 undergraduate & postgraduate projects. She has published/presented more than 14 papers in National / International Conferences /Journals.

Dr.R.R.Sedamkar is presently the Professor, Dean Academics & HOD in Computer Engineering Department ,Thakur College of Engineering & Technology, Mumbai. He has guided more then 50 undergraduate & Post graduate Project. He has published/ presented more than 15 papers in National / International / Journals. He has also visited USA and UK for Collaborative Programmes and headed Engineering Programmes of the Kingston University, London and has independently set up NMIMS University's off-campus centre at Shirpur.