IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

Enhancing Automatic PPT Generation Technique through NLP for Textual Data

Pooja Belote[1], Sonali Bidwai[2], Snehal Jadhav[3], Pradnya Kapadnis[4], Nakul Sharma[5]

Department of Information Technology, Sinhgad Academy of Engineering, Pune, India[1,2,3,4,5]

Abstract: Proposed system come with an idea of automatically generate presentation slides from textual data. The slides mostly incorporate of only the important points related to the topic. There are various tools and systems are available in the market which only address with formatting of the slides but not the content. But it having performance issues regarding of sentence extraction, so this paper proposes an idea of automatically ppt generation from text. This will finally help in reducing a great amount of the presenter's time and efforts. The intended system works on NLP rules to classify data for the desired slides.

Keywords: Pre-processing, Tokenization, Steaming, Feature Extraction, Fuzzy Classification.

I. INTRODUCTION

This Software Requirements Specification furnishes a complete description of all the functions and restraints of the **"**Automatic PPT slide generation system for the given document using NLP features and Fuzzy logic**"**. The document describes the issues regarding to the system and what actions are to be performed by the development team in order to arise with a better solution. The basic idea of Automatic PPT generation system comes from the fact that due to high increase of the digital data, in depth study of the same always takes more time than of estimation. So it is a hard task to gather proper points to generate PPT slides .so proposed system put forwards an idea of Automatic PPT generation system using feature extraction by applying strong NLP protocols and then these features are classified by using fuzzy logic to get the best PPT slide points out of the given document.

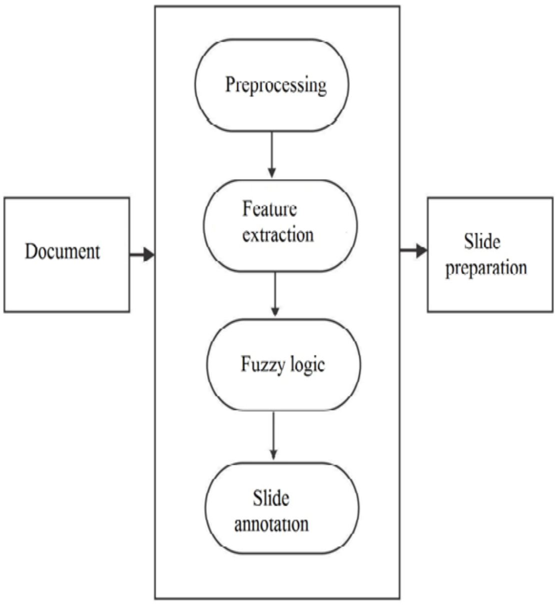


Fig 1 . Data Flow Diagram

Pre-processing:
Pre-processing reduces overhead and increases processing speed. It consist of removal of stop words, stemming and tokenization.

Feature Extraction:
In this stage important features are extracted by the following means.

It includes numerical data, term weight, proper noun and sentence to sentence similarity.

Fuzzy logic:
Classification is done into 5 types: Very low-0 Low Medium High Very high-1 Generated scores of the sentences are checked according to the above classification. A score is termed as Very low if it has a score 0 and is termed as Very high if it has a score 1. Hence if the score is 0, the sentence is less important and if the score is 1 then the sentence is important. Thus, importance of a sentence can be obtained.

This paper dedicates section II for related work and section III for narration of implementing idea. Section IV evaluates the performance of our system and finally section V concludes this paper with feature extension possibilities.

II. LITERATURE SURVEY

1. Automatic Slide Generation Based on Discourse structure analysis:
In this slides, there are itemizing topic/non topic parts that are taken from the text based on syntactic/case analysis. The items are identifying by controlled according to the discourse structure and this is detected by cue phrases, identication of word chain and similarity between two sentences.

DOI10.17148/IJARCCE.2017.63237

In Figure 2.1,An example of a text is shown and in Figure 2.2 an example slide that is generated from the text is shown (the translated slide is shown in Figure 2.3).

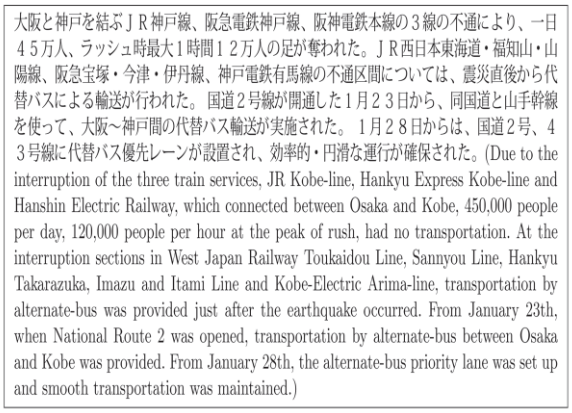
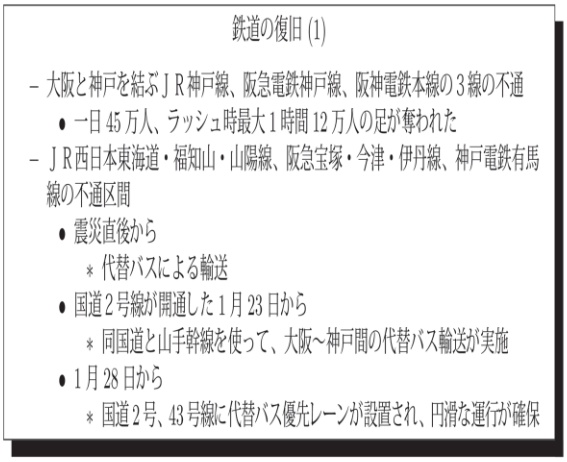

大阪と神戸を結ぶＪＲ神戸線、阪急電鉄神戸線、阪神電鉄本線の３線の不通により、一日４５万人、ラッシュ時最大１時間１２万人の足が奪われた。ＪＲ西日本東海道・福知山・山陽線、阪急宝塚・今津・伊丹線、神戸電鉄有馬線の不通区間については、震災直後から代替バスによる輸送が行われた。国道２号線が開通した１月２３日から、同国道と山手幹線を使って、大阪～神戸間の代替バス輸送が実施された。１月２８日からは、国道２号、４３号線に代替バス優先レーンが設置され、効率的・円滑な運行が確保された。(Due to the interruption of the three train services, JR Kobe-line, Hankyu Express Kobe-line and Hanshin Electric Railway, which connected between Osaka and Kobe, 450,000 people per day, 120,000 people per hour at the peak of rush, had no transportation. At the interruption sections in West Japan Railway Toukaidou Line, Sannyou Line, Hankyu Takarazuka, Imazu and Itami Line and Kobe-Electric Arima-line, transportation by alternate-bus was provided just after the earthquake occurred. From January 23th, when National Route 2 was opened, transportation by alternate-bus between Osaka and Kobe was provided. From January 28th, the alternate-bus priority lane was set up and smooth transportation was maintained.)

Fig 2.1An example of a text


鉄道の復旧 (1)

– 大阪と神戸を結ぶＪＲ神戸線、阪急電鉄神戸線、阪神電鉄本線の３線の不通
  • 一日 45 万人、ラッシュ時最大１時間 12 万人の足が奪われた
– ＪＲ西日本東海道・福知山・山陽線、阪急宝塚・今津・伊丹線、神戸電鉄有馬線の不通区間
  • 震災直後から
    ∗ 代替バスによる輸送
  • 国道２号線が開通した１月 23 日から
    ∗ 同国道と山手幹線を使って、大阪～神戸間の代替バス輸送が実施
  • １月 28 日から
    ∗ 国道２号、43 号線に代替バス優先レーンが設置され、円滑な運行が確保

Fig2.2 An example of a slide

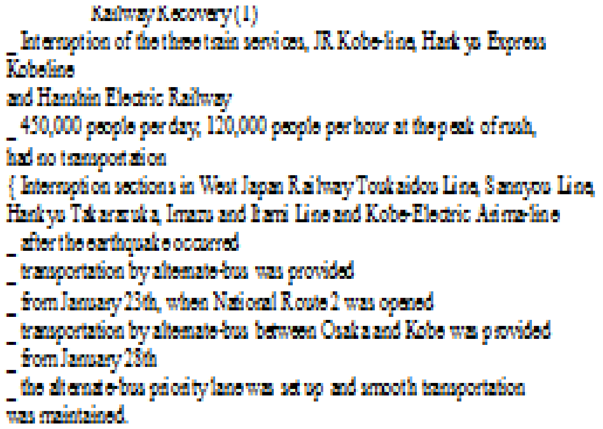

Railway Recovery (1)
_ Interruption of the three train services, JR Kobe-line, Hankyu Express Kobeline
and Hanshin Electric Railway
_ 450,000 people per day, 120,000 people per hour at the peak of rush, had no transportation
{ Interruption sections in West Japan Railway Toukaidou Line, Sannyou Line, Hankyu Takarazuka, Imazu and Itami Line and Kobe-Electric Arima-line
_ after the earthquake occurred
_ transportation by alternate-bus was provided
_ from January 23th, when National Route 2 was opened
_ transportation by alternate-bus between Osaka and Kobe was provided
_ from January 28th
_ the alternate-bus priority lane was set up and smooth transportation was maintained.

Fig 2.3An example of a slide (in English)

2. A Review on Feature Selection and Feature Extraction for Text Classification:

The problem of high dimensionality of feature space is text classification. This problem is solved by feature selection and feature extraction methods and improves the performance of text categorization.

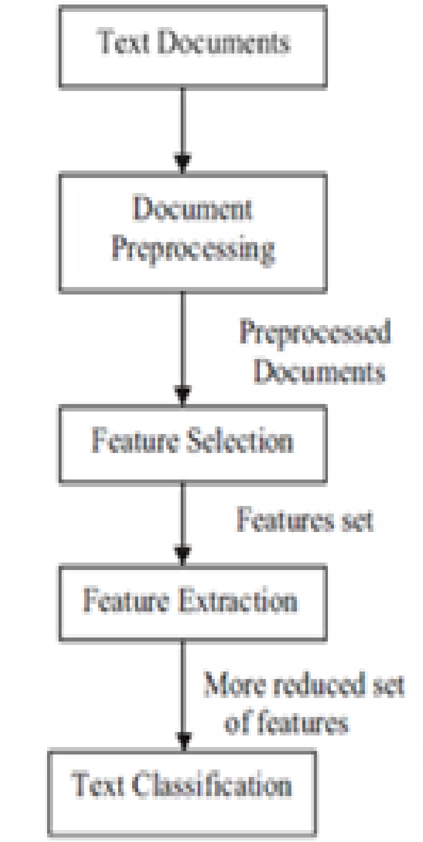


Fig.3Text classification flow

The irrelevant features from the text documents are removed feature selection and feature extraction techniques and decrease the dimensionality of feature space.

In Fig3 flow of text classification includes four steps: document preprocessing, feature selection, feature extraction and text classification.

3. Feature-Based Sentence Extraction Using Fuzzy Inference rules:

This paper shows the automatic text summarization by sentence extraction. Extraction is the identification of important features done by summarization. For extract the sentences used important features based on fuzzy logic. In this experiment, they used 30 test documents in DUC2002data set. Each document is made by preprocessing process. They calculate their score for each sentence using 8important features. They introduce a method using fuzzy logic for sentence extraction.

4.Slides Gen: Automatic Generation of Presentation Slides for a Technical Paper Using Summarization:

In this paper, documents is LATEX form which is rich in structural and semantic information. They used them as input to their system. These documents are converted to XML format. Then, this XML file is parsed and information is extracted. To generate slides used query specific extractive Summarizer.

DOI10.17148/IJARCCE.2017.63237

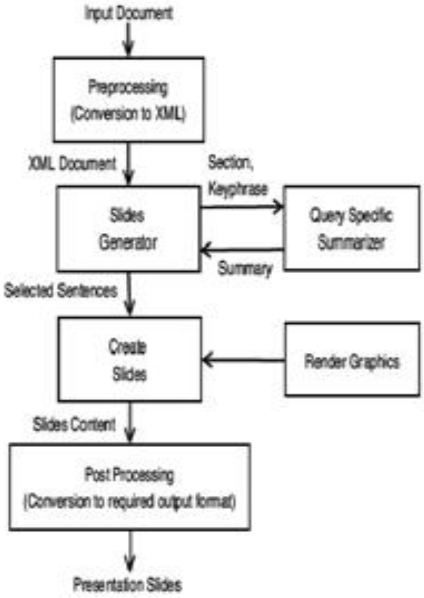


Fig 4.System Architecture

5. PPSGen: Learning-Based Presentation Slides Generation for Academic Papers:

In this paper regression method used to learn the importance scores of the sentences in an academic paper and to generate well-structured slides by selecting and aligning key phrases and sentences using integer linear programming (ILP). Results of test set of 200 pairs of papers and slides collected on the web demonstrate. This is their proposed PPSGen system can generate slides with better quality. In stemming handle a word is conveyed to its base frame. By doing this overhead is lessened and exactness is expanded.

e.g.: Engineering will be lessened to engine Tokenization. It this procedure words are trimmed, spaces are expelled,
tokens are created and are then placed in an cluster.

For survey work by S. L. Bangare et al was also studied in relation to study few image processing techniques [6] [7] [8] [].

III.PROPOSED METHODOLOGY

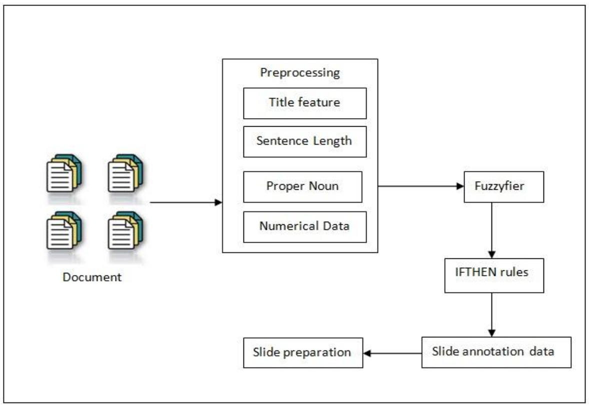


Fig 3: Proposed System Workflow

It has dependably been a think about how introduction slides are produced consequently from content. Many existing frameworks are yielding low semantics. Proposed framework takes a shot at NLP principles to characterize the information for coveted slides. At the point when content is given as contribution to the framework it experiences the accompanying stages before creating introduction slides.

Proposed System implements phase based approach for automated slide generation **Phase 1: Input to system:** PDF, DOC file is been given as input to system and complete data is been retrieved in string format.

Phase 2: Pre-processing: Input String data is been sent to preprocess phase for data cleansing and Data filtration process this process consist of three sub process. Data cleansing stop word and special symbol removal. Data alteration i.e stemming brings all terms to base form.

Stop words are those words, which when expelled won't modify the coveted importance of the sentence. Consequently stop words are evacuated with a specific end goal to expand the handling speed. Copy words are recognized and expelled from the sentence

Phase 3: Feature Data Extraction: feature data is vital data

Numerical information The sentence that holds numerical information is critical and it is most presumably incorporated into the archive synopsis. The numerical score of each sentence is figured. This score is acquired by ascertaining the number of numbers happening in a sentence. Contingent upon the score it is chosen whether to incorporate the sentence or no .**Term weight** It is the recurrence of the term rates inside a record which has been utilized for computing the rank of the sentence. The score of a sentence can be proposed as the whole of the score of words in the sentence. Term weight is the quantity of times a specific word has happened in a sentence. tf_isf(Term frequency, Inverse sentence recurrence) strategy is connected to ascertain the score of the term. If the score is high then that term is thought to be essential. Formal person, place or thing the sentence that holds greatest number of formal people, places or things (name substance) is a fundamental sentence and is well on the way to be incorporated into the report synopsis. The score for this component is the quantity of formal people, places or things happening in a sentence over the length of the sentence.

Sentence to **Sentence similarity** This feature finds the similarity between sentences. For each sentence S, the similarity between S and other sentence sis computed by the cosine similarity measure with a value resulting between 0 and 1.

Fuzzy Classification: classification theory that divides data into five parts.

- **Very low-0**
- **Low**
- **Medium**
- **High**
- **Very high-1**

DOI10.17148/IJARCCE.2017.63237

Produced scores of the sentences are checked by the above grouping. A score is named as Very low in the event that it has a score 0 and is named as Very high on the off chance that it has a score 1. Consequently if the score is 0, the sentence is less imperative and if the score is 1 then the sentence is critical. Accordingly significance of a sentence can be acquired. **Fuzzy inference engine** It is utilized to remove adjust conclusions from estimated information. Tenets are composed to recognize titles for each slide.

Fuzzy If Then or Fuzzy restrictive proclamations are articulations of the shape "If A Then B", where An and B are names of fluffy sets portrayed by proper enrollment capacities. The created scores are checked with the if - then condition. If the score is high then the sentence is vital. In the event that the score is low then the sentence is not critical.

Algorithm to Pre-processing

```
Step 0: Start
Step 1: Get contents of Query
Step 2: split in Words
Step 3: Remove Special Symbols
Step 4: Identify Stopwords
Step 5: Remove Stopwords
Step 6: Identify Stemming Substring
Step 7: Replace Substring to desire String
Step 8: Concatenate Strings
Step 9: Preprocessed String
Step 10: Stop
```

Algorithm to find top words

```
Step 0: Start
Step 1: Read string
Step 2: divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: for i=0 to N (Where N is length of V)
Step 5: for i word of N check for its frequency
Step 6: Add frequency in List Called L
Step 7: end of for
Step 8: return L
Step 9: stop
```

Algorithm to find noun

```
Step 0: Start
Step 1: Read string
Step 2: divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: for i=0 to N (Where N is length of V)
Step 5: for i word of N check for its occurrence in Dictionary
Step 6: if present then return true
Step 7: else return false
Step 8: stop
```

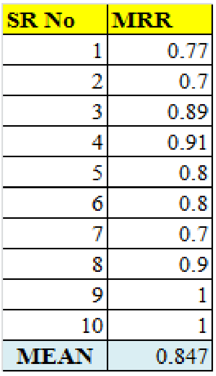
The proposed model is created on java based windows machines which utilizes Netbeans as IDE. For the analyses and assessment proposed framework utilizes distinctive elements to make PPT for the given content by utilizing improved characteristic dialect preparing (NLP) which is been controlled with fluffy rationale. Created framework is put under sledge in numerous situations to demonstrate its validness and precision in producing PPT's in beneath tests.

We tested our proposed framework more than 10 times for various information content records. For estimation of the exactness we utilized MRR as the best measuring strategy, where from the outcomes we separated the best one and relegates a rank as indicated by its execution out of 5 that will be in the end a conclusion. The positions are allocated as 1,1/2, 1,3,1/4, 1/5 and 0 as per the conclusions and after that they are naming as Reciprocal Rank (RR), The mean complementary rank (MRR) can be given as the mean rank over number of Runs.

$$MRR = \frac{\sum_{i=1}^{N} 1/Rank_i}{N} \qquad \ldots\ldots\ldots\ldots..(1)$$

We performed try different things with various info content documents for producing the PPT and rate the supposition allotting positions out of 5 for 10 number of runs, Which gives MRR as recorded in the beneath table.

TABLE I Result of the Experiment



SR No	MRR
1	0.77
2	0.7
3	0.89
4	0.91
5	0.8
6	0.8
7	0.7
8	0.9
9	1
10	1
MEAN	0.847

The above figure 2 gives the normal MRR of 0.847 for various number of keeps running for the information content documents. This outcome can be say great in our first endeavour of PPT Generation utilizing NLP rules.

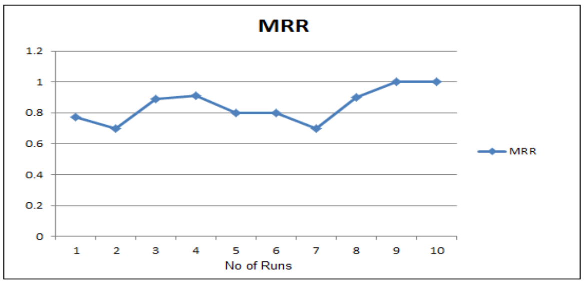


Fig 2. MRR for the Different Runs

IV. CONCLUSION AND FUTURE SCOPE

The Very much organized slides are created. Moderator's chance and endeavors are spared as it were. Slides will incorporate imperative key expressions and sentences identified with them. It is a tedious job to create presentation slides.

Thus our system will save a huge amount of the user's time and efforts. Presentation slides are generated in an efficient and quicker way after using the above methods. The system can be enhanced to work on all cross platforms.whereas the Microsoft Word templates are self-contained.

REFERENCES

[1] Yue Hu and XiaojunWan, "PPSGen: Learning-Based Presentation Slides Generation for Academic Papers", IEEE Transactions on knowledge and data engineering, vol. 27, no. 4, april 2015.
[2] Foram P. Shah and VibhaPatel ,"A Review on Feature Selection and FeatureExtraction for Text Classification", IEEE WiSPNET 2016 conference.
[3] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan "Feature-Based Sentence Extraction Using Fuzzy Inference rules," 2009 International Conference on Signal Processing Systems.
[4] Tomohide Shibata and SadaoKurohashi," Automatic Slide Generation Based on DiscourseStructure Analysis" R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 754–766, Springer-Verlag Berlin Heidelberg 2005.
[5] M. Sravanthi, C. RavindranathChowdary and P. Sreenivasa Kumar, "SlidesGen: Automatic Generation of Presentation Slides for a Technical Paper Using Summarization," Proceedings of the Twenty-Second International FLAIRS Conference (2009).
[6] Sunil L. Bangare et al. "Implementing Tumor Detection and Area Calculation in MRI Image of Human Brain Using Image Processing Techniques."International Journal of engineering Research and Applications 1.5: 60-65.
[7] Sunil L. Bangare, et al. "Reviewing Otsu's Method for Image Thresholding."International Journal of Applied Engineering Research 10.9 (2015): 21777-21783.
[8] Pallavi S. Bangare, et al. "Implementation of abandoned object detection in real time environment." International Journal of Computer Applications Vol.57, Issue.12 (2012).
[9] Sunil L. Bangare, et al. "Quality measurement of modularized object oriented software using metrics." Proceedings of the International Conference & Workshop on Emerging Trends in Technology. ACM, 2011.
[10] Magar, Anand Mohanrao, and Nilesh J. Uke. "Use of AST for Translating Executable UML Models to Java Code in Eclipse and Testing Strategy for UML Models." International Journal of Innovative Technology and Exploring Engineering (IJITEE) 3.7 (2013): 157-160.
[11] Yalla, Prasanth, and Nakul Sharma. "Integrating Natural Language Processing and Software Engineering." International Journal of Software Engineering and Its Applications 9.11 (2015): 127-136.
[12] Yalla, Prasanth, and Nakul Sharma. "Utilizing NL Text for Generating UML Diagrams." Proceedings of the International Congress on Information and Communication Technology. Springer Singapore, 2016.
[13] Mane, Tushar, and Laxman Deokate. "THE REAL SOCIAL NETWORKING WEBSITE.", International Journal of Advanced Technology & Engineering Research (IJATER), 3.1 (2013), pp.36-42
[14] Lathkar, Shilpa G., Nilima A. Kavitke, and Abhay N. Adapanawar. "Online Digital Advertising on Public Display." International Journal of Computer Applications 66.10 (2013).
[15] KOTWAL, PRIYA A., et al. "A Location Tracer With Social Networking Services." International Journal of Engineering and Technology (IJET) 4.1 (2012).
[16] Magar, Anand Mohanrao, and Nilesh J. Uke. "Use of AST for Translating Executable UML Models to Java Code in Eclipse and Testing Strategy for UML Models." International Journal of Innovative Technology and Exploring Engineering (IJITEE) 3.7 (2013): 157-160.
[17] Rathod, Aakash, Nitika Sinha, and Mrs Pankaja Alappanavar. "Extraction of Agricultural Elements using Unsupervised Learning." Imperial Journal of Interdisciplinary Research 2.6 (2016).
[18] Jain, Jyoti, et al. "Sentiment Analysis Using Supervised Machine Learning."Imperial Journal of Interdisciplinary Research 2.6 (2016), ISSN: 2454-1362.
[19] Nihar Suryawanshi, et al. "Sentiment Analysis Using Machine Learning."Imperial Journal of Interdisciplinary Research 6.1 (2016), pp.451-453.
[20] Kumar, Samarth, et al. "Stock Market Forecasting using Hybrid Methodology." Imperial Journal of Interdisciplinary Research 2.6 (2016).
[21] A. Sutar and P. Y. Pawar, "Load Balancing With Third Party Auditor", International Engineering Research Journal (IERJ), 2.3 (2016), pp.1156-1160.
[22] A. Mishra, et al., "DDOS Attack Detection and Multivariate Correlation Analysis and Clustering", IJSTE, 2.12 (2016).
[23] M. K. Nivangune et al, "Secured Online Voting System over the Network", International Engineering Research Journal (IERJ), 2.3 (2016), pp.1380-1383.
[24] Sunil L. Bangare, et al. "Automated API Testing Approach." International Journal of Engineering Science and Technology 4.2 (2012).
[25] Sunil L. Bangare, et al. "Automated Testing in Development Phase."International Journal of Engineering Science and Technology 4.2 (2012).

DOI10.17148/IJARCCE.2017.63237