Slide Creation Approach through Content Categorization

K.AMUTHA[1],C.SUGASINI[2], S.KRISHNANARAYANAN[3]
Department of Computer Science and Engineering, Arjun College of Technology, Coimbatore, Tamilnadu, India.

ABSTRACT

The wide availability of web documents in electronic forms requires an automatic technique to label the documents with a predefined set of topics, what is known as automatic Text Categorization (TC).Over the past decades, it has been witnessed a large number of advanced machine learning algorithms to address this challenging task. The generated presentation slides can be used as drafts to help the presenters prepare their formal slides in a quicker way. Documents are usually represented by the “bag-of-words”: namely, each word or phrase occurs in documents once or more times is considered as a feature. JFSC (joint feature selection and classification) learns both shared features and label specific features by considering pairwise label correlations, and builds the multilevel classifier on the learned low-dimensional data representations simultaneously. It 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. We train a sentence scoring model based on SVR 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.

I. INTRODUCTION

Slides have been an effective and popular means of presentation of information. In many conferences and meetings, a presenter takes the aid of slides to present his work in a systematic way (pictorial). In recent years with the availability of many software tools like Microsoft PowerPoint, Open office Presenter etc., for easy preparation of slides, their usage has increased tremendously. But these tools help only in the formatting of content (stylizing, bullet points etc), but not in preparing the content itself. A user has to start from scratch and it is a time consuming task. In this work, we propose 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/upgraded as final presentation. 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.

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 with making presentation slides, it is still cumbersome to make them from scratch.

Slides contain the summarized version of a technical report. They contain the vital points of the report arranged in a systematic way, including graphic elements like figures and tables for easy illustration of the idea. Given a document, “Automatic generation of presentation slides” becomes a nontrivial task because of challenges like segmental Copyright c 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved, summarizing content of each topic and aligning these topics to one or more slides and placing necessary graphical content like figures, graphs and tables in appropriate slides at appropriate locations.

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.

The slide rule, also known colloquially in the United States as a slapstick,[is a mechanical analog computer. The slide rule is used primarily for multiplication and division, and also for functions such as roots, logarithms and trigonometry, but is not normally used for addition or subtraction. Though similar in name and appearance to a standard ruler, the slide rule is not ordinarily used for measuring length or drawing straight lines.

Slide rules come in a diverse range of styles and generally appear in a linear or circular form with a standardized set of markings (scales) essential to performing mathematical computations. Slide rules manufactured for specialized fields such as aviation or finance typically feature additional scales that aid in calculations common to those fields.

Academic papers always have a similar structure. They generally contain several sections like abstract, introduction, related work, proposed method, experiments and conclusions. Although presentation slides can be written in various ways by different presenters, a presenter, especially a beginner, always aligns slides sequentially with the paper sections when preparing the slides. Each section is aligned to one or more slides and one slide usually has a title and several sentences. These sentences may be included in some bullet points. Our method attempts to generate draft slides of the typical type mentioned above and helps people to prepare their final slides.

Automatic slides generation for academic papers is a very challenging task. Current methods generally extract 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. Slides usually not only have text elements but also graph elements such as figures and tables. But our work focuses on the text elements only.

In this paper we concentrate on generating slides for research papers that are in accordance with standards of conference/journal proceedings. By and large, conference papers have an almost similar structure. They have an abstract and the sections present in them can be broadly classi- fied into presenting the introduction, the related work, actual work (model), the experiments, the conclusions and the bibliography. Most of the times, the presenter preserves the order of the paper in slides and each section is allotted one or more slides. 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, we address the problem of automatic generation of presentation slides by exploiting the structure of a conference paper. Here after we use terms “conference paper”, “technical paper”, “document” and “report” interchangeably.

In this study, we propose a novel system called PPSGen to generate well-structured presentation slides for academic papers. In our system, the importance of each sentence in a paper is learned by using the support vector regression (SVR) model with a number of useful features, and then the presentation slides for the paper are generated by using the integer linear programming (ILP) model with elaborately designed objective function and constraints to select and align key phrases and sentences.

Experiments on a test set of 200 paper-slides pairs indicate our method can generate slides with better quality than the baseline methods. Using the ROUGE toolkit and the pyramid evaluation, the slides generated by our method can get better ROUGE scores and pyramid scores. Moreover, based on a user study, our slides can get higher rating scores by human judges in both content and structure aspects. Therefore, our slides are considered a better basis for preparing the final slides.

II. RELATED WORK

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%.

DISADVANTAGES

1. To prepare the paper presentation they use much software such as Microsoft Power- Point and Open Office to help researchers prepare their slides.
2. 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.
3. A user has to start from scratch and it is a time consuming task.

III. PROPOSED WORK

The generated presentation slides can be used as drafts to help the presenters prepare their formal slides in a quicker way.It 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.We train a sentence scoring model based on SVR and use the ILP method to align and extract key phrases and sentences for generating the slides. Motivated by the core idea of linear discriminant analysis Experimental results show that our method can generate much better slides than traditional methods

ADVANTAGES

1. Each sentence in a paper is learned by using the support vector regression (SVR) model.
2. The presentation slides for the paper are generated by using the integer linear programming (ILP) model.
3. The slids are generated automatically from the academic papers.
4. The generated slides for the presentation will have only important points and all necessary figures, tables and graphs.

A. Load Dataset and Preprocessing:

In this module we want to load the input document which we want 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 to this we 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.

B. Frequent Mining:

In this module we get the non-stop words as input and calculate the count of words and find the repeated occurrence of each and every word from the non-stop words.

C. Similarity Clustering:

From the maximum frequents word we 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 we going to group the words as clusters.

D. Topic Modeling:

In this module we are going to create the topics for the clusters, each cluster have n number of similar words using this words we going to find the topic for that cluster with the help of lexical analysis.

E. Sentence Extraction:

Here we get 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.

F. Slide Creation:

Finally we are going to create the slides using the topics of cluster as titles of the slides and sentence as slide points.

DATA FLOW DIAGRAM TO CONVERT PDF TO PPT

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be the considered to be the most critical in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods

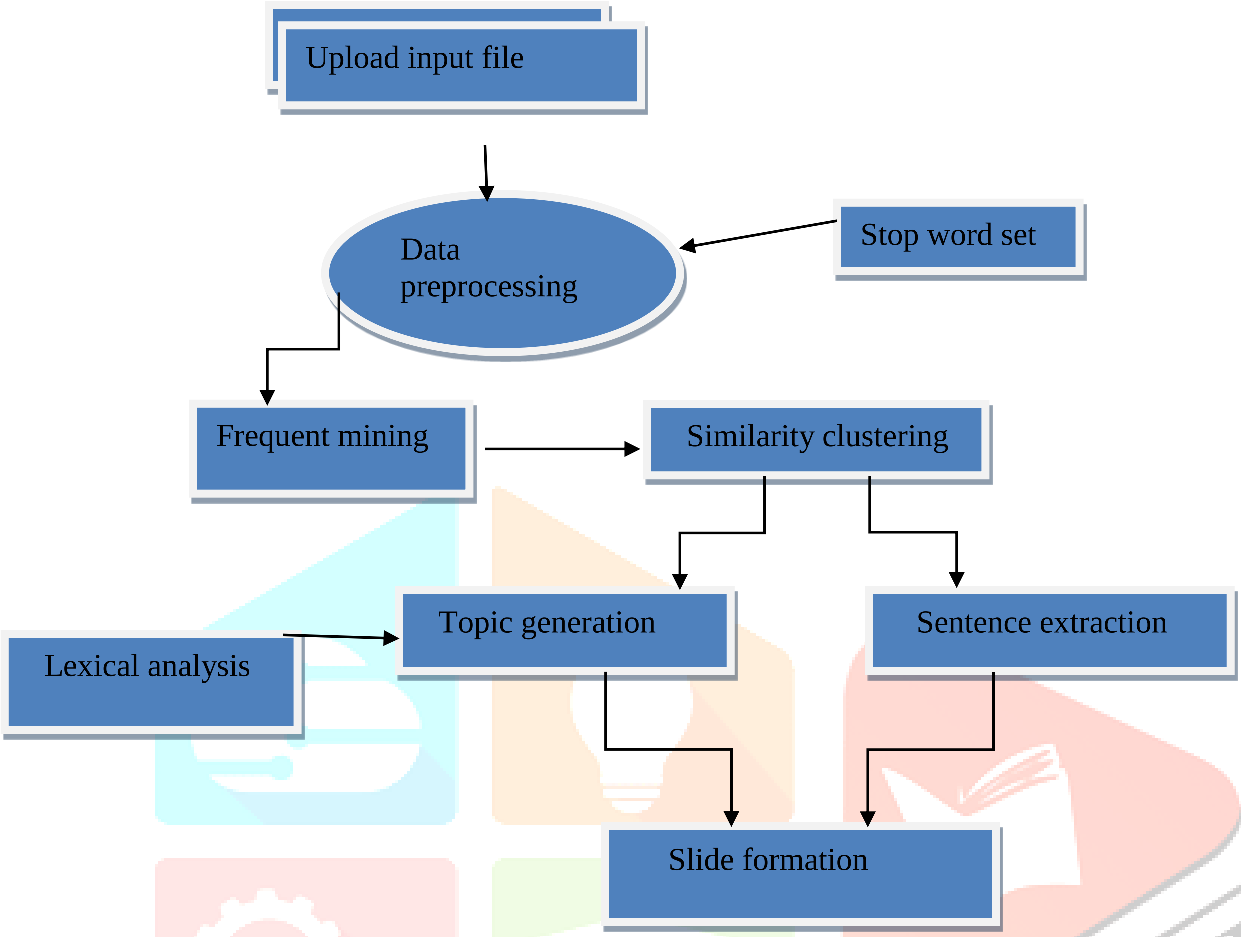

Figure 1 PDF TO PPT CONVERSEN

IV CONCLUSION AND FUTURE ENHANCEMENT

This paper proposes a novel system called PPS Gen to generate presentation slides from academic papers. We train a sentence scoring model based on SVR 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 in this paper, we only consider one typical style of slides that beginners usually use. In the future, we will 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.We 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. We will consider this issue in the future.

REFERENCE

1) M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," IEEE Trans. Knowl. Data Eng., vol. 18, no. 10, pp. 1338–1351, Oct. 2006.
2) Y. Xia et al., "Weakly supervised multilabel clustering and its applications in computer vision," IEEE Trans. Cybern., vol. 46, no. 12, pp. 3220–3232, Dec. 2016.
3) R. Hong et al., "Image annotation by multiple-instance learning with discriminative feature mapping and selection," IEEE Trans. Cybern., vol. 44, no. 5, pp. 669–680, May 2014.

4) F. Kang, R. Jin, and R. Sukthankar, “Correlated label propagation with application to multi-label learning,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., New York, NY, USA, 2006, pp. 1719–1726.

5) S. Pan, J. Wu, X. Zhu, G. Long, and C. Zhang, “Task sensitive feature exploration and learning for multitask graph classification,” IEEE Trans. Cybern., to be published. Aug. 2014.

6) J. Read, B. Pfahringer, and G. Holmes, “Multi-label classification using ensembles of pruned sets,” in Proc. IEEE Int. Conf. Data Min., Pisa, Italy, 2008, pp. 995–1000.

7) H. Liu, X. Li, and S. Zhang, “Learning instance correlation functions for multilabel classification,” IEEE Trans. Cybern., vol. 47, no. 2, pp. 499–510, Feb. 2017.

8) K. Nag and N. R. Pal, “A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification,” IEEE Trans. Cybern., vol. 46, no. 2, pp. 499–510, Feb. 2016.

9) M.-L. Zhang and L. Wu, “Lift: Multi-label learning with label-specific features,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 1, pp. 107–120, Jan. 2015.

10) E. Gibaja and S. Ventura, “Multi-label learning: A review of the state of the art and ongoing research,” Wiley Interdisc. Rev. Data Min. Knowl Disc., vol. 4, no. 6, pp. 411–444, 2014.