MAT JOURNALS

Journal of Computer Science Engineering and Software Testing

Volume 2 Issue 3

Generating Presentation Slides for Academic Paper using SVR and ILP Technique

V. Suresh, G. Selva Priya

Department of CSE, Dr. N. G. P. Institute of Technology, Coimbatore, India
E-mail: suresh@drngpit.ac.in

Abstract

PowerPoint Presentation is a common means of mechanism for a person to project his or her view meaningfully and pictorially. Accordingly, generating slides to make effective presentations is a tedious work in present days. In this view, a novel system PPS Gen is used to generate presentation slides which can be used as a draft by all the stakeholders. The slides not only have text elements but also has graphical element named as figures and tables. The existing work focuses only on text elements. This paper proposes a model that focuses on graphical elements additionally. The model first uses the Support Vector Regression (SVR) method to learn the strength of relationship between the sentences to make effective presentations, further the method of Integer Linear Programming (ILP) used to select, align key phrases and sentences. The final slides will have good structure and content quality from academic papers.

Keywords: *SVR, effective presentation, ILP, slides*

INTRODUCTION

Presentation slides are an effective means of transferring the information in many fields including business and education among others. Text mining is also referred to text data mining. The process of deriving high-quality information from text is called as text analytics. Text mining tasks include text categorization, text clustering, concept/entity extraction, document summarization and entity relation modelling. The overarching goal is essentially to turn text into data for analysis via application of Natural Language Processing and analytical methods [1]. The researchers always use a slide to present their work in a pictorial way on the conferences. The creation of slides is a difficult work now-a-days and it is a time consumption process. To overcome this, the system aims to generate a well-structured slide for an academic paper [2, 3]. The draft slide is produced to reduce the time and effort of presenters. For preparing, their final slide [2].

In this work, the proposed tool that generates slides for presentation with important points and all necessary figures, tables and graphs from technical paper. The slide includes the summarization content in each topic and aligning these topics to one or more slides and placing necessary graphical content at appropriate locations [4, 5]. Automatic slide generation for academic papers is a very challenging task [1]. Slides can be divided into an ordered sequence of parts. Each part addresses a specific topic and these topics are relevant to each other. The novel system called PPS Gen is proposed to generate a well-structured presentation slides. SVR model is used to find the importance of each sentences of a paper [6–12]. It also used to avoid over fitting in slide generation. Based on the regression of the word a score is generated. Where an SVR is trained on a corpus collected on the web. The presentation slides are generated regressively by using the ILP model [3, 10, 12].

This model is also used to select and align key phrases and sentences in an appropriate slide. To split the data into a training dataset the KNN can used to make predictions and a test dataset that we can used to evaluate the accuracy of the model. Therefore, our slides are considered to be a better basis to prepare the final slides.

RELATED WORK

In automatic generation of presentation slides reduces the presenter's effect [1]. It also helps in creating structure in more effective way. The framework of novel system is proposed for creating slides. Input to the system is LATEX document. XML file is passed to extract information. Easier presentation slides are available in recent years as many software tools like Microsoft power point. Open-office presenter but all these used only for content formatting instead of preparing content itself. The proposed concept explains generating slides for research papers with standards. The problem is the use of terms such as conference paper, technical paper, document and report. Good quality presentation and starting point for preparation of final presentation. This can be done using natural language presenting technique to compress the sentence. To prepare slides with ease by constructing slide layouts from the text for expression styles of words the skeleton generation method is designed. The expression styles of the words presented in the slide from text book are analysed then it extract context-role of the words by the difference between the words in the text and slides [2].

It also derives the words from pre-existing text and slides. The words are expressed in different ways in different slides by creating or exploring slide skeleton. The document structures are derived from the text by focusing on their logical units and also focusing the levels of indentation of slides text there are often used for the better organize their slide contents. By using, a

large set of actual texts and slide pairs. We can improve the algorithm for the generation of skeleton of slides. By Metadata is extracted to evaluate the tools performance [4]. Metadata based on arXiv collection, GROBID and Mendely Desktop. By using various tools PDF document is obtained. Two stage of SVM are used to solve the metadata extraction problem: 1) Comprehensive metadata 2) Fundamental methods. Fundamental methods used for extracting the metadata are stylistic analysis, machine learning and use of knowledge bases. Stylistic analysis extract title, machine learning supports SVM and Hidden HMM and Knowledge base is helpful to use database.

From summarization algorithm is used to combine content model with discourse model for generating coherent survey and readable summary for scientific document [9]. The set of input papers related to question and answering is used for generating coherent summary. Human survey does not exist for all topics. Hence automated system is build for developing a summary. For flexible relationship Minimum Independent Discourse Context (MIDC) is used. The summarization approaches are content model and discourse model which includes supervised and unsupervised word sense disambiguation for detailed overview that includes subtopic using Hidden Markov Model (HMM). Discourse model where common problems are overcome by extraction summary. The sentence from original document is not always easy to understand when it is pulled out as it is.

To avoid this MIDC is used. MIDC is calculated for each sentence by initializing the first topic as input document set followed by the subtopics using HMM. HMM is used to identify whether the topic in the input document set is valid or invalid MIDC. The author introduced the effort to automatically align the spoken utterances in recorded lectures with the content of the slides used. A set of approaches considering the problem that words helpful for such alignment are spare and noisy. The assumption of that presentation of a slide is usually smooth and top-down across the slide. This includes the structured SVM learning from local and global features [6].

By multimodal system is present for aligning scholarly document to corresponding presentation in a fine-grained manner [5]. This method improves upon a state-of-the-art baseline that employs only textual similarity. A three pronged alignment system that combines image, textual and ordering information based on an analysis of errors made by the baseline. Scholarly document and presentation slides are the two media that is used to provide a clear function. The images can be classified in to two categories such as Natural (photographs) and Synthetic (computer generated images). Synthetic images are maps images, icons, figures, cartoons and art

work. The presentation slides should have relative size, text alignment, lack of smaller images on vertical grids and width to height ratio. The issue here is slide type classification such as pure text slide, outline, drawing, result slides. Performance alignment accuracy is increased.

The ILP framework bigram based supervised method are proposed to extract the document summarization [3]. The frequency of bigram regression model is used. Sentence selection problem is formulated in the time of testing. TAC evaluation is performed to demonstrate the performance of ILP system. Sentence selection process is extracted from one or multiple documents using many methods such as supervised approaches to predict summary sentences, Graph based approaches to rank the sentences, ILP and Sub modular methods. These methods are powerful to select the important sentences and remove redundancy.

The proposed one is to find a candidate summary to evaluate the ROUGE metric used for summarization. Bigram frequency is used to extract the summarization and also act us language concept. This regression model is used to calculate word level and sentence level features. Supervised learning method is used to determine bigram. ILP based summarization is widely adopted for state-of-the-art performance [10]. Two new modifications are used for updating. It measures both the salience and the novelty of words in sentences. The first feature is to predict the weights, second is to generate preliminary sentence candidates and finally re-ranking them. This method evaluates different TAC and is applied for multilingual news summarization in addition to generative models. A novel non parametric Bayesian approach is proposed. Further, the idea of evolutionary clustering is borrowed and three levels of HDP model is used to represent the diversity and commonality. The importance and novelty of the bigram concepts used in the ILP model. The pilot research point out new directions for generic or update summarization based on the ILP framework.

Standard graph ranking algorithm is used and it has two layers such as Sentence layer and Topic layer which are used to improve the summary performance [7]. MDS is used to facilitate users to grasp the main idea of the documents. The score of the given query is used to rank each sentences. The sentences are picked into a summary based on the ranking. Graph based semi supervised learning algorithm is an effective method to impose a query's influence on sentences. Positive score is assigned to the query and zero is assigned to the remaining node. Then the nodes are spread into neighbour nodes this is repeated until all nodes obtain their final ranking score. The issue arise here is sentences are ranked without considering topic level information. By using traditional clustering algorithm better results are achieved.

The system used plagiarism detection or research paper recommendations are used. Manually constructing the whole queries is more difficult. Instead of motivating Simseer X-Search Engine is used to retrieve the whole document as queries. This method is able to work with multiple similarity functions and document collections [8] .In this paper, the model present SimseerX, a similar document search engine framework. It can be used to find similar files in a collection of documents and it can support many different types of similarity scoring functions and document collections. It can also be used to compare and evaluate new similarity functions that can be plugged into the system. The system was designed to allow the user to submit full document in a collection. Thus a future feature could involve displaying the results from different similarity functions side by side or combining them into a single ranked list.

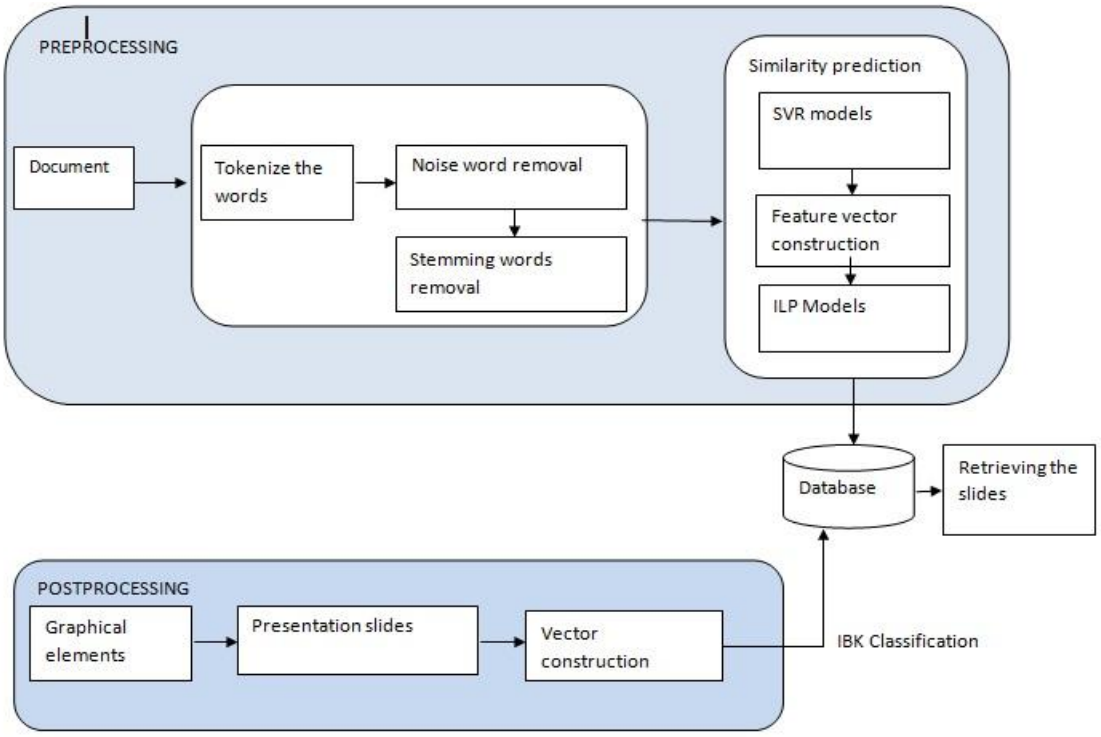

Fig. 1: *Proposed Architecture.*

Identified Modules
The implementation section comprises of:

1. Data Acquisitions.
2. Preprocessing.
3. Clustering.
4. Post Processing.

Dataset Acquisitions

The system used unstructured (Documents) information to extract meaningful numeric indices from the text. It can analyze all types of text documents to improve the quality level. Even temporal information may be informative for mining processing.

Pre-Processing

Data pre-processing is an important step in the data mining process.It has been used to detect irrelevant and redundant or noisy information and unreliable data present. Tokenize the words stripping and prefix stripping. Noise words are known as stop words.

Clustering

The module analyzed the slide information to eliminate noises and improve quality of text documents. Clustering data sources are used in group data. It provides improved results in slide complexity overhead. Combination of the text and side-information used on creating a cluster-based model.

Post Processing

Document classification categorization on slide retrieves the datasets. The constrains for graphical elements for Stemming words analysis is made. Initialization phase cluster training data are retrieving. The presented methods for mining graphical data with the academic papers of slide-information are included. The present results on real data sets illustrating the effectiveness of our approach.

CONCLUSION

This paper proposes a novel system called PPSGen to generate presentation slides from academic papers. SVR algorithm is used to train a sentence scoring model and ILP algorithm is used to extract aligned key phrases and sentences. To generate a good quality presentation with a good starting point, a technical paper is given in LATEX format and introduces a method of skeleton-generation for making expression styles of words. And this holds ILP method as a core component in our summarization system to propose a supervised learning method and to minimize bigram gain. GROBID has the advantage of producing score accuracy and non-lost information in the PDF format. Topic modelling techniques are used for sentence clustering and further graph construction. To focus on different kinds of information such as background or document specific information, two LDA

topic model extensions are used. Structured SVM is considered for sparse/noisy word problem and the sequential smoothness assumption for slide presentation. Supervised ILP framework is used for the update summarization task. Experimental result shows that our method can generate much better slides than traditional methods.

REFERENCES

1. M. Sravanthi, C. Ravindranath Chowdary, P. Sreenivasa Kumar. Slide Gen: Automatic generation of presentation slides for a technical paper using summarization. *International FLAIRS Conference*; 2009.
2. Yunayuan Wang, Kazutoshi Sumiya. Generating slides skeletons based on expression styles for presentation contents; 2012.
3. Chen Li, Xian Qian, Yang Lin. Using supervised bigram-based ILP for extractive summarization; 2013.
4. Bamdad Bahrani. Multi model alignment of scholarly documents and their presentation; 2013.
5. Mario Lipinski, Kevin Yao, Corinna Breitinger, et al. Evaluation of header metadata extraction approaches and tools for scientific PDF document. *ACM/IEEE-CS Joint Conference on Digital Libraries*; 2013.
6. Han Lu, Sheng-syun Shen, Sz-Rung Shiang, et al. Alignment of spoken utterances with slide content for easier learning with recorded lectures using structured SVM; 2014.
7. Kyle Williams, Jian Wu, C. Lee Giles. SimseerX-A similar document search engine; 2014
8. Yanran Li, Sujian Li. Query focused multi-document summarization: combining a topic model with graph-based semi supervised learning; 2014.
9. Chen Li, Yang Liu, Lin Zha. Improving update summarization via supervised ILP and sentence reranking; 2015.
10. Kazunari Sugiyama, Min-Yen Kan. A comprehensive evaluation of scholarly paper recommendation using potential citation papers; 2015.
11. Rahul Jha, Reed Coke, Dragomir Radev. Surveyor-A system for generating coherent survey article for scientific topics; 2015.
12. Yue Hu, Xiaojum Wan. PPSGen: Learning to generate presentation slides for academic papers. *International Joint Conference on Artificial Intelligence*; 2015.