Finding Similar Content Posts Using Semantic Textual Similarity Based On Text Segmentation Through Natural Language Processing

Rohan C. Tadvi, Vrushali A. Chakkarwar

Abstract: Posts in the forums are dispersed in database where determining the congruence among the text posts in web forums is cumbersome task. Congruence is relevant property while text clustering and text classification. Traditionally the documents were searched with the collation of keywords or set of terms from the posts. Proposed system posts are contemplated as corpus of words where each entity in corpus has some individual weightage where terms and words are also found in another corpuses as well. To fulfill the objective with common goal there should be some relatedness among the corpus of different posts in different or same forum which provides the similar motive the user needed to deliver. Congruence is calculated by applying a score to common terms calculated in preprocessing. Semantic relatedness score of corpus differs for every corpus depending on the relatedness in corpuses. Posts are divided into segments at particular instances. Of these segments the corpuses are created and text features are extracted and monitored by identifying congruence of keywords. The common terms extracted are evaluated using process by combination of different Semantic Textual algorithms. After calculating the similarity most identical posts are displayed to user on threshold basis.

Index Terms: congruence, relatedness, clustering, classification, semantic, corpuses, forum, preprocessing.

1 INTRODUCTION

The phenomenon of manipulating automatically the Natural Language content like text or speech is called Natural Language Processing (NLP).Natural Language can be referred as any communication way where people exchange information with each other by speech, E-mails, posts , SMS, web pages, blogs, signs, letters etc. In computer science NLP is the potential of computer to understand the human language spoken as it is. Due to different linguistic structure, accent, dialects, social context, complex variables and regional impact there are many ambiguities which makes its strenuous to process. For any language syntax which is grammatical sense of sentence and semantics which is meaning of words are necessary to make it understand. All syntax and semantics techniques are explained in the later part of the paper. Previously NLP were rule-based where some algorithms of machine learning were applied where if certain specific phrases or words occurred and response was given accordingly. When two words are similar or related in a given context human brain can effortlessly find it .When there is functional association among the given of terms other than the lexical relation then those two terms are in semantic relatedness. Nowadays NLP is based on deep learning where massive amount of labeled data is trained to identify the correlations between the data for analysis purposeInternet forums are the message boards on Internet where people provide exclusive answers to the queries asked by other users added on that forum to get the views, reviews and added solutions to the questions asked regarding their specific problems.

Online forums have emerged as an extensive solution providers for different user communities in various domains related to technology, medicine, product quality, tourism, law, medicine, education etc. by exploring other user's experience of their view on that particular products and services. Business organizations use this existing data as a market place to improve their services by analyzing the trending situations about their products, which helps them to aid their customer base. Organizations classify posts in forums according to their different aspects related to reviews, support for customer, topics with through specialization still with vast database it becomes cumbersome and time-consuming process for user, when he tries to search a given topic with keyword. Still keywords play an important role in searching the relevant information there are some cases where user intents to search the past using the related keyword or the brevity of the keyword itself is not sufficient to search the user specified information. For better support to users in forums an important functionality is to initiate to deliver relevant posts related to their topic based on the similarity of the posts without browsing for long duration and generating complex queries. For e.g. Searching the hotels nearby Railway Station and Finding the hotels near the station should display similar results. Normal users are not technically competent so they use synonym as a keyword. With aid of such functionality user having difficulty in finding the related posts can search related posts in other forums as substitute solutions to similar situations. Due to thematic categorization of contents under specific categories of some technical forums like finding faults or reviews of particular accessory like mobiles, printers, tablets or laptops the similarity finding is not that effective or in tourism forums it sometimes become difficult to find the hotels at nearby places on current location of user there may be substitute solution in other forum. The traditional search methods are related to searching in hierarchical form from generalization to specific category (Computers->hardware->printers->brand->model->specific problems). There can be n kind of problems with multiple suggestions from the k users in the forums but still user is not able to find the appropriate solution to the given problem. In the proposed system the relatedness of the forum posts is estimated using NLP treating the posts as corpus of

- *Rohan Chandan Tadvi is currently pursuing masters degree program in Computer Science and Engineering in Government College of Engineering Aurangabad, 431005, India. E-mail: rtadvi98@gmail.com*
- *Vrushai A. Chakkarwar has Completed masters degree program in Computer Science and Engineering. Current designation is Assistant Professor of Computer Science and Engineering in Government College of Engineering Aurangabad, 431005, India. E-mail: vrsuh..a143@gmail.com*

segments. Corpus consists of constituent pieces of natural language documented on base of its origin of that language. To find the relatedness among the given segments NLP framework will play an important role. Semantic similarity can be document level, paragraph level, sentence level and at word level.

2 RELATED WORK

Development of applications related to systems like Text-Classification, Information-Retrieval, Document-Clustering, Topic-Tracking, Topic-Detection, Machine-Translation, Text-Summarization, Question-Answering systems necessiate large-scale research where similarity in texts play a vital role. These text similarity measures can broadly be classified on String-based, Corpus-Based and Knowledge based [2]. Researchers combined knowledge based and corpus based methods to get better performance results known as Semantic Text Similarity models [5]. Online user forums provide an input for business analytics to improve the service by pointing out the user search on specific products where Association Rule Mining becomes aided from it[1]. Dimitri Papadimitrio , G. Koutrika etc. the authors in their work has given emphasis on segmentation phase by segmentation of posts by different possible segments related to tense, subject ,thematic ,intention based, change from first person to another[1]. Segmentation of posts determine coherence, depth and border selection which is helpful for searching the intent based segments. These segments of posts of total posts were clustered in one group using K-means algorithm where the similarity is found on the clustered data using different term based similarity measures[1].On semantic words and semantic phrases or semantic sentences the semantic relatedness is dependent .In research of Zhao Zingling , Zhang Huiyan the lexical knowledge is based on human knowledge[4]. The correlation is measured using semantine comparisons on the effect of word order and sentence structure.Jiyi Li, Toshiyuki Shimizu proposed a document evaluation approach by combining different semantine similitude representation models on a collection of documents called Fused model[6].They investigated the candidate model and its fused version to demonstrate the relationship between models to boost STS models output. H.P. Costa used the Distribution Similarity Measures to evaluate the relationship between comparable entities by measuring the common entities [2]. The statistical methods like SCC and Chi-Square were analyzed using the comparable corpora where the author got high precision results which was additionally used to calculate documents ranks according to the estimated similarity[2]. S. Poomagal has used Cosine Similarity on the textual documents and used Cosine Similarity measure to calculate the Page Rank by calculating thennumber of in-links from the similar documents[9].

3 SYSTEM ARCHITECTURE

System design is one of the main aspects for practical implementation of research. This section presents an overview to System Architecture to find the related forum posts to the user based on input post to the system. The aim of the system is to show accurate results with the help of NLP models. System design is one of the main aspects for practical implementation of research. This section presents an overview to System Architecture to find the related forum posts to the user based on input post to the system. The aim of the system is to show accurate results with the help of NLP models.

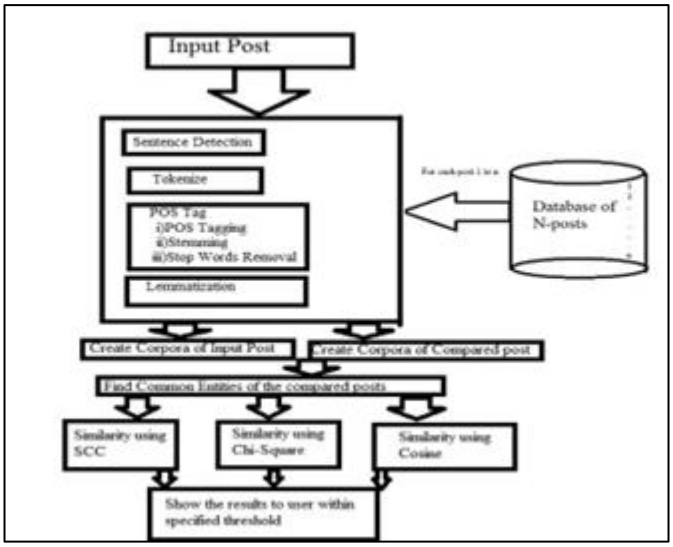


Figure 1

A.OpenNLP models:

Natural Language text is processed using OpenNLP library. It services are tokenizing, sentence detection, segmentation, stemming, lemmatization, Named-Entity Recognition, Summarization etc. This set comprise models of different real world languages. For implementation of NLP projects all services are not compulsory to implement

This set comprise models of different real world languages. For implementation of NLP projects all services are not compulsory to implement but are required according to the user requirement. In out work we have used the following. Tokenization is the process of creating tokens which are individual words in any natural language. In tokenization process text or set of text is broken into individual words which are used as input for analyzing and parsing which is to tag the syntactic relation between the words. POS Tagging is used for processing purpose sentence is converted in list of tuples and these words are tagged according to the Part of Speech in grammar which denotes whether a word is noun, verb, ad jective, pronoun etc. This tagging is necessary for building parse trees where extracting relations between words become useful. Stemming is the process of converting a word to its base or root form .for e.g beginning, begun to begin, in our work we have used Snowball stemmer to stem the sentences. StopWords removal is process of removing the most common words used in natural language to designate subject or for conjunction the two words like a, an and the or on, at etc. Lemmatization is the process of converting the word to its morphological form. For e.g humorous, ridiculous can be reduced to funny. Groups together different inflected forms of words called Lemma. It is similar to stemming like mapping of several words to one common root. Output of lemmatization is proper word. After preprocessing a corpus is created for the common entites. This corpus is created on the mean score of ncommonlemma, ncommonstems, ncommontokens. The common entities are the co-occurrence of the terms in the matrix created while preprocessing. Entropy is the amount of information a text segment have it is the number of bit multiplied by length of message. For entropy calculation frequencies are measured for words in corpus the more the words are different it becomes difficult to predict the content. In our work we have calculated Shannon Entropy Index its formula is given in figure 2.

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

Figure 2

Spearman's Rank Correlation Coefficient(SCC) is used to find the similarity between two documents of varying size. For calculating SCC common entities are calculated between the comparing documents say post1 with post2.In our work we have compared the first document with all the remaining documents in the database. These common entities are ranked in ascending order so that if highest frequency entity will be assigned higher rank and lower frequency entity will get lower ranks for both the posts. All the entities in the documents are ranked individually in their own corpus and then these corpuses are compared such that the difference is computed on ranking order for entities.

$$\mathrm{SCC}(\mathrm{post}_1, \mathrm{post}_2) = \frac{1-6\Sigma(\mathrm{di})^2}{n(n^2-1)}$$

Figure 3

where d_i is the difference computed between ranks of entities in 2 posts and n is the number of common entities. Chi-Square Test is for high performance and robustness Chi-square method is used, it denotes whether relationship between variables in sample also reflect relationship in real. Like SCC common entities are calculated for both the posts. Chi-Square test works on the set of Observed value and Expected Value. Lets say the size of doc_1 is N_1 and the size of doc_2 is N_2 and e_j is common entity then the Observed Value is calculated as $O(e_{j,} doc_{1})$, similarly the observed value of doc_2 will be $O(e_{j,} doc_{2})$.The Expected Values for doc_1 is given is figure 4 and similarly for doc_2 in figure 5.

$$e_j\,doc_1 = \frac{N_1 * (O(e_j, doc_1)) + (O(e_j, doc_2))}{N_{1+} N_2}$$

Figure 4

Chi-Square score is the distance between doc_1 and $doc_{2.}$ In Figure 6

$$e_j\,doc_2 = \frac{N_2 * (O(e_j, doc_1)) + (O(e_j, doc_2))}{N_{1+} N_2}$$

Figure 5

Vector Space Model with Cosine Similarity:

A document can be described flexible and powerful when there are multi-set terms which is also called Bag-of-words. Some words may appear multiple of times at different instances in a paragraph or sentence for e.g k- words may have i^{th} position we say k_i The position of words is not an important factor in this type of model. For e.g S={raj goes to school} and S={school goes to raj} will have the same value in this model. The grammatical arrangement of words may define the different meaning to sentence but in this model it will be treated as same case. In the given e.g Raj and School both the terms are called as Content words.Documents are described in term of vectors in this model. While representing the algebraic of Vector Space Model there are two steps. First is to develop the vector of words and then secondly is to transform that vector in the numerical format. In our work we applied NLP so we get the refined vector of words so the overhead required to remove the stopwords is reduced. In the second step, term document matrix is to be developed for each document d. Vector $d=<w_1,w_2,\ldots,w_n>$ and vector $q=<v_1,v_2,\ldots,v_n>$ where w in vector d represents words in document and vector q represents weight v for each corresponding word in the document. The weighting scheme is given in binary format where 1 represents presence of word in the post and 0 represents absence of word. The Term-Frequency(TF) is calculated by frequency of times word appeared in the document and IDF is calculated by dividing total number of documents by the number of documents containing the word.

$$\chi^2\ (\mathrm{doc}_1, \mathrm{doc}_2) = \Sigma\ (O-E)^2/E$$

Figure 6

In our work we have used Cosine Similarity in correspondence with SCC and Chi-Square verification for similarity. Advantage of Cosine Similarity is apart from size of each documents if one document contains all the words in the other document both documents will be still be shown similar for e.g Doc1 ->He runs to river and Doc2-> He runs to river to get water both will be treated as same documents. Cosine Similarity projects Cosine angle between two vectors in multidimensional space. When both documents are similar then the value will be closer to 1.In figure below A is document1 and B is document respectively.

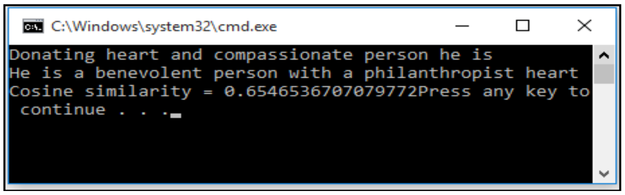
$$\mathrm{TF\text{-}IDF=TF*IDF}$$

Figure 7

For implementation of the system we have used Java jdk 12.0 with Apache Open NLP toolkit on Windows 10 OS. The System will require 4GB RAM with 10 GB of free storage and i3 2GHz processor. For storage and database purpose we have used XAMPP version 3.2.2.The dataset is scraped using WebHarvy data scraper. We have used dataset of Amazon product reviews of different accessories like Laptops, Mobiles, Tablets etc. For experimental results we have used some Tripadvisor reviews and Quora posts.

4.0 EXPERIMENTATION AND PERFORMANCE EVALUATION

For performance evaluation and accuracy we have performed certain experiments on other algorithms for small sets of text data where human brain can decide whether these sentences are similar or not also we have performed some experiments on text data with use of Natural Language Processing aided techniques with Cosine Similarity and without use of NaturalLanguage processing techniques. For experimental purpose we have used a large dataset of Amazon reviews of 1000 posts where performance on real time system is determined for storing, processing. The results given below are of same statements with Cosine Similarity without NLP. The further given sentences are same with change in sequence and used with other synonyms 1) "He is a benevolent person with a philanthropist heart" 2)" Donating heart and compassionate person he is" Figure 9 illustrates the result of Cosine Similarity without NLP.

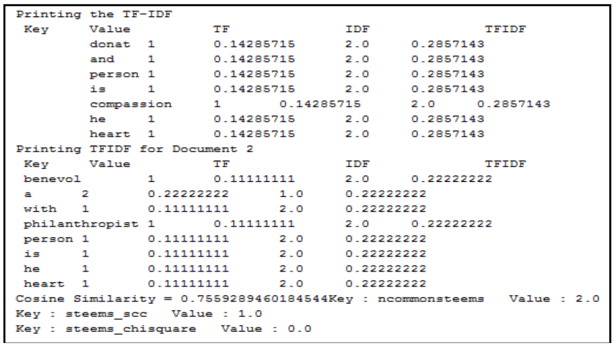


C:\Windows\system32\cmd.exe

```
Donating heart and compassionate person he is
He is a benevolent person with a philanthropist heart
Cosine similarity = 0.6546536707079772Press any key to
 continue . . .
```

Figure 8

Using our system we got the following results for small experimentation where result for SCC Similiarity measure score was 1.0 which is both the sentences are equal and for Chi-Square Similarity measure was 0.0 which is also the equal value when scaled for Chi-Square. Figure 10 illustrates the Cosine Similarity with NLP and results of our system with SCC and Chi-Square similarity measures.



```
Printing the TF-IDF
 Key      Value          TF              IDF             TFIDF
          donat  1       0.14285715      2.0      0.2857143
          and    1       0.14285715      2.0      0.2857143
          person 1       0.14285715      2.0      0.2857143
          is     1       0.14285715      2.0      0.2857143
          compassion     1       0.14285715      2.0      0.2857143
          he     1       0.14285715      2.0      0.2857143
          heart  1       0.14285715      2.0      0.2857143
Printing TFIDF for Document 2
 Key      Value          TF              IDF             TFIDF
 benevol         1       0.11111111      2.0      0.22222222
 a        2      0.22222222      1.0      0.22222222
 with     1      0.11111111      2.0      0.22222222
 philanthropist 1        0.11111111      2.0      0.22222222
 person 1        0.11111111      2.0      0.22222222
 is       1      0.11111111      2.0      0.22222222
 he       1      0.11111111      2.0      0.22222222
 heart  1        0.11111111      2.0      0.22222222
Cosine Similarity = 0.7559289460184544Key : ncommonsteems   Value : 2.0
Key : steems_scc   Value : 1.0
Key : steems_chisquare   Value : 0.0
```

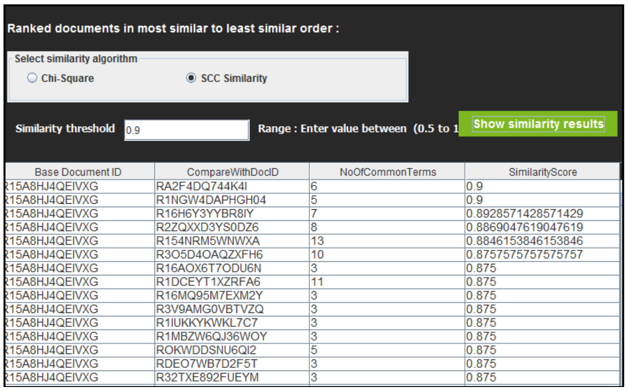
Figure 9

For test purpose we have used the 1000 reviews of variable sizes from Amazon dataset of Laptop reviews. The loading data from the JSON files and calculating its amount of information takes nearly 7 seconds with finding the segments for each text review on an average each post can be divided into 8 segments. For storing the data in database into MYSQL database takes about 80 seconds. The details of computation are in Table 1.

Table 1

No. of Reviews	Time required for Segmenting and Extracting	Time required for storing in MYSQL	Time taken SCC, Chi-Square and Cosine Similarity
1180	10 secs	60 seconds	180 seconds

For retrieving of the actual data from the database we have used the 2 methods Rating-Scale approach and Top-n approach both are query based. The results of the Rating scale approach can be found directly from the system with threshold as parameter in which all the posts Similar posts in the specified range of SCC form 0.5 to specified threshold within 1.0 will be displayed to the user of the system. Figure 11 illustrates the results of Rating-Scale approach using our System where user has to pass the SCC value and for Chi-Square we have decided the threshold in between 0.0 to 3.0 which will display the top similar posts to user from the database.



Ranked documents in most similar to least similar order :

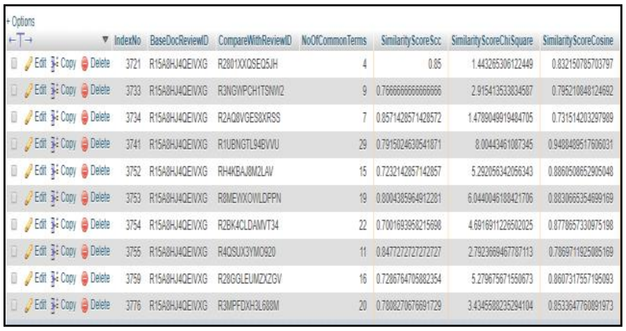
Select similarity algorithm
○ Chi-Square ● SCC Similarity

Similarity threshold 0.9 Range : Enter value between (0.5 to 1

Base Document ID	CompareWithDocID	NoOfCommonTerms	SimilarityScore
R15A8HJ4QEIVXG	RA2F4DQ744K4I	6	0.9
R15A8HJ4QEIVXG	R1NGW4DAPHGH04	5	0.9
R15A8HJ4QEIVXG	R16H6Y3YYBR8IY	7	0.8928571428571429
R15A8HJ4QEIVXG	R2ZQXXD3YS0DZ6	8	0.8869047619047619
R15A8HJ4QEIVXG	R154NRM5WNWXA	13	0.8846153846153846
R15A8HJ4QEIVXG	R3O5D4OAQZXFH6	10	0.8757575757575757
R15A8HJ4QEIVXG	R16AOX6T7ODU6N	3	0.875
R15A8HJ4QEIVXG	R1DCEYT1XZRFA6	11	0.875
R15A8HJ4QEIVXG	R16MQ95M7EXM2Y	3	0.875
R15A8HJ4QEIVXG	R3V9AMG0VBTVZQ	3	0.875
R15A8HJ4QEIVXG	R1IUKKYKWKL7C7	3	0.875
R15A8HJ4QEIVXG	R1MBZW6QJ36WOY	3	0.875
R15A8HJ4QEIVXG	ROKWDDSNU6QI2	5	0.875
R15A8HJ4QEIVXG	RDEO7WB7D2F5T	3	0.875
R15A8HJ4QEIVXG	R32TXE892FUEYM	3	0.875

Figure 10

The Top-n approach can be directly found from the database where the top matching posts in SCC will be displayed. Figure 12 illustrates the posts using the Top-n approach.



	IndexNo	BaseDocReviewID	CompareWithReviewID	NoOfCommonTerms	SimilarityScoreScc	SimilarityScoreChiSquare	SimilarityScoreCosine
□ Edit Copy Delete	3721	R15A8HJ4QEIVXG	R2801XXQSEQ5JH	4	0.85	1.443265306122449	0.832150785703797
□ Edit Copy Delete	3733	R15A8HJ4QEIVXG	R3NGWPCH1TSNW2	9	0.7666666666666666	2.915413533834587	0.795210848124692
□ Edit Copy Delete	3734	R15A8HJ4QEIVXG	R2AQ8VGES8XRSS	7	0.8571428571428572	1.4789049919484705	0.731514203297989
□ Edit Copy Delete	3741	R15A8HJ4QEIVXG	R1UBNGTL94BVVU	29	0.7915024630541871	8.00443461087345	0.9488489517606031
□ Edit Copy Delete	3752	R15A8HJ4QEIVXG	RH4KBAJ8M2LAV	15	0.7232142857142857	5.292056342056343	0.8860508652905048
□ Edit Copy Delete	3753	R15A8HJ4QEIVXG	R8MEWXOWLDFPN	19	0.8004385964912281	6.0440046188421706	0.8830665354699169
□ Edit Copy Delete	3754	R15A8HJ4QEIVXG	R2BK4CLDAMVT34	22	0.7001693958215698	4.6916911226502025	0.8778657330975198
□ Edit Copy Delete	3755	R15A8HJ4QEIVXG	R4QSUX3YMO920	11	0.8477272727272727	2.7923669467787113	0.7869711925085169
□ Edit Copy Delete	3759	R15A8HJ4QEIVXG	R28GGLEUMZXZGV	16	0.7286764705882354	5.279675671550673	0.8607317557195093
□ Edit Copy Delete	3776	R15A8HJ4QEIVXG	R3MPFDXH3L688M	20	0.7808270676691729	3.4345588235294104	0.8533647760891973

Figure 11

Use of XAMPP and MySQL in the project gives the feature of data summarization on the basis of Query where the analysis using the different charts is not a tedious task .Figure 13 illustrates it.

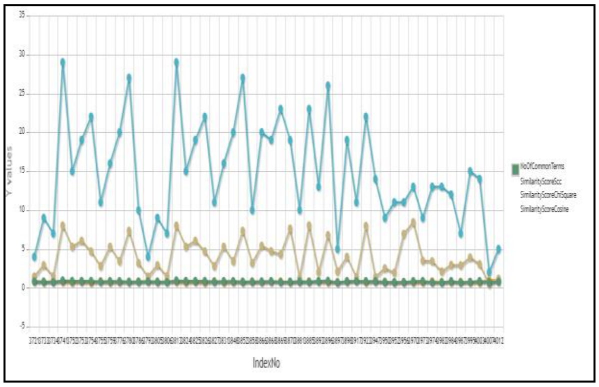


Figure 12

The results of the summarized data displays results of SCC ,Cosine Similarity, Chi-Square Similarity and Number of Common Terms where the SCC and Cosine Similarity will always be below 1 and the Chi-Square Similarity and Number of Common Terms will display the result above 1 because the numbers are always larger as compared to the SCC and Cosine Similarity measure.

4 CONCLUSION

By using the Natural-Language-Processing functionality to Similarity measures the efficiency of system has improvised. The synonym detection capability has detected the similar sentences and has shown better results than the previous semantic measures. With XAMPP MySQL the Summarization of data has become easy for plotting the graphs without external coding. Due to easy database connectivity and processing large database with the use of Java framework there was boost for quick results from database. There has been a challenge in scaling the similarity scores between SCC and Chi-Square because of difference in variation of resulting parameters like SCC result will be between 0-1 and Chi-Square results will always be above 0.Though most of the results of same scaling parameters like SCC and Cosine similarity are 0-1 in most cases if SCC has shown equal to better scores than Cosine Similarity and thus Chi-Square has also shown equal results to SCC when both are mapped. In future with appropriate recommendation algorithms we will build the recommendation system on this given model.

5 REFERENCES

[1]. D. Papadimitriou, G. Koutrika,Y.Velegrakis and J. Mylopoulos, “Finding Related Forum Posts through Content Similarity over Intention-Based Segmentation” in IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 9, pp. 1860-1873, Sep 2017.

[2]. Hernani Costaa, Gloria Corpas Pastora and Ruslan Mitkovb, "Measuring the Relatedness between Documents in Comparable Corpora" in Conference: 11th Int. Conf. on Terminology and Artificial Intelligence (TIA'15). Granada, Spain. pp.29-37., At Granada, Spain Nov 2015.

[3]. Ming Liua,1, Bo Lang a and Zepeng Gu, "Calculating Semantic Similarity between Academic Articles using Topic Event and Ontology" arxiv 2017.

[4]. Zhao Zingling ,Zhang Huiyan,Cui Baojiang ,"Sentence Similarity Based on Semantic Vector Model",2014 Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing.

[5]. Wael H. Gomaa, Aly H. Fahmy "A Survey of Text Similarity Approaches" International Journal of Computer Applications (0975 – 8887)Volume 68– No.13, April 2013.

[6]. Jiyi Li, Toshiyuki Shimizu, Masatoshi Yoshikawa "Document Similarity Intention based Segmentation Approach".

[7]. Muthuselvi M,Annie John,Anslin Jenisha S,Archana S,Manimegalai M,Student, Department of CSE, Univer"Text Mining from CMS Forums- an Intention based Segmentation Approach"Journal of Network Communications and Emerging Technologies (JNCET)Volume 8, Issue 4, April (2018).

[8]. V. Sowmya Vishnu Vardhan B, Bhadri Raju M S V S,"Influence of Token Similarity Measures for Semantic Textual Similarity", 2016 IEEE 6th International Conference on Advanced Computing.

[9]. .S. Poomagal,T. Hamsapriya “Cosine Similarity based Page Rank Calculation” Int. J. WebScience Vol. 1. Nos. 1/2, 2011.

[10]. www.youtube.com/edureka/Introduction to Natural Language Processing.

[11]. Yue Feng , Ebrahi Bagheri , Faezah Ensan2 and Jelena Jovanovic “The state of the art in Semantic Relatedness: A framework for comparision” The knowledge engineering review 2017.