Scientific Journal of Impact Factor(SJIF): 4.14

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Special Issue for ICPCDECT 2016,Volume 3 Issue 1

Auto Generation Of Presentation Slides From Text

Sandhya Budar [1], Durgadevi Jadhav [2], Snehal Malawadkar[3], Akshata Shinde [4]

[1]*Computer Engineering, All India Shri Shivaji Memorial Society's Institute Of Information Technology,Pune,*

Abstract- *In this paper, we have come up with a system that will automatically generate presentation slides from text.Presentation slides are widely used to communicate information to the audience. Presentation slides are prominently used in I.T. sector, business meetings ,colleges. The slides mostly comprise of only the important points related to the topic. They are a powerful means of presenting a topic to the audience, as the important points also known as bullet points are covered in the slides and are explained by the presenter in depth. There are various tools available in the market which only deal with formatting of the slides but not the content. Our system aims at automatically generating presentation slides from text. This will eventually help in reducing a great amount of the presenter's time and efforts. The proposed system works on NLP rules to classify data for the desired slides.*

Keywords- *Pre-processing, Tokenization, Feature Extraction, Fuzzy Classification, Crisp values*

I. INTRODUCTION

Presentation slides are used to present a topic or a new concept before the audience. They are used because they are efficient and also easy to understand. Presentation slides are widely used in the corporate world, I.T. industries, schools, colleges. They can be used to address a large audience. They are especially used when a new product is to be launched in the market. The essential features of the product are highlighted by using presentation slides. Various tools are available in the market which tackle only with the formatting of the slides. The content has to be inserted by the user himself. For this the user has to study the topic throughly and also spend a lot of his time and efforts. This problem is taken care of by our system. Our system is efficient because it generates presentation slides automatically when text is given as input to the system. Important keyphrases and points related to the topic are extracted and displayed on the slides.

II. LITERATURE REVIEW

Yue Hu and Xiaojun Wan [1] used SVR based sentence scoring model to assign important scores to each sentence and ILP model to generate structured slides from academic papers.
R. Jha, A. Abu-Jbara, and D. Radev [2] investigate the problem of automatic generation of scientific surveys starting from keywords which are provided by the user. The system takes a topic query as input and generates a survey of the topic. It selects a set of relevant documents and then selects relevant sentences from the documents to generate the survey. Content models namely Centroid , Lexrank , C-Lexrank are used.
A. Abu-Jbara and D. Radev [3] proposed an approach which focuses on coherence and readability aspects of the problem. It produces citation-based summaries in three stages: pre-processing, extraction and post processing. These summaries are better than several summarization systems.
M. Sravanthi, C. R. Chowdary, and P. S. Kumar [4] concentrate on generating slides from research papers. Latex documents which have rich structure and semantic information are given as input to the system. The documents are initially in XML format. The XML file is first parsed and then information in it is extracted . QueSTS Summarizer,a query specific extractive summarizer is used to generate slides. All graphical elements are placed at appropriate locations in the slides.
M.Y. Kan [5] present their work on SlideSeer , a customized digital library which comprises of an offline discovery, alignment and indexing system and an online web user interface. Three major system components of SlideSeer are discussed namely 1) resource discovery, 2) fine-grained alignment , 3) the user interface.
M. Utiyama and K. Hasida [6] in their paper discuss how to automatically generate slides shows. The system inputs documents which are annotated with the GDA tagset and XML tagset. These tagsets allow machines to automatically infer the semantic structure underlying the raw documents.
M. Sravanthi, C. R. Chowdary, and P. S. Kumar [7] present a system called QueSTS , which does effective extraction of query relevant information which is present within documents on the web. This is done by filtering aggregating important query relevant sentences which are distributed across documents.

III. PROPOSED SYSTEM

It has always been a wonder how presentation slides are generated automatically from text. Many existing systems are yielding low semantics. Proposed system works on NLP rules to classify the data for desired slides. When text is given as input to the system it goes through the following stages before generating presentation slides.

Reading:
Text is given as input to the system. It can be in pdf or doc format. The text is read in string format.

3.1. Preprocessing

The input is preprocessed by the following methods. Preprocessing reduces overhead and increases processing speed.

3.1.1. Remove stopwords. Stopwords are those words, which when removed will not alter the desired meaning of the sentence. Hence stopwords are removed in order to increase the processing speed. Duplicate words are identified and removed from the sentence.
e.g.: India is a country of rich heritage. Stopwords in this sentence are : is, a, of.

3.1.2. Stemming. In stemming process a word is brought to its base form. By doing this overhead is reduced and accuracy is increased. e.g.: Engineering will be reduced to engineer.

3.1.3. Tokenization. In this process words are trimmed, spaces are removed, tokens are generated and are then put in an array.

3.2. Feature Extraction

In this stage important features are extracted by the following means.

3.2.1. Numerical data. The sentence that holds numerical data is important and it is most probably included in the document summary. The numerical score of each sentence is calculated. This score is obtained by calculating the number of numbers occurring in a sentence. Depending on the score it is decided whether to include the sentence or not.

3.2.2. Term weight. It is the frequency of the term incidences within a document which has been used for calculating the rank of the sentence. The score of a sentence can be intended as the sum of the score of words in the sentence. Term weight is the number of times a particular word has occurred in a sentence. tf_isf (Term frequency, Inverse sentence frequency) method is applied to calculate the score of the term. If the score is high then that term is considered to be important.

3.2.3. Proper noun. The sentence that holds maximum number of proper nouns (name entity) is an essential sentence and is most likely to be included in the document summary. The score for this feature is the number of proper nouns occurring in a sentence over the length of the sentence.

3.2.4. Sentence to Sentence similarity. This feature finds the similarity between sentences. For each sentence *S*, the similarity between *S* and other sentences is computed by the cosine similarity measure with a value resulting between 0 and 1.

3.3. Fuzzy Classification

3.3.1. Fuzzification. Classification is done into 5 types:
Very low-0
Low
Medium
High
Very high-1
Generated scores of the sentences are checked according to the above classification. A score is termed as Very low if it has a score 0 and is termed as Very high if it has a score 1. Hence if the score is 0, the sentence is less important and if the score is 1 then the sentence is important. Thus, importance of a sentence can be obtained.

3.3.2. CRISP values. These are exact values that are obtained after Fuzzification.

3.3.3. Fuzzy inference engine. It is used to extract correct conclusions from approximate data. Rules are written to identify titles for each slide.

3.3.4. If-then rules. Fuzzy If-Then or fuzzy conditional statements are expressions of the form "If A Then B", where A and B are labels of fuzzy sets characterized by appropriate membership functions. The generated scores are checked with the if-then condition. If the score is very high then the sentence is important. If the score is very low then the sentence is not important.

3.4. Slide Generation

APACHE API is used for generating the slides at the end.

IV. BLOCK DIAGRAM

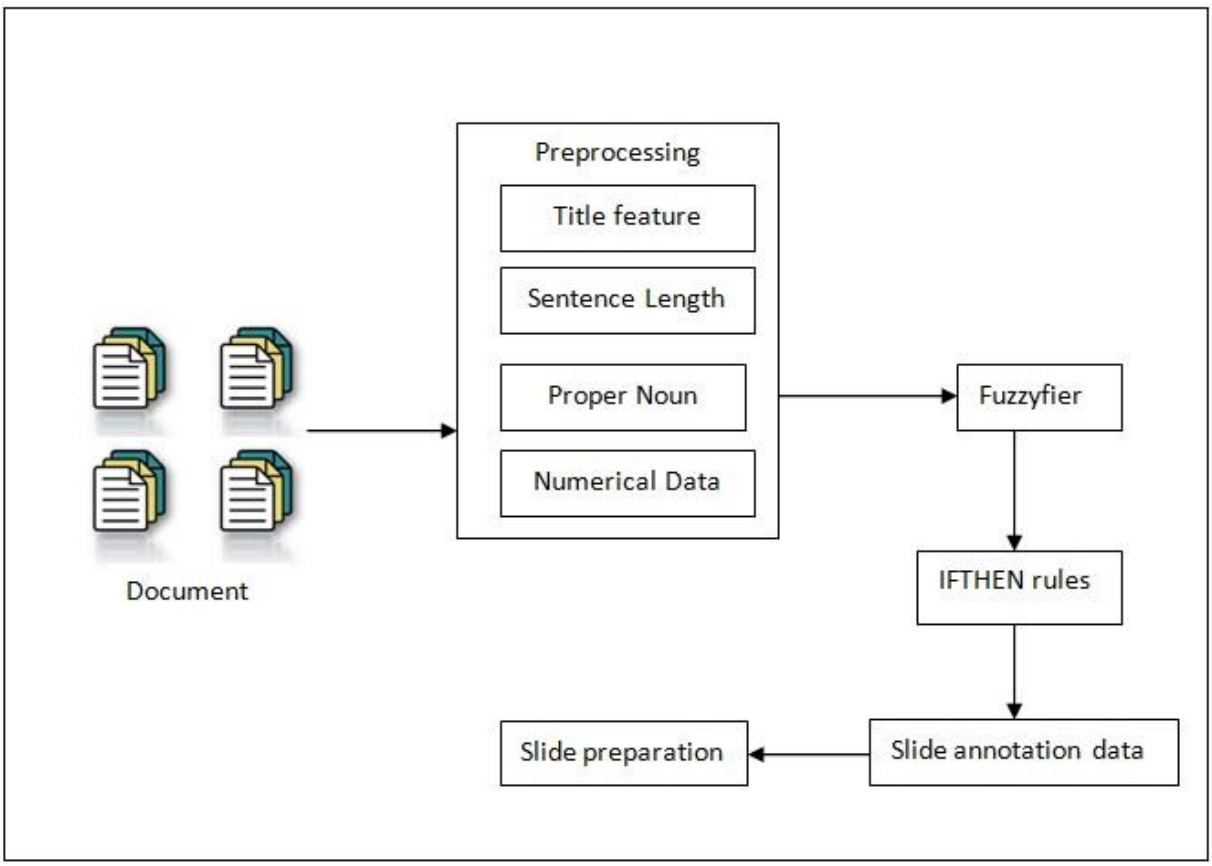

Figure 1. Block Diagram

V. ALGORITHM

5.1. Algorithm for Preprocessing

Step 0: Start
Step 1: Get contents of Query
Step 2: split in Words
Step 3: Remove Special Symbols
Step 4: Identify Stopwords
Step 5: Remove Stopwords
Step 6: Identify Stemming Substring
Step 7: Replace Substring to desire String

Step 8: Concatenate Strings
Step 9: Preprocessed String
Step 10: Stop

5.1.1. Algorithm to find stop words

Step 0: Start
Step 1: Read string
Step 2: divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: for i=0 to N (Where N is length of V)
Step 5: for i word of N check for its frequency
Step 6: Add frequency in List Called L
Step 7: end of for
Step 8: return L
Step 9: stop

5.2. Feature Extraction

5.2.1. Algorithm to find noun

Step 0: Start
Step 1: Read string
Step 2: divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: for i=0 to N (Where N is length of V)
Step 5: for i word of N check for its occurrence in Dictionary
Step 6: if present then return true
Step 7: else return false
Step 8: stop

Formulae used in Feature Extraction

1) Title length feature = Number of title words in S(Sentence) / No of words in sentence
2) Sentence length = Number of words in sentence / Number of words in longest sentence
3) Term weight = Number of times term appeared in the document

VI. ADVANTAGES

1. Well structured slides are generated.
2. Presenter's time and efforts are saved to a great extent.
3. Slides will include important key phrases and sentences related to them.

VII. CONCLUSION

It is a tedious job to create presentation slides. Thus our system will save a huge amount of the user's time and efforts. Presentation slides are generated in an efficient and quicker way after using the above methods.

VIII. FUTURE SCOPE

The system can be enhanced to work on all cross platforms.

REFERENCES

[1] Yue Hu and Xiaojun Wan,"PPSGen: Learning-Based Presentation Slides Generation for Academic Papers", ieee Transactions on knowledge and data engineering, vol. 27, no. 4, april 2015.
[2] R. Jha, A. Abu-Jbara, and D. Radev, "A system for summarizing scientific topics starting from keywords," ACM Comput. Surv,vol. 40, no. 3, p. 8, 2013.

[3] A. Abu-Jbara and D. Radev, "Coherent citation-based summarization of scientific papers," in Proc. 49th Annu. Meeting Assoc.Comput.Linguistics: Human Lang. Technol.-Volume 1, 2011,pp. 500–509.
[4] M. Sravanthi, C. R. Chowdary, and P. S. Kumar, "SlidesGen:Automatic generation of presentation slides for a technical paper using summarization," in Proc. 22nd Int. Flairs Conf.,2009, pp. 284–289.
[5] M.Y. Kan, "SlideSeer: A digital library of aligned document and presentation pairs," in Proc.7th acm/ieee-cs Joint Conf. Digit.Libraries, Jun. 2006, pp. 81–90.
[6] M. Utiyama and K. Hasida, "Automatic slide presentation from semantically annotated documents," in Proc. ACL Workshop Conf.Its Appl., 1999, pp. 25–30.
[7] M. Sravanthi, C. R. Chowdary, and P. S. Kumar, "QueSTS: A query specific text summarization approach, in Proc. 21st Int.Flairs Conf., 2008, pp. 219–224.

