NOVATEUR PUBLICATIONS
International Journal of Research Publications in Engineering and Technology [IJRPET]
ISSN: 2454-7875
VOLUME 3, ISSUE 4, Apr.-2017

TEXT SUMMERIZATION USING WEIGHTED ARCHETYPAL ANALYSIS

PROF. VIJAY ANIL KULKARNI
(vijaykulkarni.soft@gmail.com), Contact no:-9420488939

IJRPET

ABSTRACT:

A nonfigurative summarization is used to acquire an understanding of the principle concepts and then individual those concepts in clear and easy language. It uses lexical methods to prove and translate the text and then to search the new concepts and classify to elaborate it by creating a new shorter form that direct the most important information from the main text document.

In this paper, we aim for text summarization based on Archetypal Analysis Algorithm (AAA) to make simple understandable summary.

An extractive summarization it can select important sentences, paragraphs etc. from the main collected document and integrate them into shorter form. The importance of sentences is decided based on statistical and lexical features of sentences.

1. INTRODUCTION:

Text summarization is a process, an important and timely tool for associate and translates text information in today's fast-growing information age. It is hard for human understanding to manually summarize large no of documents of text. There is an prosperity of text material available on the internet. However, we can get the more information through the Internet that we needed. Therefore, a duplex problem is detect: searching for suitable documents through an amazing number of documents is their presence, and interesting a large quantity of suitable information. The goal of automatic text summarization is keeping the source text into a shorter version preserving its information content and overall meaning.

A good summary system should return the several points of the document while keeping repetition to a minimum. Summarization tools may also search for title and other markers of subtypes in order to classify the key points of a document. Microsoft Word's summarize method is a example of text summarization.

Text Summarization has two types of methods-:

1. extractive 2. Abstractive.

An extractive summarization it can select important sentences, paragraphs etc. from the main collected document and integrate them into shorter form. The importance of sentences is decided based on statistical and lexical features of sentences.

2. LITERATURE REVIEW:

In the previous research, different techniques were presented for Text Summarization.

Manuel J.A. Eugster, And Friedrich Leisch has presented the paper on "Weighted and robust archetypal analysis" at (Sciencedirect.com).

Archetypal analysis represents observations in a multivariate data set as convex combinations of a few extremely points lying on the boundary of the convex hull. Data points which vary from the majority have great influence on the solution; in fact one outlier can break down the archetype solution. The original algorithm is adapted to be a robust M-estimator and an iteratively reweighted least squares fitting algorithm is presented. As a required first step, the weighted archetypal problem is formulated and solved. The algorithm is demonstrated using an artificial example, a real world example and a detailed simulation study

Christian Seiler 1, Klaus Wohlrabe has presented the paper on "Archetypal scientists" at Springers.com.
We introduce archetypal analysis as a tool to describe and categorize scientists. This approach identifies typical characteristics of extreme ('archetypal') values in a multivariate data set. These positive or negative contextual attributes can be allocated to each scientist under investigation. In our application, we use a sample of seven biblio metric indicators for 29,083 economists obtained from the Repec database and identify six archetypes. These are mainly characterized by ratios of published work and citations. We discuss applications and limitations of this approach. Finally, we assign relative shares of the identified archetypes to each economist in our sample.

Morten Mørup n, LarsKai Hansen has presented the paper on "Archetypal analysis for machine learning and data mining" at sciencedirect.com.

Archetypal analysis (AA) proposed by Cutler and Breiman(1994) [7] estimates the principal convex hull (PCH) of a dataset. As such AA favors features that constitute representative 'corners' of the data, i.e., distinct aspects or archetypes. We currently show that AA enjoys the interpretability of clustering– without being limited to hard assignment and the uniqueness of SVD – without being limited to orthogonal representations.

IJRPET

3. PROBLEM STATEMENT:

Summarization is a useful tool for selecting relevant texts, and for extracting the key points of each text. A good summary system should reflect the diverse topics of the document while keeping redundancy to a minimum. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document. Input to automatic text summarization is news articles from the Internet with various topics such as technology, sports, and world news to train the network. The aim of text summarization is convey the important information in limited sentence length from the source document.

4. EXISTING SYSTEM:

The primary intension of this technique is to design and develop an efficient and archetypal analysis approach for the text summarization. The document utilized for text summarization is prepared by set of pre-processing steps namely, sentence segmentation, tokenization, stop words and word stemming.

5. PROPOSED SYSTEM:

The proposed weighted archetypal analysis system combines the advantages characteristics f both feature and structure based method to obtain better summary and at the same time, compactness of the sentences can also be preserved. The input the approach is large document that has to be summarized, The pre-processed document is given to feature extraction which involve the identification of significance features.

6. OBJECTIVES & SCOPE:

6.1 OBJECTIVE:

Main objective of Document summarization is an automatic procedure aimed at producing a generic or a query-focused compressed summary of a document or a set of documents, sharing the same or similar topics, by reducing the document(s) in length.

6.2 SCOPE:

6.2.1 INPUT:

I used 30 news articles from the Internet with various topics such as technology, sports, and world news to train the network.

6.2.2 OUTPUT:

Summary of document or set of document(s).

7. METHODOLOGY:

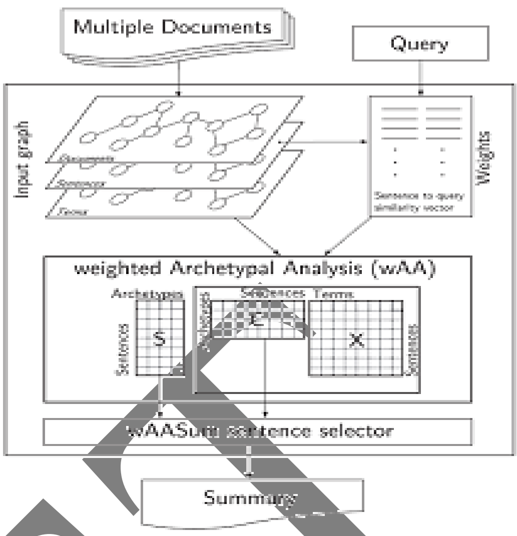

Fig: - Text summarization using WAA (Weighted Archetypal Analysis)

8. SYSTEM DESIGN:

System design is the conceptual model that defines the structure, behaviour, and more view of a system. Architecture description is a formal description and representation of a system, organized in way that that support reasoning about the structure of system which comprises system components.

8.1 DATA FLOW DIAGRAM (DFD):

A DFD is graphical representation of the "flow" of data through an information system, modeling its process aspects. A DFD is often used as a preliminary step to create an overview of the system, which can later be elaborated .A DFD shows what kind of information will be input to and output from the system, where the data will come from and go to , and where the data will be stored. It deed not show information about whether processes will operate in sequence or in parallel.

Level 0

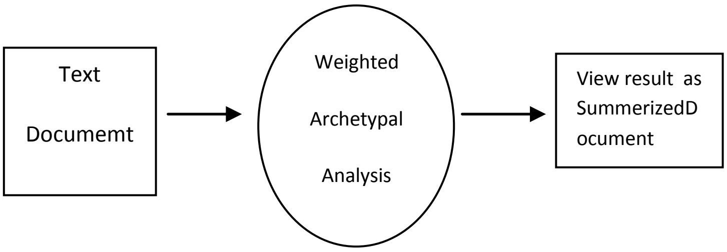

Fig: - Data flow diagram level 0

8.2 ARCHETYPAL ANALYSIS:

Archetypal Analysis (AA) is data point in a data set as a mixture of points of pure, not necessarily observed, types or archetypes. The archetypes themselves are restricted to being sparse mixtures of the data points in the data set, and lie on the data set boundary, i.e., the convex hull. AA model can naturally be considered a model between low-rank factor type approximation and clustering approaches, and as such offers interesting possibilities for data mining.

Order to classify the key points of a document. Microsoft Word's summarize method is a example of text summarization.

Text Summarization has two types of methods:

1. Extractive 2. Abstractive.

An extractive summarization it can select important sentences, paragraphs etc. from the main collected document and integrate them into shorter form. The importance of sentences is decided based on statistical and lexical features of sentences.

An Abstractive summarization is used to develop an understanding of the main concepts and then individual those concepts in clear n simple language. It uses lexical methods to prove and translate the text and then to search the new concepts and classify to describe it by creating a new shorter form that conduct the most important information from the main text document.

In this paper, we propose text summarization based on Archetypal Analysis Algorithm (WAA) to make simple understandable summary.

11. RESULT AND CONCLUSION:

The paper has formalized the problem of the document summarization as the weighted archetypal analysis problem. Additionally, paper has presented our study of how to Incorporate information in the own nature of AA and how to use weighted version of AA for simultaneous sentence clustering and ranking. We have examined the proposed method on several input matrix modelling configurations, where the paper reports the best results on the multi-element graph model. The paper has found that wAASum is an effective summarization method.

Experimental results on the DUC2002 or DUC2005 datasets demonstrate the effectiveness of the proposed approach, which compares well to most of the existing matrix factorization methods in the literature.

REFERENCES

1) Alguliev, R. M., Aliguliyev, R. M., & Hajirahimova, M. S. (2012). *Gen Doc Sum + MCLR: Generic document summarization based on maximum coverage and less redundancy.*

2) Chan, B.-H.-P. (2003). *Archetypal analysis of galaxy spectra. Monthly Notices of the Royal Astronomical Society*, 338(3), 790–795.

3) Cutler, A., & Breiman, L. (1994). *Archetypal analysis. Technometrics.*

4) Lee, J.-H., Park, S., Ahn, C.-M., & Kim, D. (2009). *Automatic generic document.*

IJRPET