IMPROVING PPT GENERATION TECHNIQUE

[1]Rohan Chaurasia, [2]Somnath Malge, [3]Ketan Sawalkar, [4]Sumit Raut, [5] S.R. Patil
[1]BE Student, [2]BE Student, [3]BE Student, [4]BE Student, [5]Asst. Professor
[1]Department of Computer Engineering, [2]Department of Computer Engineering, [3]Department of Computer Engineering, [4]Department of Computer Engineering, [5]Department of Computer Engineering
Sinhgad Institute of Technology, Lonavala, Pune, India

Abstract— The most common occurring problem while making PPT is time consumption as it takes time to extract important points from documents. In this study, we developed Automatic Generation of PPT. The different methodologies for Automatic PPT Generation are Fuzzy Logic, Text Mining and Natural Language Processing. Most of the methodologies in Automatic PPT Generation are having some performance issues regarding accuracy of the system, to enhance this paper purpose a novel idea of Automatic PPT Generation. The basic idea of Automatic PPT Generation system comes from the fact that due to high increase of the digital data, in depth study of the same always takes more time than of estimation. So it is hard task to gather proper points to generate PPT slides. So proposed system put forwards an idea of Automatic PPT Generation system using feature extraction by applying strong NLP protocols and then these features are classified by using fuzzy logic to get the best PPT slide points out of the given document.

Index Terms— Preprocessing, Fuzzy Logic, Natural Language Processing, Text Mining.

INTRODUCTION

Presentation slides is one of the most effective method to deliver key message to the audience Microsoft power point and open office are used for formatting slides but not context extracting key points from documents is time consuming process. This paper presents on effective way to generate presentation slides with the context. The key point can be extracted from documents using different data mining algorithms which reduces the time consumption of individuals.

Pre-processing converts raw data into useful information. Real world data has many errors such as incomplete, inconsistent and lacking in certain behavior. Pre-processing is used to remove such errors. The working of pre-processing can be explained by using four key points. Data cleaning, data integration, data transformation, data reduction. Data cleaning process detects and removes the errors and inconsistencies and it helps to improve the quality of data. Data quality problems arise due to misspellings during data entry, missing values or any other invalid data. Data integration is prepared by collecting data from multiple sources like multiple databases, data cubes or flat files. Operational databases keep changing with the requirements. A data warehouse integrating data from multiple sources faces the problem of inconsistency. To deal with these inconsistent data, transformation process can be applied. It consist of activities like smoothing, aggregation,generalization, normalization. Data reduction reduces the number of attributes like data cube aggregation, removing irrelevant attributes, numeric attributes. Data reduction includes techniques like data cube aggregation, dimensionality reduction, data compression and numerosity reduction. The purpose of pre-processing is it converts rough information into understandable form. It makes understanding the complex information easy.

Feature extraction is the process where textual features are identified with their score. Feature identification protocol is used. Feature is extracted and stored in a vector. Feature extraction consists of four key points such as numerical data, term weight, sentence to sentence similarity, proper noun. 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 number occurring in a sentence. Depending on the score it is decided whether to include the sentence or not. 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 for feature extraction. 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. 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. 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.

Fuzzy logic is a technique which is based on "degrees of truth" instead of "true or false" (1 and 0). It gives us accurate number between 1 and 0. Fuzzy 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. These are exact values that are obtained after fuzzification is called Crisp values. It is used to extract correct conclusions from approximate data. Rules are written to identify titles for each slide. 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.

"In this paper, section-2 is dedicated for background work and section-3 is elaborates the proposed techniques in depth. The evolution of system is carried out in section-4 and finally this paper is concluded with feature announcement tracers in section-5".

LITERATURE SURVEY

[2]Introduces the concept of providing pre-processing strategies on data before applying Sparse Auto-Encoders (SA), which make the earlier strategy suitable for all kind of datasets, be it large, complex or varied. Also, use of these strategies have improved the classification accuracy by up to 3% and reduced the rule base over 40 times. Strategies varies for real attribute data and categorical attribute data. In real data, the pre-processing can be done for all datasets but the number of features after pre-processing would increase. Whereas using K-modes clustering method all nominal type feature for pre-processed for categorical attribute data. Hence improving performance due to both pre-

processing and supervised learning of SA's. Perception for data pre-processing produces is needed because learning of SA's is dependent on data quality and optimization procedures and helps in increasing the performance of SA's.

[3]Proposes several of technologies of data pre-processing. Techniques related to too much data, data filtering and data elimination throw away data, whereas in data sampling important data sets are selected. Also noise modelling and principal component analysis help in summarizing or comprising the data. Each technique used have several strength and weaknesses. The limitations are mainly due to quality and completeness of data. Data filtering, data ordering, data editing and noise modelling are used to transform data from various domain. When all data characteristics are not known limited and uncompleted results are obtained. Which can be done using data visualization, data elimination and data selection. There are drawbacks because data from real world is never perfect. If proper techniques is not selected it can lead to loss or change of information. Certain data elimination or selection needs to be done iteratively for best results. Regarding the future scope of these techniques, many efforts are being made to analyze data using commercially available tools or by developing an analysis tool that meets the requirements of a particular application.

[4]This paper presented an analytical model for feature extraction in a massively parallel distributed environment. It also represented a framework that combines the speedup from both parallel and pipelined execution in a model. This model represents two components such as Analytical component and Empirical component. Feature extraction algorithms are developed in HDDI. This model is suitable for web based mega cluster and large document sets. This model is developed for parallel distributed environment hence it's not suitable for serial environment.

[5]This paper proposes new technique such as feature selection. Feature selection is a new stage before feature extraction. There is new set of features are generated into feature selection stage. The methods like Document frequency, Information gain, Gini index, Chi-square statistics, Best term, Ambiguity measure & Distinguishing feature selector are involved in feature selection stage. Feature extraction stage is extract a new features from feature set for given text documents. Such feature selection and feature extraction techniques can be used as a pre-processing stage for text classification to handle large amount of text documents. It is linear model hence it is not solution for non-linear models and it is not reliable for low frequency terms.

[6]In this method predicting state of health battery system has been developed. This method use for to analyze data obtained by spectroscopy and coulomb counting techniques. Fuzzy logic provides a powerful modelling complex, non-linear system. The major problem in medical science involves the diagnosis of disease, based upon various tests performed upon the patient. In data mining application, medical tests and patient information are stored in predictive manner. The patient received only those tests that are allows us. A data mining application call Data comber was built using the union rule configuration (URC). A mathematical construct that allows this program to process very large data sets in linear time, work simultaneously with hundreds or thousands of predictive elements without resorting to pruning techniques, and incorporate records with any number of missing values without requiring missing-value interpolation or the elimination of sparsely populated records.

[7]In this method fuzzy logic has mathematical and physical science to visible and more substantive. Fuzzy logic is the main methodology for computing with words. The point of this note is that fuzzy logic play an important role in computing word and vice versa. Computing involves number and symbol, fuzzy logic equal to computing words. The mean disadvantage of fuzzy logic is that given a collection of proposition expressed in natural language with the initial data set. There are two major imperatives for computing with words. When there is a tolerance for imprecision which can be exploited to achieve robustness and low solution cost.

[8]Proposes a method by Paul Dickson, W. Richards Adrian and Allen Hanson that captures and save images to create accurate visual record of computer based presentations. System uses visual inspection techniques to scan the screen capture stream to identify points to save. System requires that instructor should connect a pc to sampling device as they would to a projector.

[9]Describes a method by Yuting Su, Zhong Ji, Xingguang Song, Rui Hua that presents a unique caption text location method. Caption text contains useful data for video annotation, indexing and searching. The window is scanned over the key frame and then texture and edge feature are extracted which are input to SVM classifier and finally vote mechanism and morphological filter are performed to locate the caption region. This method has a drawback that it cannot localize the no horizontally aligned caption text.

PROPOSED SYUSTEM

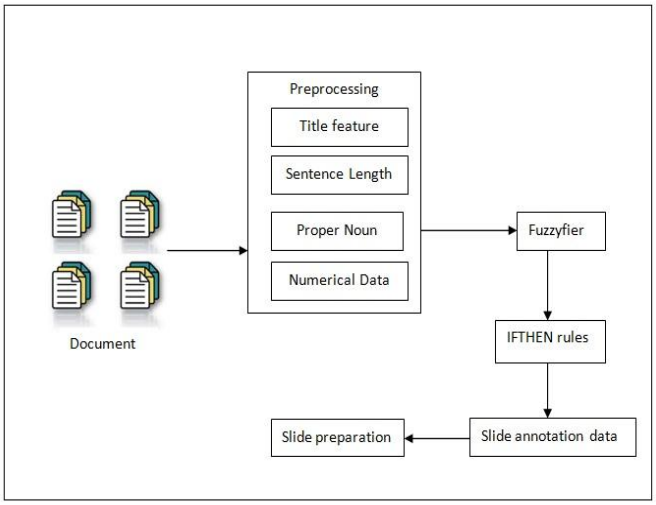

Figure 1: **Proposed System Workflow**

It has dependably been a think about how introduction slides are produced consequently from content. Many existing frameworks are yielding low semantics. Proposed framework takes a shot at NLP principles to characterize the information for coveted slides. At the point when content is given as contribution to the framework it experiences the accompanying stages before creating introduction slides. Proposed System implements phase based approach for automated slide generation.

Phase 1: Input to system: PDF, DOC file is been given as input to system and complete data is been retrieved in string format.
Phase 2: Preprocessing: Input String data is been sent to preprocess phase for data cleansing and Data filtration process this process consist of three sub process. Data cleansing stop word and special symbol removal. Data alteration i.e. stemming brings all terms to base form. Stopwords are those words, which when expelled won't modify the coveted importance of the sentence. Consequently stopwords are evacuated with a specific end goal to expand the handling speed. Copy words are recognized and expelled from the sentence.
Phase 3: **Feature Data Extraction**: feature data is vital data **Numerical information** sentence that holds numerical information is critical and it is most presumably incorporated into the archive synopsis. The numerical score of each sentence is figured. This score is acquired by ascertaining the number of numbers happening in a sentence. Contingent upon the score it is chosen whether to incorporate the sentence or no. **Term weight** is the recurrence of the term rates inside a record which has been utilized for computing the rank of the sentence. The score of a sentence can be proposed as the whole of the score of words in the sentence. Term weight is the quantity of times a specific word has happened in a sentence. tf_isf (Term frequency, Inverse sentence recurrence) strategy is connected to ascertain the score of the term. If the score is high then that term is thought to be essential. Formal person, place or thing. The sentence that holds greatest number of formal people, places or things (name substance) is a fundamental sentence and is well on the way to be incorporated into the report synopsis. The score for this component is the quantity of formal people, places or things happening in a sentence over the length of the sentence. Sentence to **Sentence similarity** This feature finds the similarity between sentences. For each sentence S, the similarity between S and other sentence sis computed by the cosine similarity measure with a value resulting between 0 and 1.
Fuzzy Classification: classification theory that divides data into five parts.

- **Very low-0**
- **Low**
- **Medium**
- **High**
- **Very high-1**

Produced scores of the sentences are checked by the above grouping. A score is named as Very low in the event that it has a score 0 and is named as Very high on the off chance that it has a score 1. Consequently if the score is 0, the sentence is less imperative and if the score is 1 then the sentence is critical. Accordingly significance of a sentence can be acquired. **Fuzzy inference engine** is utilized to remove adjust conclusions from estimated information. Tenets are composed to recognize titles for each slide.
Fuzzy If Then or Fuzzy restrictive proclamations are articulations of the shape "If A Then B", where An and B are names of fluffy sets portrayed by proper enrollment capacities. The created scores are checked with the if-then condition. If the score is high then the sentence is vital. In the event that the score is low then the sentence is not critical

ALGORITHM to 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

Algorithm to find top 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

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

RESULTS AND DISCUSSIONS

The above figure 2 gives the normal MRR of 0.847 for various number of keeps running for the information content documents. This outcome can be say great in our first endeavor of PPT Generation utilizing NLP rules.

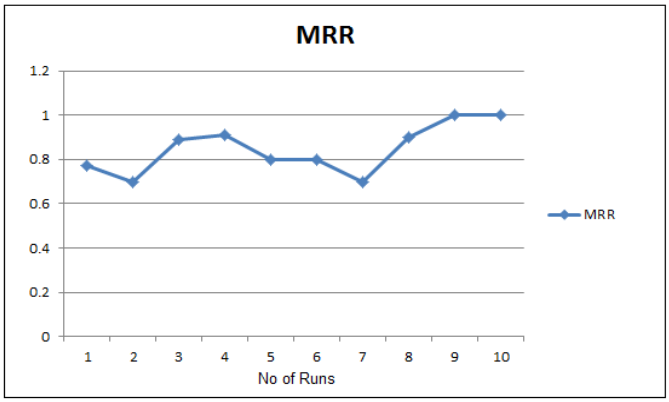

Figure 2. MRR for the Different Runs

CONCLUSION AND FUTURE SCOPE

Very much organized slides are created. Moderator's chance and endeavors are spared as it were. Slides will incorporate imperative key expressions and sentences identified with them. 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. 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] Rahul K. Sevakula, Abhi Shah and Nishchal K. Verma, "Data Preprocessing methods for Sparse Auto-encoder based Fuzzy Rule Classifier".

[3] A. Famili, Wei-Min Shen, Richard Weber, Evangelos Simoudis, "Data Preprocessing and Intelligent Data Analysis," intelligent Data Analysis, Vol. 1, No. 1, http:/www.elseviel:co&ocate/ida, 0 1997 Elsevier Science B.V. All rights reserved.

[4] Jirada Kuntraruk and William M. Pottenger, "Massively Parallel Distributed Feature Extraction in Textual Data Mining Using HDDI," 0-7695-1296-8/01 $10.00 0 2001 IEEE.

[5] Foram P. Shah and Vibha Patel, "A Review on Feature Selection and Feature Extraction for Text Classification," 978-1-4673-9338-6/16/$31.00 ©2016 IEEE.

[6] William E. Combs, "Using Fuzzy Logic in Large, Complex Data Mining Applications," 0-7803-7278-6/02/$10.00 Q2002 IEEE.

[7] Lotfi A. Zadeh, "Fuzzy Logic=Computing with Words", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 4, NO. 2, MAY 1996

[8] Paul Dickson, W. Richards Adrion, and Allen Hanson, "Automatic Capture of Significant Points in a Computer Based Presentation", Proceedings of the Eighth IEEE International Symposium on Multimedia(ISM'06)0-7695-2746-9/06$20.00©2006.

[9] Yuting Su, Zhong Ji, Xingguang Song, Rui Hua, "Caption Text Location with Combined Features Using SVM", 978-1-4244-2251-7/08/$25.00 ©2008 IEEE.