International Journal of Engineering Research and Modern Education (IJERME)
ISSN (Online): 2455 - 4200
(www.rdmodernresearch.com) Volume I, Issue I, 2016

EXTRACTING MULTI-DOCUMENT SUMMARIZATION WITH INTEGER LINEAR PROGRAMMING

N. Vennila* & M. Rajakumar**

* PG Scholar, Department of Master of Computer Applications, Dhanalakshmi Srinivasan Engineering College, Perambalur, Tamilnadu

** Associate Professor & Head, Department of Master of Computer Applications, Dhanalakshmi Srinivasan Engineering College, Perambalur, Tamilnadu

Abstract:

"Extracting multi document summarization with Integer Linear Programming" is used create an automatic slide generation summary for slides using text. On other hand it also generates well structured slides by selecting and aligning the key phrases and sentences. The Integer linear programming (ILP) method is used to align and extract key phrases and sentences for generating the slides. The ILP method includes the text but also graph, figures and tables in the slides but this project focus only on the text and key phrases. The Support Vector Regression (SVR) method is used to find the similarity and over all similarity with the sentences present in the slides.

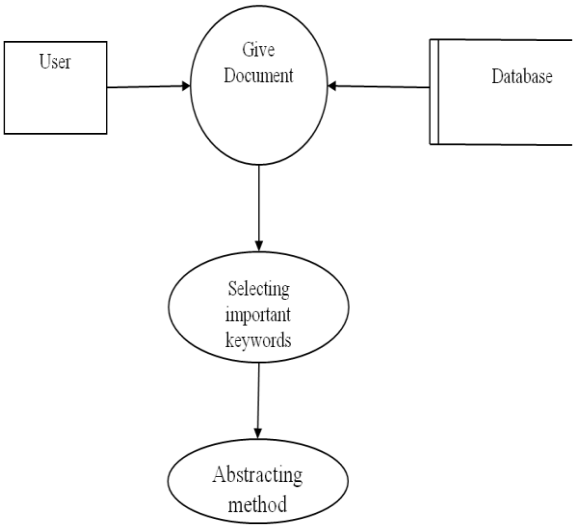
Index Terms: Text Summarization, Integer Linear Programming & Support Vector Regression

1. Introduction:

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 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 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.

Data Flow Diagram:

DFD Level 0:



DFD Level 1:

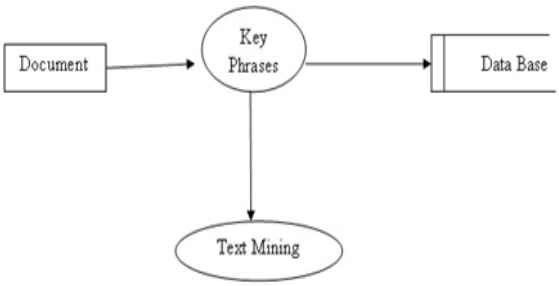


Figure 1: Data Flow Diagram

2. Related Works:

The work of constructs summaries by maximizing the importance of the selected sentences and minimizing their pair wise similarity.

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.

3. Proposed Work:

In this work propose a method of automatically generating presentation slides for academic papers. 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. 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.

4. Experimental Analysis and Results:

Various life cycle processes such as requirement analysis, design phase, verification, testing and finally followed by the implementation phase result in a successful project management. System implementation is an important stage of theoretical design is turned into practical system.

Similarity with the Titles: We consider three types of titles: paper title, section titles and section titles. Only the titles of the section and section which contain the sentence are used. We use the cosine similarity values between the sentence and different types of titles as different features. Stop words are removed and all the words are stemmed in

the similarity calculation. Intuitively the sentences that have higher similarity with the titles should be more likely to be selected.

Word Overlap with the Titles: It is the number of words shared by the sentence and the set of words of all titles, including all three types of titles mentioned above.

Sentence's Parse Tree Information: The features are extracted from the sentence' sparse tree. It includes the number of noun phrases and verb phrases, the number of sub-sentences and the depth of the parse tree.

Stop Words Percentage: It is the percentage of the stop words in the total word set of the sentences. Intuitively the sentences that have high percentage of the stop words are less likely to be important.

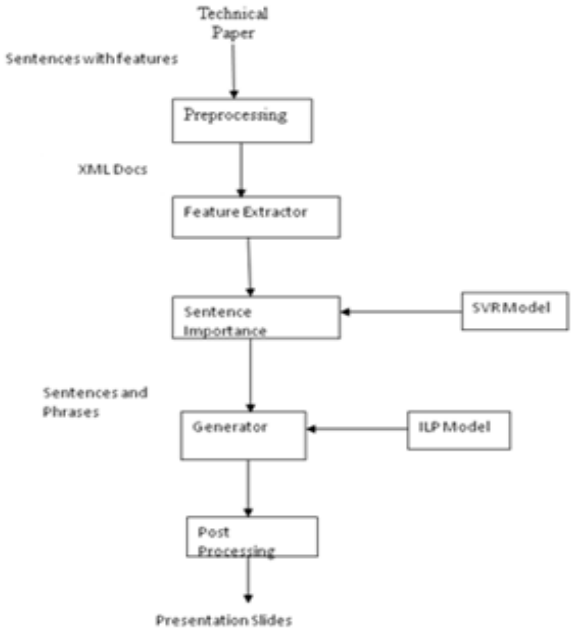
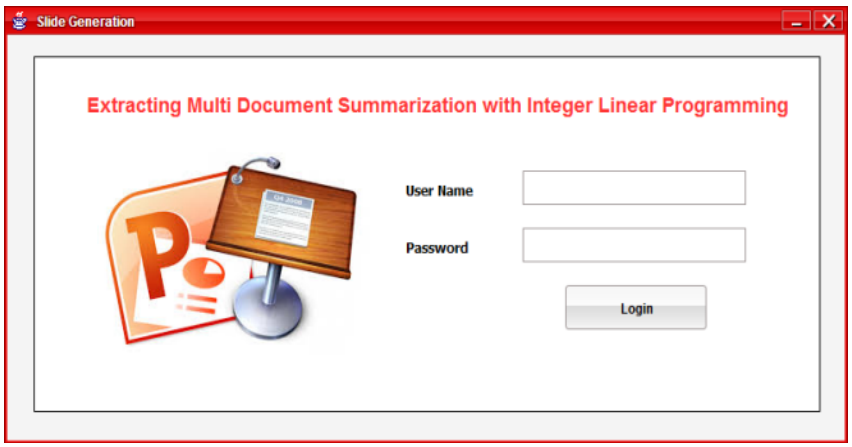
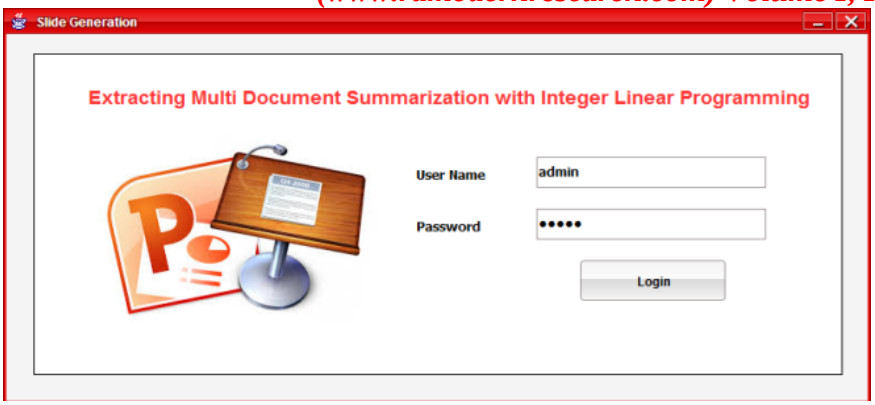


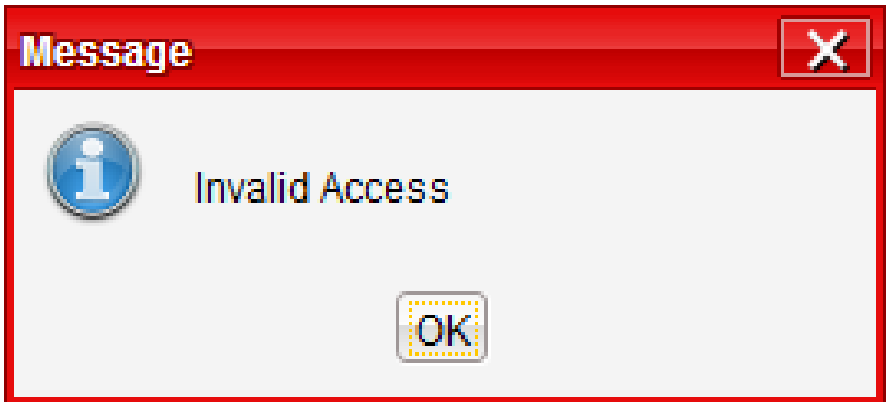
Figure 2: Sentiment Analysis Reviews Based on Commands

Other features including the length of sentences, the numbers of words after removing stop words and the average length of sentences of the section, section or paragraph that contains the sentence. All the features mentioned above are scaled into [−1, 1]. Based on the features and importance scores of the sentences in the training data, we can learn an SVR model, and then apply the model to predict an importance score for each sentence in any paper in the test set. The score indicates the possibility of a sentence to be selected for making slides.

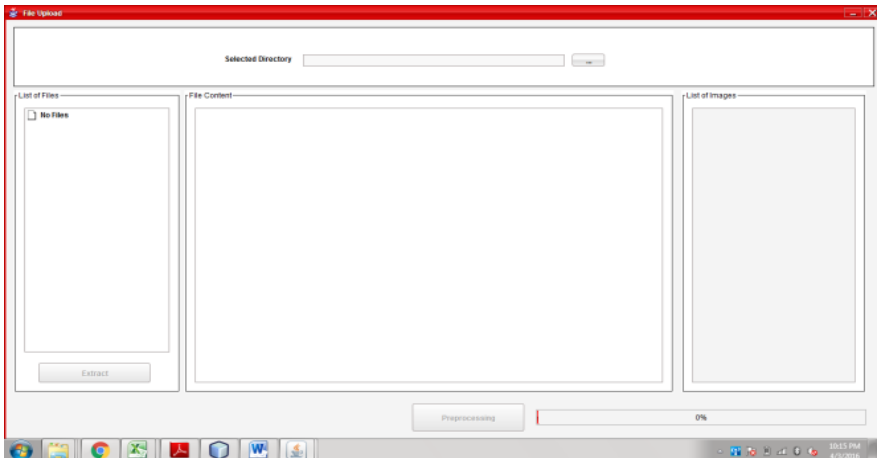
User Login:

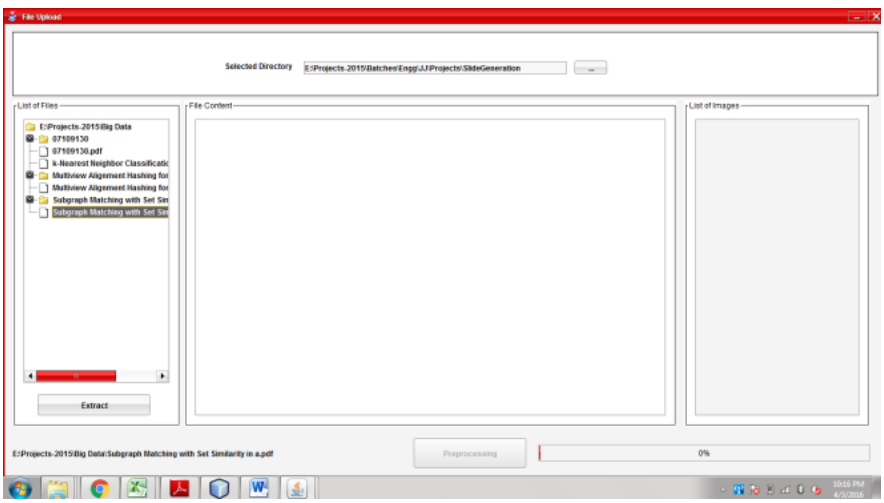



Slide Generation
Extracting Multi Document Summarization with Integer Linear Programming
User Name
admin
Password
Login

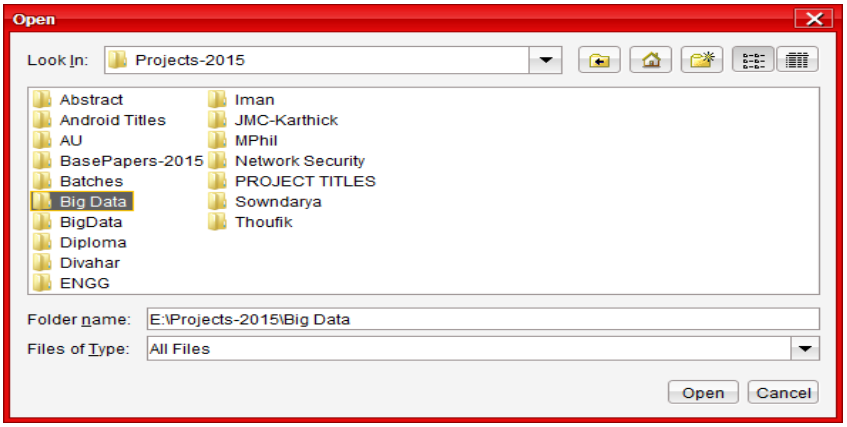

Message
Invalid Access
OK

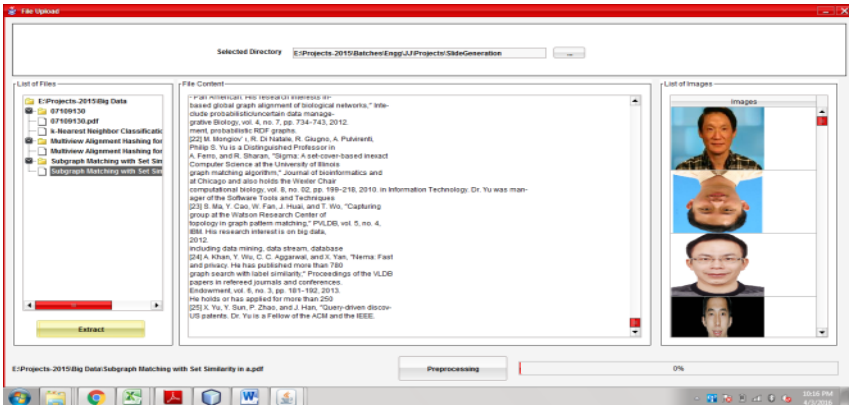
File Upload:


File Upload
Selected Directory
List of Files
No Files
File Content
List of Images
Extract
Preprocessing
0%

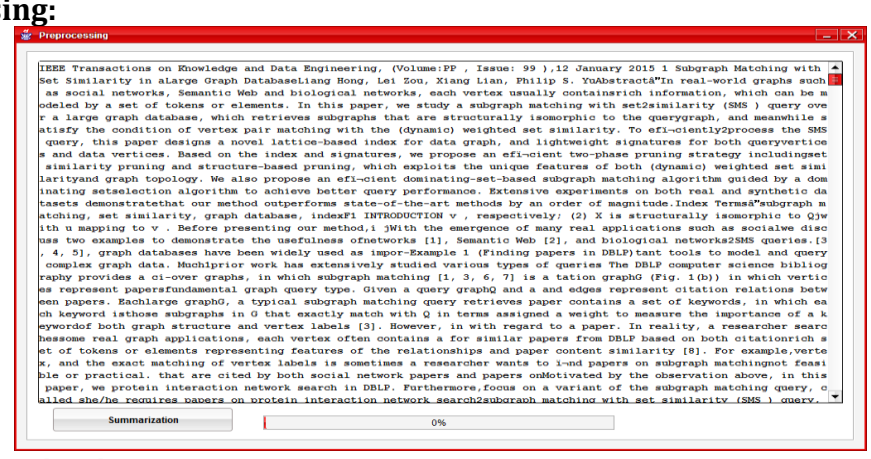

File Upload
Selected Directory
E:\Projects-2015\Batches\Engg\JJ\Projects\SlideGeneration
List of Files
File Content
List of Images
Extract
Preprocessing
0%

File Open:

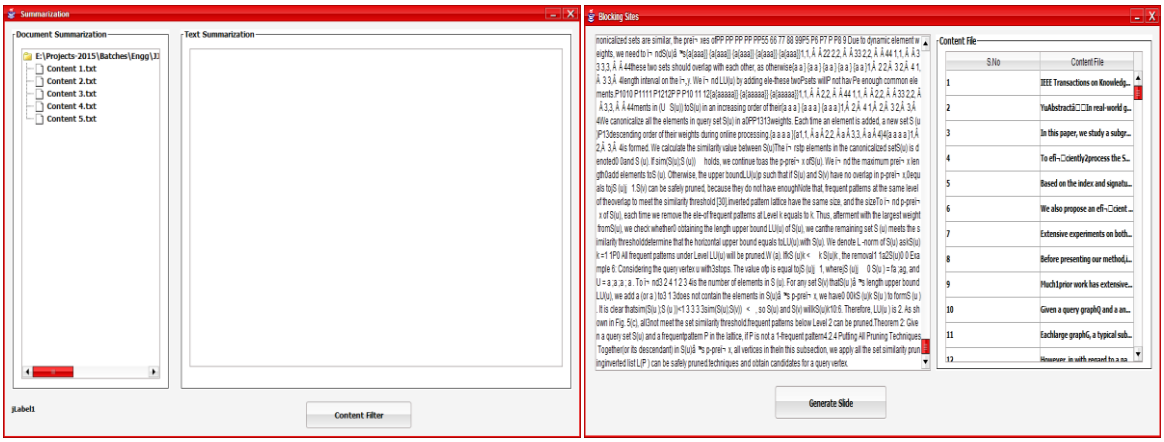




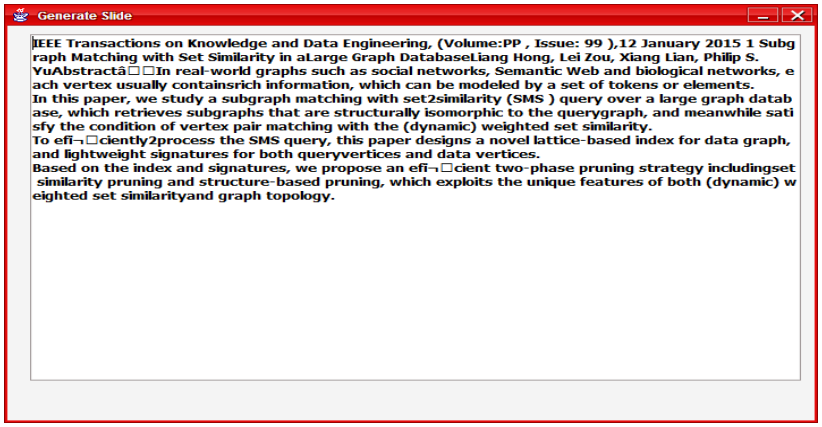
Preprocessing:



Text Summarization:



Generate Slide:



5. Conclusion:

In this project trains that a sentence scoring model based on SVR and use the ILP method to align and extract key phrases and sentences for generating the slides. The relationship between the text elements and the graphical elements also needs to be identified. The presenter need to know which sentences are most relevant to a graphical element and which graphical elements should be selected to generate the slides. The presenter can use rule-based methods or machine learning based methods to solve the above problems. Then can simply attach the tables and figures we select to the most relevant sentences in the slides. Additional information such as other relevant papers and the citation information can be used to improve the generated slides. By considering more complicated styles of slides such as styles that slides are not aligned sequentially with the paper and styles that slides have more hierarchies.

6. Future Enhancement:

In this work propose a method of automatically generating presentation slides. 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. 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.

7. References:

1. Beamer .B and Girju .R, (2009) 'Investigating automatic alignment methods for slide generation from academic papers', in Proc. 13th Conf. Comput. Natural Lang. Learn., Jun., pp. 111–119.
2. Hayama .T, Nanba .H and Kunifuji .S, (2005) 'Alignment between a technical paper and presentation sheets using hidden Markov model', in Proc. Int. Conf. Active Media Technol., pp. 102–106.
3. Kan .M.Y, (2006) 'Slide Seer: A digital library of aligned document and presentation pairs', in Proc. 7th ACM/IEEE-CS Joint Conf. Digit. Libraries, Jun., pp. 81–90.
4. Masum .M.A, Ishizuka .M, and Islam .M.T, (2005) 'Auto-presentation: A multi-agent system for building automatic multi-modal presentation of a topic from World Wide Web information', in Proc. IEEE/ WIC/ACM Int. Conf. Intell Agent Technol., pp. 246–249.