International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 4, Issue 6 (Nov-Dec 2016), PP.199-203

AUTOMATIC SLIDE GENERATION FOR ACADEMIC PAPER USING PPSGEN METHOD

[1] Parvez Javed Shaikh [2] Prof. R A Deshmukh
[1, 2] Department of Computer Engineering, Rajarshi Shahu College of Engineering, Pune, India.
[1] pshaikh10@gmail.com, [2] radesh19@gmail.com

***Abstract*— To present a slide for papers and save the author's time when organizing presentations we proposed a method for automatic slide generation. Generally to generate the slides from start, it takes a lot of time. These slides contains data about base paper objective from conceptual details of system i.e. introduction and additionally approach utilized, writing audit from related work segment, exploratory results and conclusions from base paper. The generated slides can be used as rough idea for further preparation. This helps presenters in preparing their formal slides in faster manner. For this, to learn the importance of the sentences we have employed regression method in an academic paper, and then exploit the integer linear programming (ILP) method. It generates well-structured slides by selecting and aligning key phrases and sentences. Slides contents both graphical and textual data and we have focused on both text and graphical data like tables and images.**

***Keywords*— GDA, PPSGEN, ILP, SVR, NLP.**

I. INTRODUCTION

Presentation slides are important for effective way of presenting and exchange information, mostly in academic conferences. To present work analysts constantly favored utilization of slides for their work in meetings. There are many software in the market, for example, Microsoft Power-Point and Open Office to help analysts organize their slides. By utilizing these devices authors can simply design their slides, not the content. Regardless it requires moderators much investment to compose the slides from the scratch. In this paper, we propose a technique to consequently making presentation slides from scholarly papers. In this paper we arranged a method which serves to consequently make very much organized slides and gives such draft slides as a premise to diminish the moderator's opportunity and exertion while setting up their last presentation. Scholastic papers dependably have a parallel structure. They for the most part have a few segments like abstract, introduction, related work, proposed technique, analyses and conclusions.

Distinctive moderators may utilize different styles for presentation slides. Be that as it may, a moderator who are learner, dependably endeavor to adjust slides according to the paper sections sequence. Every area is related to at least one slides and one slide regularly has a title and diverse sentences. These sentences might be coordinated in some visual cues. This system utilizes machine learning technique to outline and make content for slides of the common sort said above and people groups to arrange their final slides. Automatic slides creation for scholastic papers is an extremely difficult task. Current techniques normally remove objects like sentences from the paper to build the slides. Rather than the short rundown removed by an outline framework, the slides are important to be a great deal more organized and any longer. Slides can be partitioned into a requested succession of parts. Every part addresses a particular theme and these points are additionally important to each other. Similarly errand of programmed slide creation from base paper is more troublesome than outline. Slides as a rule have content components as well as chart components, for example, figures and tables. So we will consider graphical elements to improve slides.

II. RELATED WORK

In this [2] paper author proposes a technique to naturally produce slides from data records delineate with the GDA label set. Where GDA labeling is utilize to encode semantic structure of presentation. The semantic relations removed for substance to speak to between sentences that discovered syntactic relations, state portray by sentence verb and linguistic relations. Initial step is to identify themes from the data reports, and after that by utilizing natural language processing vital sentences are extracted relevant to the points to create slides.

In this paper [3] author acquainted a system to produce slides from specialized papers. The information considered for this framework is scholarly papers in LATEX design. Proposed strategy ascertains the weights of vital sentences or terms in the paper utilizing TF*IDF measures. Using the term weights sentences and tables are also weighted and used to determine the number of objects for each section to generate the slides.

In this paper [5] author proposed a technique to consequently produce slides from crude data writings. Clauses and sentences are considered as phonetic unit and connectedness between the section like list, contrast, topic-chaining and cause are identified. Out of which some of statements will be recognized as point parts and remaining will be non-theme. These contents are utilized to create the last slides taking into account the identified semantic challenge and some heuristic tenets.

In these [4] papers, creator investigate the issue of changing particular papers and presentation slides. Assortment of the Hidden Markov Model (HMM) is used to change the substance in the slides to a fragment in the paper; they have similarly used the additional information of titles and position openings. Kan connected a balanced most prominent closeness system to do the relating to game plans and arranged a classifier for recognizing slides which are not to be balanced. Beamer and Girju took a gander at and evaluated four differing plan systems that were merged with methodologies, for instance, TF-IDF term weighting and question augmentation. Moreover,

SVR and ILP have been used extensively as a part of the errand of delineating.

There are some limitations and challenges in previous methods, e.g. Extraction of image and tables from the base paper. To overcome those we proposed a method which selects number of important sentences, images, tables and the phrases from the corresponding base paper. Calculation of the importance of sentence with image reference is challenging task.

III. PROPOSED SYSTEM

As shown in figure [1], various academic papers in various formats are taken as an input to the system which is later converted in to XML format. These XML documents are preprocessed using XML document parser so as to tag various important aspect of the document. These important aspects are processed and features are extracted from the document so as to generate a text file. Various features are matched with the generated text file so as to cover various important points from the document. The attributes are separated from the sentences parse tree. It incorporates the number of noun and verb states, the number of sub-sentences and the profundity of the parse tree.

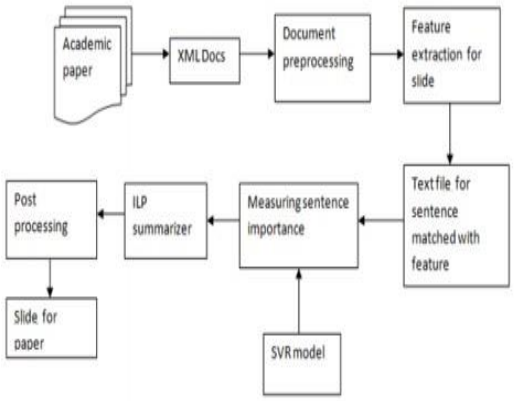


Fig. 1. System Architecture

We utilize the Stanford NLP library for sentence parsing. For making slides it's critical to spot significance of each sentence from the made archive. Support machine regression (SVR) model is utilized to calculate the importance of each sentence in the document. Sentence importance score is calculated and matrix is generated for each sentence. SVR is utilized to prepare and take in the sentence imperative score. We require foreseeing the significance score of each sentence to shape slide for academic paper presentation. SVR model is better than classification model because regression score is much finer to use for important sentence selection strategy for generating slide than a classification which gives coarse score.

By concerning generated matrix sentence significance score is calculable that later provided as an input to the integer linear programming (ILP) summarizer. ILP summarizer module plays terribly very important role in summarizing these sentences therefore on detect only very important topics connected with every sentence.

In our system, Support Vector Regression (SVR) model is used to learn the importance of each sentence in a paper, and then Integer Linear Programming (ILP) model is used to select and align key phrases and sentences for generated slides. Here we are using two algorithms for processing input data to generate slide are SVR model and ILP method.

A. SVR

1) Sentence Position: Here position of sentence is calculated using equation

$$SP(s) = position(s, doc(s)) / \left|d(s)\right| \quad (1)$$

Where position(s, doc(s)) is the sentence order of sentence s in its document doc(s), and $\left|d(s)\right|$ is the number of sentences in doc(s).

2) Content Similarity:

$$\boldsymbol{cosine_1(a,b)} = \frac{\sum_{i,j}^{N} s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^{N} s_{ij} a_i a_j}\sqrt{\sum_{i,j}^{N} s_{ij} b_i b_j}} \quad (2)$$

Where S_{ij} = similarity ($feature_i$, $feature_j$).

If there is no similarity between features (Sii = 1, Sij = 0 for $i \neq j$), the given equation is equivalent to the conventional cosine similarity formula.

3) Word Overlap*:* It is number of words shared by the input query sentence q with sentence s consider finding matching. This is accomplished by removing stop words and redundant words from both q and s.

B. ILP

Important points identified by ILP are now needed to be highlight and convert in to slides format. Post processing of the document is done based on the required slide format for all the important topics and slides are generated. It constructs summaries by minimizing their pair wise similarity and maximizing the importance of the selected sentences, as shown below which of the form.

$$Max\ x,y \sum_{i=1}^{n} imp(si).xi - \sum_{i=1}^{n} \sum_{j=i+1}^{n} sim(si,sj).yi,j \quad (3)$$

$$\sum_{i=1}^{n} l_i .x_i \leq L_{max} \quad (4)$$

For (i = 1, …. n & j = i + 1, … n)

Where

$$y_{i,j} - x_i \leq 0 \quad (5)$$

$$y_{i,j} - x_j \leq 0 \quad (6)$$

$$y_i + x_i - y_{i,j} \leq 1 \quad (7)$$

n - Number of sentences in the input documents

$imp(S_i)$ - Importance score of sentence S_i

l_i - Length of S_i

$sim(S_i, S_j)$ - Similarity of sentences S_i and S_j

L_{max} - Maximum allowed length

The Xi variables are represented in binary. This variable indicates whether the corresponding sentences Si is part of the summary. The yi,j variables, in binary form gives idea whether both Si and Sj are part of the summary.

Condition 3 gives assurance that total maximum length of sentence is not exceeded.

Condition 4–6 assured that the x_i , x_j , and $y_{i,j}$ weight are consistent (e.g., if $y_{i,j} = 1$, then $x_i = x_j = 1$; and if $y_{i,j} = 0$, then $x_i = 0$ or $x_j = 0$).

	ROUGE-1	ROUGE-2	ROUGE-SU
Max-SVR	0.41212	0.13175	0.17613

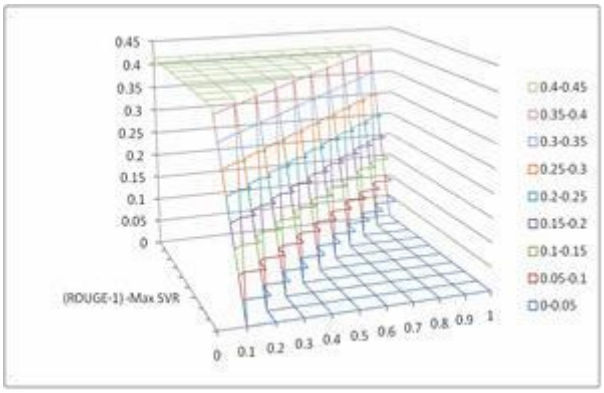


Fig. 2. Max_SVR-ROUGE-1

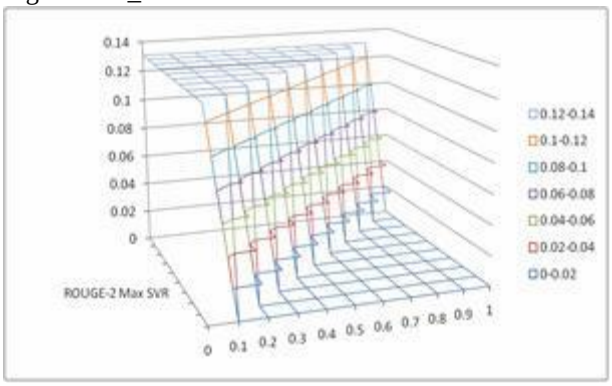


Fig. 3. Max_SVR-ROUGE-2

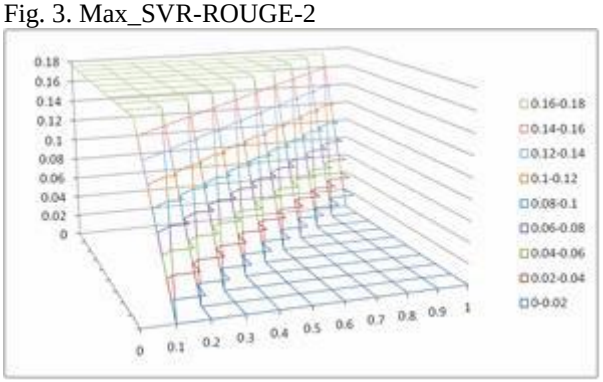


Fig. 4. Max_SVR-ROUGE-SU4

Table beneath demonstrates the normal scores evaluated by human judges for every strategy. The slides created by our technique clearly have preferred general quality over the baseline strategies.

TABLE II. OVERALL COMPARISON

Judge	Structure	Content	Overall
1	3.8	3.6	3.7
2	3.75	3.6	3.675
3	3.75	3.5	3.625
4	3.4	3.3	3.35
Avg	3.675	3.5	3.587

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar.

CONCLUSIONS

In this paper we have presented a framework to create presentation slides from scholastic paper. The most essential steps in this examination are report grouping and key expression extraction. We additionally depict essential sentence extraction and separating. Similarity algorithm cosine similarity is utilized to pick one sentence from every set of similar sentences and disregarding the rest. The significant commitment which can be pushed by the present study is another model for extracting pictures and tables which will enhance presentation slides. The aftereffects of investigations demonstrated the proposed model gives better performance in contrast with other works.

FUTURE WORK

Our system generates slides based on one given paper. Further knowledge corresponding to other related papers and the citation understanding can be used to beef up the generated slides. We will be able to consider this hindrance in the long run.

ACKNOWLEDGMENT

We would like to thank the researchers and publishers for resources. We additionally thank the college authority for giving the required foundation and backing. At long last, we want to extend an ardent appreciation to loved ones individuals.

SAMPLE SLIDES

Agenda

- INTRODUCTION
- ALGORITHM
- DATASET
- SENTIMENT ANALYSIS
- MODEL LEARNING AND PREDICTION
- PORTFOLIO MANAGEMENT
- CONCLUSIONS AND FUTURE WORK

Abstract

- In this paper, we apply sentiment analysis and machine learning principles to find the correlation between "public sentiment" and "market sentiment".
- We use twitter data to predict public mood and use the predicted mood and previous days' DJIA values to predict the stock market movements.
- In order to test our results, we propose a new cross validation method for financial data and obtain 75.56% accuracy using Self Organizing Fuzzy Neural Networks (SOFNN) on the Twitter feeds and DJIA values from the period June 2009 to December 2009.
- We also implement a naive protfo- lio management strategy based on our predicted values.
- Our work is based on Bollen et al's famous paper which predicted the same with 87% accuracy.

INTRODUCTION

- Stock market prediction has been an active area of research for a long time.
- We perform sentiment analysis on pub- licly available Twitter data to find the public mood and the degree of membership into 4 classes - Calm, Happy, Alert and Kind (somewhat like fuzzy membership).
- The authors used Self Organizing Fuzzy Neural Networks to predict DJIA values using previous values.
- Section 4 discusses the sentiment analysis technique developed by us for the purpose of this paper.
- Section 5 includes in detail, the different machine learning techniques to predict DJIA values using our sentiment analysis results and presents our findings.

SENTIMENT ANALYSIS

- Sentiment analysis was an important part of our solution since the output of this module was used for learning our predictive model.
- In order to do automate this analysis for tweets, the word list needs to be appro- priately extended.
- We cross validated the results of our sentiment analysis technique by comparing the values returned by our algorithm around significant events like Thanksgiving day and Michael Jackson's death.
- Granger Causality analysis finds how much predictive information one signal has about another over a given lag period.
- In the next section, we use the results of our sentiment analysis algorithm to learn a model that can predict the stock index and its movement.

DATASET

- In this project, we used two main datasets- • Dow Jones Industrial Average (DJIA) values from June 2009 to December 2009.
- The data was obtained using Yahoo!
- Finance and includes the open, close, high and low values for a given day.

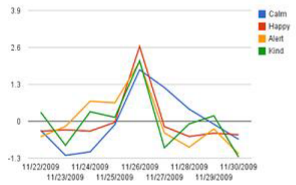
SENTIMENT ANALYSIS

- Table 1: p-values obtained using Granger causality analysis with different lags (in days)

Lag	Calm	Happy	Alert	Kind
1	0.0207	0.4501	0.0345	0.0775
2	0.0336	0.1849	0.1063	0.1038
3	0.0106	0.0658	0.1679	0.1123
4	0.0069	0.0682	0.3257	0.1810
5	0.0100	0.0798	0.1151	0.1157

SENTIMENT ANALYSIS

- Figure 2: Cross validation of our sentiment analysis by analyzing moods on some important events



Referances

- *[1] J. Bollen and H. Mao. Twitter mood as a stock market predictor. IEEE Computer, 44(10):91–94.*
- *[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.*
- *[3] G. P. Gang Leng and T. M. McGinnity. An on-line algorithm for creating self-organizing fuzzy neural networks. Neural Networks, 17(10):1477–1493.*
- *[4] A. Lapedes and R. Farber. Nonlinear signal processing using neural network: Prediction and system modeling. In Los Alamos National Lab Technical Report.*
- *[5] A. E. Stefano Baccianella and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In LREC. LREC.*

REFERENCES

[1] Yue Hu and Xiaojun Wan, PPSGen: Learning-Based Presentation Slides Generation for Academic Papers, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 4, APRIL 2015

[2] Automatic Generation of Related Work Sections in Scientific Papers: An Optimization Approach ,Yue Hu and Xiaojun Wan, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 16241633, October 2529, 2014, Doha, Qatar. c 2014 Association for Computational Linguistics

[3] Biju P. Smitha C.S. "A Support System for Making Presentation Slides", Transactions of the Japanese Society for Artificial Intelligence 01/2003; 18:212-220

[4] Investigating Automatic Alignment Methods for Slide Generation from Academic Papers,Brandon Beamer and Roxana Girju, Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 111119

[5] T. BergKirkpatrick, D. Gillick, and D. Klein, "Jointly learning to extract and compress, in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Lang. Technol., 2011, pp. 481490

[6] Abstractive Summarization of Line Graphs from Popular Media Charles F. Greenbacker Peng Wu Sandra Carberry Kathleen F.

McCoy Stephanie Elzer, Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, pages 4148, Portland, Oregon, June 23, 2011. c 2011 Association for Computational Linguistics

[7] V. Qazvinian, D. R. Radev, S. M. Mohammad, B. J. Dorr, D. M.Zajic, M. Whidby, and T. Moon, "Generating extractive summaries of scientific paradigms," J. Artif. Intell. Res., vol. 46, pp. 165201, 2013.

[8] A. Abu-Jbara and D. Radev, "Coherent citation-based summarization of scientific papers," in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Lang. Technol.-Volume 1, 2011, pp. 500–509.

[9] V. Qazvinian, D. R. Radev, S. M. Mohammad, B. J. Dorr, D. M. Zajic, M. Whidby, and T. Moon, "Generating extractive summaries of scientific paradigms," J. Artif. Intell. Res., vol. 46, pp. 165– 201, 2013

[10] V. Qazvinian and D. R. Radev, "Identifying non-explicit citing sentences for citation-based summarization," in Proc. 48th Annu. Meeting Assoc. Comput. Linguistics, Jul. 2010, pp. 555–564.

[11] M. A. Whidby, "Citation handling: Processing citation texts in scientific documents," Doctoral dissertation, Dept. Comput. Sci., Univ. Maryland, College Park, MD, USA, 2012.

[12] R. Jha, A. Abu-Jbara, and D. Radev, "A system for summarizing scientific topics starting from keywords," ACM Comput. Surv, vol. 40, no. 3, p. 8, 2013.