ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

AUTOMATED SUMMARIZATION OF ACADEMIC PAPERS AND GENERATION OF PRESENTATION SLIDES

Deni Avinash.A.B, Senthilkumaran U, Manikandan N
School of Information Technology, VIT University, Vellore, India.
Email: usenthilkumaran@vit.ac.in

Received on 25-10-2016 *Accepted on 02-11-2016*

Abstract

Presentation slides have been a popular and effective means to present and transfer information, especially in academic conferences. There are many software such as Microsoft Power- Point 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 presenter much time to write the slides from scratch. In this work, we propose a method of automatically generating presentation slides for academic papers using some major datamining techniques with the help of WordNet and Lexical Analysers for Extraction and Summarization. Thus, we aim to automatically generate well-structured slides and provide such draft slides as a basis to reduce the presenter's time and effort when preparing their final presentation slides.

Introduction

It is difficult for researchers and presenters to read and summarize academic papers and create slides for presentation purpose from the scratch, so we try to develop a system that automatically summarizes the academic papers and generates draft slides that will be useful to create the final set of slides. Engineering students and researchers spend most of their times in reading the research papers for their academic purpose and it is always difficult for them to read the entire paper from the scratch. To make their work easier and to get an idea of a particular paper, a tool that would assist this task is needed.

Background: A slide is a single page of a presentation. Collectively, a group of slides may be known as a slide deck. In the latter part of the 20th century, a presentation slide was created on a transparency and viewed with an overhead projector. 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 markup 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. It's is an easy tool to explain the audience about what our work is about.

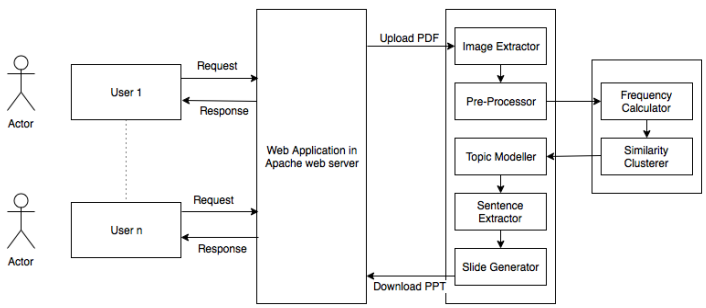
Innovation: A tool that generates slides for the presentation with important points and all necessary figures, tables and graphs from a technical paper. As it is evident, such kind of a tool saves time and reduces the effort by providing a basic presentation, which can be further tuned or upgraded as final presentation.

Literature Survey: Proposed system that can generate a good quality presentation, from a technical paper given in LATEX format. These slides provide a good starting point for the preparation of final presentation. We make a good use of the richness of the LATEX document markup language and generate slides using only statistical processing[2]. Generating slides consists of discourse structure analysis, extraction of topics/non-topic parts and displaying them based on discourse structure. While they are improving the accuracy of detecting discourse structure and reducing the non-topic parts, they are planning to integrate text-to-presentation system with embodied conversational agents to enhance the presentation contents[3]. They investigated the task of automatic slide to paper alignment. We built a corpus of slide-paper pairs and used four presentations from it to evaluate four aligners which utilize methods such as TF-IDF term weighting and query expansion[4]. Proposed the notion of an integrated graph that represents inherent structure present in a set of related documents by removing redundant sentences[5].

Limitations: It is difficult for researchers or students to go through the entire content of research papers and come up with the summarized idea to present it. Existing summarization techniques does not deal with the images and graphs in the research papers. The existing text summarization techniques have the problems of Alignment of the extracted text and do not focus on extracting the important graphs, diagrams and stats to the generated slides.

Proposed System: We propose a system to automatically generate slides that have basic structure and content quality from academic papers using data mining techniques which has some important points in it and segregates all necessary figures, tables and graphs from a technical paper and saves it in a local folder.

Proposed Framework



A. Image Extractor and Pre-Processor-

Load the input PDF document which we want as power point presentation. First the images in the document is extracted and stored in the local disk. Pre-processing is data cleaning for which we use stop word removal method, this method reads word from the input file and checks with stop word dataset, if the word exist in that dataset then this method ignores that word. Set of non-stop words are generated.

B. Frequency Calculator-Set of non-stop words are preprocessed. Next we are calculating word count and finding the repeated occurrence of each and every word from the non-stop words. Words with their Count is generated.

C. Similarity Mining- Words with their count is sent and Unwanted Affix and Suffix of words are removed using Stemmer and Words with similar meaning Clustered together using WordNet. Cluster of Words eligible to become topics are generated.

D. Topic Modeling- Cluster of Words are sent and we are Comparing the words with the database table that helps in identifying the topics and also get the keywords from paper as topics. Set of Topics are generated.

E. Sentence Extractor PDF file and Clusters with Topics is sent as input then we are Splitting the pdf file line by line and check if the topic word occurs in the line and extract it.

F. Slide Generator- Topics of cluster as titles of the slides and sentence as slide points are sent in it. Generation of presentation slides using topics as titles and sentences as slide points is generated as Presentation slides.

Methodology

Convert the entire pdf contents in to buffered stream and copy the texts into another file also extract the images and store it into the local disk. Preprocess the file with buffered texts by comparing it with stop words set and extract the non-stop words into another text file. Now count the frequency of each word in that text file, by now we get an idea on what the paper concentrates on. Now we take the top twenty words and apply stemming algorithm and remove the suffix and identify the synonymous words using WorldNet API and find the words eligible to become topics. Then we we compare the words respective to each topic in database and identify the final set of topics. We also include the keywords as topics. Now we have set of topics with us. Then we are splitting the pdf file line by line and check if the topic word occurs in the line and extract it. Then we are generating presentation slides using those topics and points.

Advantages of the system

We propose a system to automatically generate slides that have basic structure and content quality from academic papers using data mining techniques which has some important points in it and segregates all necessary figures, tables

and graphs from a technical paper and saves it in a local folder. This technique overcomes the problem of wrong text alignment and extracts the graphs, diagrams and important stats that can be used as assistance for creating final presentation.

Conclusion and Future Work

Engineering students and researches spend most of their time in getting an idea on academic papers of their interest and get ready for a presentation. In this system some Important points from the academic paper chosen, is put as points in draft slides with respective titles giving a basic idea on what the paper deals about. The images from the paper is stored in local disk which could be used while preparing the final presentation slides, thus providing an assistance to complete their task much easier. Being the first of its kind work, the system does partial summarization resulting in some important points and images saved in a folder, which needs to be further tuned to be used for final presentation. The future work would be on summarizing the texts as accurate as possible, so that it can be used as a final presentation slide without requiring any manual corrections later.

References

1. Sravanthi.M, Ravindranath Chowdry (2013).SlidesGen: Automatic Generation of Presentation Slides for a Technical Paper Using Summarization. AAAI Publications, Twenty-Second International FLAIRS Conference.
2. Tomohide Shibata, Sadao Kurohashi(2012). Automatic Slide Generation Based on Discourse Structure Analysis. Springer-Verlag Berlin Heidelberg.
3. Brandon Beamer, Roxana Girju(2012), Investigating Automatic Alignment Methods for Slide Generation from Academic Papers. Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL).
4. Sravanthi.M, Ravindranath Chowdry (2012). QueSTS: A Query Specific Text Summarization System. . AAAI Publications, Twenty-Second International FLAIRS Conference.
5. Yue Hu, Xiaojun Wan(2015). 'PPSGen- Learning Based-Presentation Slides Generation for academic papers.' IEEE Transactions on Knowledge and Data Engineering, Vol:27 Issue:4.