A Slide Generation Approach Via Content Categorization

V. Sivaranjani, G. Shanthi, A. Kumaresan, K. Vijayakumar

Department Of Computer Science Engineering, SKP Engineering College, Tiruvannamalai.

Received 28 January 2017; Accepted 22 March 2017; Available online 28 April 2017

ABSTRACT

The wide accessibility of web documents in electronic forms requires an automatic technique to label the documents with a predefined set of area, what is recognized as automatic Text Categorization (TC). Over the earlier period, it has been observed a large number of advanced machine learning algorithms to address this tricky task. Created presentation slides can be used as drafts to help the presenters prepare their formal slides in a more quickly way. A novel scheme called PPS Gen is projected to address this task. Documents are frequently represented by the "bag-of-words": that is to say, every word or phrase take place in papers once or more times is considered as a aspect. It innovative make use of the regression method to learn the importance scores of the sentences in an academic paper, and then utilize the integer linear programming (ILP) method to generate well-structured slides by selecting and aligning key phrases and sentences. Recommend a novel system called PPS Gen to generate presentation slides from academic papers. Prepare a sentence scoring model based on SVR and then classification each and ever word through the SVM and use the ILP method to align and extract key phrases and sentences for generating the slides. Experimental results prove that our method can make much better slides than traditional methods.

KEYWORDS: Phrases, Scoring model, Text mining, Classification method, Sentences

INTRODUCTION

Slides have been an effective and popular means of presentation of information. In several conferences and meetings, a presenter takes the assist of slides to present his work in a systematic way (pictorial). In modern years with the ease of use of many software tools like Microsoft PowerPoint, Open office Presenter etc., for unproblematic preparation of slides, their usage has increased enormously. But these tools facilitate only in the formatting of content (stylizing, bullet points etc), except not in preparing the content itself. A user has to begin from scratch and it is a time consuming task. In this effort, I recommend a tool that generates slides for the presentation with important points and all required figures, tables and graphs from a technological paper. As it is manifest, such kind of a tool keep time and reduces the effort by providing a basic presentation, which can be additional tuned/ upgraded as final presentation. Aim to automatically generate well-structured slides and provide such draft slides as a basis to lessen the presenter’s time and effort while preparing their final presentation slides.

A presentation with slides is so effective to pass information to people in any situations, such as an academic conference or business. Although some software’s, such as PowerPoint and Keynote, help us by means of making presentation slides, it is tranquil cumbersome to make them from scratch. These sentences may be included in some bullet points. Our method stab to generate draft slides of the distinctive kind mentioned above and facilitate people to prepare their final slides. Automatic slides generation for academic papers is a very challenging task. Current methods generally extort objects like sentences from the paper to

construct the slides. In contrast to the short summary extracted by a summarization system, the slides are required to be much more structured and much longer.

Slides can be divided into an ordered sequence of parts. Each part addresses a specific topic and these topics are also relevant to each other. Generally speaking, automatic slide generation is much more difficult than summarization. A slide has a title and contains some bulleted points which are important in that section.

Observing the similarity present between conference paper and human written slides for the paper, I address the problem of automatic generation of presentation slides by exploiting the structure of a conference paper. PPS Gen to generate well-structured presentation slides for academic credentials. In our system, the importance of each sentence in a paper is learned by using the support vector machine (SVM) model with a number of useful features, moreover then the presentation slides for the paper are generated by using the integer linear programming (ILP) model with ornately designed intention function and constraints to select and align key phrases and sentences.

II Related Work:

Automatically generate slides from input documents annotated with the GDA tag set. GDA tagging can be used to predetermined semantic structure. They first detect topics in the input documents and then extract important sentences appropriate to the topics to generate slides.

Introduced a support system for making slides from technical papers. The inputs of the system are academic papers in LATEX format. The system gauges the weights of the terms in the paper using TF-IDF scores. Using the term weights, objects in the paper like sentences, tables etc. are also prejudiced.

proposed a method to automatically generate slides from raw texts. Clauses and sentences are considered as discourse units and coherence relations between the units such as list, contrast, topic-chaining and cause are identified. These different parts are used to generate the final slides based on the detected discourse structure and some heuristic rules. The problem of aligning technical papers and presentation slides, used a variation of the Hidden Markov Model (HMM) to align the text in the slides to the most likely section in the paper, compared and evaluated four different alliance methods that were collective by methods such as TF-IDF term weighting and query expansion.

Proposed a system named automatic report to presentation (ARP) which constructs a topic-specific report and a presentation on a topic or search phrases given by a user. The system retrieves web pages relevant to the disambiguated query using multiple search engines.

III. Problem Definition:

Most existing filter approaches first calculate class dependent feature scores, i.e., the feature importance for each class is measured. One major disadvantage is that using the combination operation may bias the feature importance for discrimination. They 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. The query expansion does not improve performance in our application and that TF-IDF term weighting is inferior to a much simpler scoring mechanism based on the number of matched terms. TF-IDF term weighting is inferior to a simpler scoring mechanism based only on the number of matched terms and query expansion degrades aligner performance. Our best aligner achieves an accuracy of 75%.

IV. System Architecture:
V. Proposed System:

After getting the predicted importance score for each sentence in the given paper, we exploit the integer linear programming method to generate well-structured slides by selecting and aligning key phrases and sentences. Selecting important sentences and placing sentences on the slides, we select both key phrases and sentences to construct well-structured slides. We use key phrases as the bullet points and sentences relevant to the phrases are placed below the bullet points.

In categorize to obtain under duress the key phrases, chunking implemented by the Open NLP library is applied to the sentences and noun phrases are extracted as the candidate key phrases.

We define two kinds of phrases: global phrases and local phrases. Any unique phrase in an article is a global phrase, and a local phrase means a global phrase in a particular section. For example, “SVR” is a global phrase of this paper, while its appearances in different sections are considered unusual local phrases. So a global phrase that appears in different sections can correspond to a few local phrases. Since an important phrase is always used in many different sections, a global phrase that corresponds to more local phrases should be regarded to be more important and more likely to be selected. Use the local phrases to generate the bullet points directly for different sections and use the global phrases to address the importance differences between different unique phrases. All the phrases are stemmed and stop words are removed. Moreover, the noun phrases that appear only once in the paper are discarded.

A. Load Dataset and Preprocessing:

Input document which is wanted to make it as power point presentation than read the input document file and want to implement the preprocessing to that input file. Preprocessing is called as data cleaning which is going to use stop word removal method, this method read word by word from the input file and it will check with stop word dataset if the word is exist in stop word dataset than this method ignore that word, this method send non-stop words only to next process.

![Fig. 1: Load Dataset](image)

B. Frequent Mining and Similarity Clustering:

Getting 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. Mining in the sense, it have to extract files to large amount of datasets. Here find out maximum number of time repeated words in our file for analyzing & know, which are the topics explained in that paper.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are further related (in some sense or another) to each other than to those in other groups (clusters). From the maximum frequents word, find the weight age of the each and every word than from the weight age value to going to calculate the similarity between the words, based on the similarity it going to group the words as clusters.

![Fig. 2: Frequent Mining](image)
C. Topic Modeling and Sentence Extraction:

Topic modeling is a form of text mining, a method of identifying outline in a corpus. You capture your corpus and run it through a tool which sets words across the corpus into 'topics'. A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Without needing to ask, given that a document is about a scrupulous topic, one would wait for particular words to become visible in the document more or less frequently. Going to create the topics for the clusters, each cluster have n number of similar words using this words going to find the topic for that cluster with the help of lexical analysis.

Getting the input file and split the file into line by line here we are going to extract the lines using words from the cluster and we keep it as points for power point presentation. Sentence extraction is a technique used for automatic summarization of a passage. In this shallow approach, statistical heuristics are used to identify the most prominent sentences of a wording. Sentence extraction is an inexpensive approach compared to more knowledge-intensive deeper approaches which require additional knowledge bases such as ontology’s or linguistic. In short “sentence extraction” works as a filter which allows only important sentences to pass.

The major downside of applying sentence extraction techniques to the assignment of summarization is the thrashing of coherence in the ensuing summary. On the other hand, sentence extraction summaries can confer valuable clues to the main points of a document and are frequently satisfactorily intelligible to human readers.

D. Slide Creation:

Going to create the slides using the topics of cluster as titles of the slides and sentence as slide points. We train a sentence scoring model based on Support Vector Regression (SVR) and use the Integer linear Programming (ILP) method to align and extract key phrases and sentences for generating the slides. Experimental results show that our method can generate much better slides than traditional methods. We will also try to extract the slide skeletons from the human-written slides and apply these slide skeletons to the automatic generated slides.
VI. Conclusion And Future Work:

This paper proposes presentation slides from academic papers. Produce a sentence scoring model based on SVM and use the ILP method to align and extract key phrases and sentences for generating the slides. Experimental results show that our method can generate much better slides than traditional methods. Only consider one typical style of slides that beginners usually use. In the future, consider more complicated styles of slides such as styles that slides are not aligned sequentially with the paper and styles that slides have more hierarchies. It will also try to extract the slide skeletons from the human-written slides and apply these slide skeletons to the automatic generated slides. Furthermore, our system generates slides based on only one given paper. Additional information such as other relevant papers and the citation information can be used to improve the generated slides.

REFERENCES