AUTOMATIC GENERATION OF PRESENTATION SLIDES FOR ACADEMIC PAPERS USING INTEGER LINEAR PROGRAMMING

1N. Yogapreethi, 2S. Maheswari

1,2 UG Scholar Department of computer science and engineering
Nanha Engineering College, Erode.

Abstract—Presentations are one of the most common and effective ways of communicating the overview of a work to the audience. Given a specialized paper, programmed era of introduction slides diminishes the exertion of the moderator and aides in making an organized synopsis of the paper. In this Project, we propose the structure of another framework that does this assignment. Any paper that has a theoretical and whose segments can be ordered under presentation, related work, model, examinations and conclusions can be given as info. This XML record is parsed and data in it is extricated. An inquiry particular extractive summarizer has been utilized to create slides. In numerous meetings and social occasions, a moderator takes the guide of slides to present his work deliberately (pictorial). The introduction slides for the paper are then made by utilizing the Integer Linear Programming (ILP) demonstrate with complicatedly organized target breaking points and goals to pick and adjust key expressions and sentences. The framework removed theme and non subject parts from the article in light of the basic talk structure. The isolated subject and non-topic parts were determined to the slides by giving fitting indents in perspective of the examination of their syntactic structure.

Keywords—Abstracting methods, Integer Linear Programming, Support Vector Regression model, text mining.

I. INTRODUCTION

Web Mining is a data mining technique to discover pattern from World Wide Web. Web Mining can be classified into three types such as Web Content Mining, Web Structure Mining and Web Usage Mining. Web Content Mining is used in extracting, integrating and mining useful information from Web page content.

Web Structure Mining is used to identify and analyze the nodes in a network connection and it is based on graph theory. Web Usage Mining is used to estimate the weblogs. It is used to understand the browsing behaviour of a user and to discover pattern for serving the needs of a web based applications. Automatic slide generation is a method of Web Content Mining which is used in extracting vital points from the document and preparing slides. A slide is a single page of a presentation. Collectively, a group of slides may be known as a slide deck.

Presentation is the process of presenting a topic to audience. It is typically a demonstration, lecture, a speech meant to inform, persuade or build a good will. A collection of pages arranged in sequence that contain text and images for presenting to an audience. It often refers to a Microsoft PowerPoint presentation. In the digital age, a slide most commonly refers to a single page developed using a presentation program such as Microsoft PowerPoint or Apple Keynote. It is also possible to create them with a document mark-up language, for instance with the Latex class Beamer. Lecture notes in slide format are referred to as lecture slides, frequently downloadable by students in .ppt or .pdf format. Presentation slides can be done by different types of software such as Microsoft
PowerPoint, Apple Keynote, Prezi, Slide Dog, ClearSlide, etc., Slides have been an effective and popular means of presentation of information. Slides provide an easy way of sharing knowledge between different groups of people.

Microsoft PowerPoint supported in Windows, Open office Presenter in Linux is used only for formatting slides but not in content generation. More recently, Google Slide had been developed by Google for creating and formatting slide in open source but not in content.

These tools help only in the formatting of content such as alignment, bulleting, font style and size etc., but not in preparing the content itself. A user has to start taking notes from the academic paper and then vital points in the paper is identified and used in presentation which is a time consuming task. But this concept propose a methodology that generates slides for the presentation with important points and all necessary figures, tables and graphs from a technical paper. It has been evident that proposed technique saves time and reduces the effort of presenter by providing a basic presentation, which can be further upgraded as final presentation.

**Basic Definitions**

Presentation slides have been a popular and effective means to present and transfer information, especially in academic conferences. The researchers always make use of slides to present their work in a pictorial way on the conferences. There are much software such as Microsoft PowerPoint and Open Office to help researchers prepare their slides. However, these tools only help them in the formatting of the slides, but not in the content. It still takes presenters much time to write the slides from scratch. In this work, we propose a method of automatically generating presentation slides for academic papers. We aim to automatically generate well-structured slides and provide such draft slides as a basis to reduce the presenters’ time and effort when preparing their final presentation slides.

**Sentiment Analysis And Polarity Shift:**

According to the levels of granularity, tasks in sentiment analysis can be divided into four categorizations: document-level, sentence-level, phrase-level, and aspect-level sentiment analysis. First, we strengthen the DSA algorithm by adding a selective data expansion procedure. Second, we extend the DSA framework from sentiment polarity classification to positive-negative-neutral sentiment classification. Third, we propose a corpus-based method to construct a pseudo-antonym dictionary that could remove DSA’s dependency on an external antonym dictionary.

**Data Expansion Technique:**

The field of natural language processing and text mining, a girre and Martinez proposed expanding the amount of labeled data through a Web search using monospermous synonyms or unique expressions in definitions from Word Net for the task of word sense disambiguation. Fujita and Fujino proposed a method that provides reliable training data using example sentences from an external dictionary. To the best of our knowledge, the data expansion technique proposed here is the first work that conducts data expansion in sentiment analysis. Different from the above mentioned techniques, the original and reversed reviews are constructed in a one-to-one correspondence.

**Objectives Of The Presentation Slides**

Presentation slides have been a prominent and powerful intends to present and exchange data, particularly in scholarly gatherings. The analysts dependably make utilization of slides to display their work pictorially on the meetings. There are numerous virtual products, for example, Microsoft PowerPoint and Open Office to help specialists set up their slides. Be that as it may, these
Devices just help them in the designing of the slides, yet not in the substance. Regardless it takes moderators much time to compose the slides starting with no outside help.

In this work, we propose a strategy for consequently creating presentation slides for scholastic papers. We expect to naturally create very much organized slides and give such draft slides as a premise to diminish the moderators’ opportunity and exertion while setting up their last presentation slides. Scholarly papers dependably have a comparative structure. They by and large contain a few areas like conceptual, presentation, related work, proposed technique, investigations and conclusions. Despite the fact that presentation slides can be composed in different courses by various moderators, a moderator, particularly a apprentice, dependably adjusts slides consecutively to the paper segments while setting up the slides.

II.RELATED WORK

Automatic slides generation for tutorial papers remains way under-investigated today. Few studies directly analysis on the subject of automatic slides generation. In M. Sravanthiet. al. [1] introduces the answer for reducing the hassle of the presenter and facilitate them in making a structured outline of paper. It helps in making slides for presentation with vital purpose and al necessary figure etc. The very important points of the paper can mention. The Latex document is provided because the input and born-again into xml format. The xml file can break down and extract the knowledge. a question specific extractors can facilitate to summarize and generate the slides.

Utiyama et al., [2] planned a brand new theme that used the GDA tagset annotated version of an editorial to find out the linguistics structure of the article. This data was used for locating the vital topics. Sentences similar to these topics are extracted that were then organized on to the output slides. Yasamura et al., [3] enforced an answer to get presentation from the LaTeX manuscript of a technical article. The TF-IDF evaluation theme was accustomed calculate the weights of all terms within the article thus on establish relevancy score for all document objects. The term weights were accustomed confirm the dimensions of every section outline. The output slides were customizable by the user.

Sravanti et al., [4] detailed a system to machine generate presentation slides from a quest manuscript. Here also, the start line of slide generation is from the raw LaTeX supply of the analysis manuscript. once the logical thinking of the logical structure from the article, every section was classified to be Introduction, connected Works, Model, Experiments and Conclusion severally. the method of automatic slide generation during this technique concerned the utilization of QueSTS summarizer [5]. Graphical parts may even be extracted from the article by the system and also the slides were engineered.

Shibata et al., [6] represented a technique for the generation of presentation slides from an editorial by the analysis of the discourse structure of the article. A clause and sentence was thought of as a discourse unit by the system and also the vital coherence relations like distinction, list, additive, elaboration etc were extracted and analyzed. Topic and non topic components were known victimization the discourse structure of the text. The output slides were created by having correct intends to the contents thus on enhance readability. The sentences are connected to the foremost similar preceding sentences. components having less importance are cropped supported some heuristic measures. K.

Gokul Prasad et al., [7] planned a brand new theme to form presentation slides for seminars and lectures. the two modules – data Extractor and Slide Generator extracts the text contents from the article and thence uses common IP operations of text segmentation and constellation to spot the noun phrases and segments. The system made an ontology tree for every phrase detected employing a chunker system. The metaphysics and weight values calculated were used for positioning key phrases and contents for bullet points and thence, the shows were generated.
Tulasi Prasad Sariki et al., [8] given a completely unique theme to get presentation slides by ab initio taking the document that the slides were to be generated. The system then enforced numerous basic preprocessing techniques like sentence division, case folding, stop word removal, stemming and lemmatization to the document. Individual sentences were thought of and a mix of in style baseline summarizers was accustomed realize relevancy score for every sentence. The system is capable of accepting keyword queries and building a presentation specific to the input question.

ShaikhMostha Al Masum et al., [9] detailed a brand new agent based mostly theme wherever within the user may offer queries as input. within the background the system collected data regarding the question by looking the net. pictures may even be additional to the output slides by the system. The system worked on numerous techniques like net information taking, website parsing and outline extraction.

Misturu Ishizuka et al., [10] mentioned a brand new theme by acceptive keywords from the user and generating a compact report and presentation by querying the net. The system worked on completely different steps, every of that completed by code agents. If the keywords were ambiguous, the disambiguated senses were additionally additional to the search keys. The summarisation theme used a vector distance for measurement the closeness between sentences. The system generated a report specific to every topic and from each report, a presentation was engineered.

Yue Hu et al., [11] approached the task of automatic slides generation by elaborating a theme that followed a corpus based mostly machine learning approach. The system worked in two phases to get slides from a quest article. There are some limitations and challenges in previous strategies, e.g. Extraction of image and tables from the bottom paper. to beat those we have a tendency to planned a technique that selects range of vital sentences, images, tables and also the phrases from the corresponding base paper. Calculation of the importance of sentence with image reference is difficult task.

### III. METHODOLOGY

PPSGen investigated the method of automatically generating presentation slides for academic papers. This method first employs the regression method to learn the importance scores of the sentences in an academic paper, and then exploits the integer linear programming (ILP) method to generate well-structured slides by selecting and aligning key phrases and sentences. Evaluation results on a test set of 200 pairs of papers and slides collected on the web demonstrated that PPSGen system can generate slides with better quality [29]. Sentence assessment is done based on Support Vector Regression (SVR) [22]. Then by using extracted sentence from the paper slide is generated. Slide generation is based on Integer Linear Programming (ILP). Then after postprocessing generated output is the presentation slide. The disadvantage of the system is that it considers only text elements into the slide. Graph elements such as tables and figures are not considered. Multi document summarization involves multiple aspects of content selection and surface realization. Multi summarization method is used to create summaries from multiple papers based only on integer linear programming (ILP). The summaries must be informative, succinct, grammatical, and obey stylistic writing conventions [19]. It learns individual aspects but optimized jointly using an integer linear programme. The ILP framework allows combining the decisions of the expert learners and to select and rewrite source content through a mixture of objective setting, soft and hard constraints. The disadvantage of this system is that it focuses on summarization of the content. It does not generate slides and it does not consider graph elements. Sentence importance assessment is not up to quality standards. SlidesGen investigated automatic generation of presentation slides from technical papers in LATEX. A query specific extractive summarizer QueSTS is used to extract sentences from the text in the paper to generate slides. QueSTS transfers the input text to an integrated graph (IG) where a sentence represents a node and edges exist between the nodes that the sentences corresponding to them are similar. The weights of the edges are calculated as cosine similarity between the sentences [18]. SlidesGen framework
Includes steps such as Preprocessing, Configuration file generation, Extracting key phrases and QueSTS Summarizer [21].

Finally slides are generated for each section in the paper and graphics are rendered in presentation. The disadvantage of the system is that it takes input only as latex format, wherein all documents cannot be considered as input. Most of the documents are in PDF format so many documents cannot be processed. Term Frequency-Inverse Document Frequency (TF-IDF) investigated a support system for making presentation slides from a technical paper. This system provides functions that assign slides to each section and put objects on a slide [28]. The system calculates weights of terms in the document by the TF-IDF method. Based on the term weights, objects in the document such as sentences, figures and tables are weighted. Finally, outputs of the system are generated as presentation slides in XHTML. An extractive multi-document summarizer [6] forms summaries by extracting or selecting sentences from the input documents, without modifying the selected sentences. The disadvantage of the system is that it calculates the weight of the sentences based on the term weight. This method is also used only for summarization and the sentence extracted using TF-IDF are less important when compared with Support Vector Regression.

Integer Linear Programming (ILP) method investigated a system to generate extractive multi-document summaries. The method uses Integer Linear Programming to jointly maximize the importance of the sentences it includes in the summary and their diversity, without exceeding a maximum allowed summary length [6]. To obtain an importance score for each sentence, it uses a Support Vector Regression model trained on human-authored summaries, whereas the diversity of the selected sentences is measured as the number of distinct word bigrams in the resulting summary. Experimental results on widely used benchmarks show that our method achieves state of the art results, when compared to competitive extractive summarizers, while being computationally efficient as well [14]. The disadvantage of the system is that it generates summaries depending on the human trained corpus. The area for this system is narrow, it generates summaries only for specific document depending on the corpus.

IV.SYSTEM ANALYSIS AND DESIGN

Existing System And Its Drawbacks

PPSGen investigated the method of automatically generating presentation slides for academic papers. This method first employs the regression method to learn the importance scores of the sentences in an academic paper,

Disadvantages of Existing system

Focuses on summarization of the content.
It does not generate slides and it does not consider graph elements.
Sentence importance assessment is not up to quality standards.
Most of the documents are in PDF format so many documents cannot be processed.
Feature selection is to select a subset of terms occurring in the training set and using only this subset as features in text categorization.

Proposed System And Its Advantages

Slides have been an effective and popular means of presentation of information. Slides provide an easy way of sharing knowledge between different groups of people. A presenter takes the help of slides to present his work in a pictorial and more convenient way in many conference and meetings. A user has to start taking notes from the academic paper and then vital points in the paper is identified and
used in presentation which is a time consuming task. But this concept propose a methodology that generates slides for the presentation with important points and all necessary figures, tables and graphs from a technical paper. Our Proposed system uses SVR and ILP technique to remove the barriers of existing one.

Advantages of Proposed System:

- Multi document summarization involves multiple aspects of content selection and surface realization. Multi summarization method is used to create summaries from multiple papers based only on integer linear programming (ILP).
- The ILP framework allows combining the decisions of the expert learners and to select and rewrite source content through a mixture of objective setting, soft and hard constraints.
- This method generate extractive summaries of a set of Question Answering (QA) and Dependency Parsing (DP) papers

Module Description

Topics are given to each cluster in the document using lexical analysis. Sentence importance assessment is done based upon Support Vector Regression (SVR) which extracts important sentence from the document and put it into slide. Slide is created by placing necessary text and graph into the slide which obeys slide writing conventions using integer Linear Programming (ILP).

Support Vector Regression (SVR)

Support Vector Regression is used in maintaining all the main features which characterize the maximal margin of the algorithm. The Support Vector (SV) algorithm is a nonlinear generalization of the Generalized Portrait algorithm developed in Russia in the sixties. The key idea is to construct a Lagrange function from the objective function which is called as the primal objective function and the corresponding constraints, by introducing a dual set of variables. It can be shown that this function has a saddle point with respect to the primal and dual variables at the solution.

Integer Linear Programming

An integer linear some or all of the variables are restricted to be integers. In many settings the term refers to integer linear programming (ILP), in which the objective function and the constraints are linear.

Stop word removal

Stop words are language specific functional words which carry no information. It may be of the following types such as pronouns, prepositions, conjunctions. Stop word list are contained in a list which contains maximum of stop words. Once the file is read it checks the uploaded file with the stop word dataset. If the words in the document and dataset are matched, then the corresponding word is removed from the document. Then the non-stop word is send as input to calculate word frequents.

Feature Extraction

Feature extraction is used to extract important features such as word frequent, sentence position, word overlap with the title and sentence parse tree information. In this system only word frequent is mined from the document. Here maximum number of repeated words in the file is considered for analyzing and is used to know which topics are explained in the paper. In this module, the system gets the non-stop words as input and calculates the count of words and finds the repeated occurrence of each and every word from the non-stop words.
Sentence Extraction is done based on the topics generated during topic modelling. Sentence Extraction is done to create summary report for the given base paper. Summaries are used to provide an overview of the given base paper. Sentence Extraction is based on Support Vector Regression.

**Slide Creation**

Using sentence extracted from the input document presentation is generated by obeying general stylistic conventions. Slide creation is done by integer linear programming (ILP). Generated slide can be downloaded from the web or it can be retrieved from the specific folder. Slide creation must incorporate

**Fig 1: Proposed System Architectures**

**V. RESULT AND DISCUSSIONS**

Conversion of PPT through the given content

**ENHANCED STYLE**
VI. CONCLUSION

This paper proposes a novel framework called PPSGen to generate presentation slides from academic papers. We train a sentence scoring model based on support vector regression and use the integer linear programming method to adjust and concentrate key expressions and sentences for producing the slides. Test comes about demonstrate that our strategy can generate much better slides than traditional methods. In future work, we will enhance our framework by utilizing both content and graphical components in the paper and make slides more understandable and clear. When managing the graphical components, we have to recognize the graphical elements in the paper first. The relationship between the content components and the graphical components likewise should be identified. We have to know which sentences are most pertinent to a graphical component and which graphical components ought to be chosen to create the slides. We can utilize govern based strategies or machine learning based techniques to take care of the above issues. At that point we can essentially join the tables and figures. we select to the most important sentences in the slides. In this paper, we only consider one typical style of slides that beginners usually use. And we will consider more complicated styles of slides such as styles that slides are aligned sequentially with the paper and styles that slides have more hierarchies.

REFERENCES


[5]. M. Sravanthi, C. R. Chowdary, and P. S. Kumar, „QueSTS: A query specific text summarization approach“, in Proc. 21st Int. FLAIRS Conf., [2008], pp.[219]-[224].


[8]. Sariki, Tulasi Prasad, Bharadwaja Kumar, and Ramesh Ragala. „Effective Classroom Presentation Generation Using Text Summarization“. 

