IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Weighted Archetypal Analysis used for Text Summarization

Miss. Vaishali Siddharam Shakhapure[1], Prof. A. R. Kulkarni[2]

Computer Science and Engineering Department, Walchand Institute of Technology, Solapur University,

Solapur, India[1, 2]

Abstract: This paper demonstrate a summarization system that generates a summary for a given documents based on sentence similarity measures using weighted archetypal analysis. Most of the former approaches for multi document summarization give the summary by using only the Query Focused methods applied on the given document. And many approaches like matrix factorization methods to search either low rank approximation method or hard/soft clustering methods to get better document summary for the given documents. In this paper we propose different method called Weighted Archetypal Analysis an efficient approach for multi document summarization using extractive type summarization which uses the term frequency for sentence importance measures: Frequency of the terms i.e. archetypes, values in the sentence. The sentences are ranked according to their respective weights (scores) and the top rank (highest weight) sentences are selected for summary. The summary is generated by using Weighted Archetypal Analysis to compute archetypes, term frequency and significant sentences to the target documents evaluation measure.

Keyword: Generic document summarization, weighted archetypal analysis, Text summarization, Matrix factorization approach, Term Frequency.

1. INTRODUCTION

Document Summarization is necessary now a day because use of internet and availability of information has been increasing day by day, it is impossible to read whole document instead we can read compressed version of that document easily. Document summarization means generating the compressed version of document in meaningful manner. Summarization can be a Single Document or Multi-document. Single Document Summarization means one larger document is condense into a compressed version, on other side Multi-document summarization means meaningful small version of set of documents.

There are two ways of document summarization namely Query Focused Document Summarization and Generic Document Summarization. Query Focused document Summarization means to select most essential sentences corresponds to the given query and Generic Text Summarization continue with primitive semantic effect of the original document.

Query focused summarization is important because use of the internet and availability of the information, it produce only the essential information required by the user. Generic summarization it must hold essential information central to original document.

2. RELATED WORK

Generic Multi-document summarization:

All over the world many researcher have been working on generic multi-document summarization with many different ways to know the best results: (Tao, Zhou, Lam, & Guan, 2008; Wan, 2008; Wang, Li, Zhu, & Ding, 2008; Wang, Zhu, Li, Chi, & Gong, 2011; Wang, Li, Zhu, & Ding, 2009). Proposed paper is based on the abstractive summarization information is expressed in different manner and generated summary consist the sentences from original documents. Abstractive summary consist all the word and phrases except original document. Extractive summary is simple and robust method for text summarization. Summary can be a query focused or generic proposed by (Dunlavy, O'Leary,Conroy, & Schlesinger, 2007; Gong & Liu, 2001; Ouyang, Li, Li, &Lu, 2011; Wan, 2008) query focused is summary should contain the information similar to query at other side generic should contain the semantic sentences from the whole document. Query expansion algorithm is used with graph based ranking approach for multi-document summarization. This algorithm is for sentence-sentence and sentence-word relation to show expansion word from the document. Because of this expansion word it can achieve information richness and query relevance.

The research in Harabagiu and Lacatusu (2010) has two main goals. First is to generate topic theme automatically of multi-document summarization and second he has used eight different methods for multi-document summarization.

There are two ways of extractive based summarization one is based on semantic similarity and another one is based on sentence similarity both are used for extracting important sentences from document. MEAD method is used to extract the sentences which are cancroids of cluster. Selected sentences are important to the cluster or target document. Gong and Liu (2001) examined latent semantic analysis (LSA) to produce a summary of highest rank sentences. Similarly some methods like NMF-based topic specification (Lee, Park, Ahn, & Kim, 2009; Wang et al., 2008, 2009) and CRF-based summarization (Shen et

DOI 10.17148/IJARCCE.2015.41151

al., 2007). In framework CRF (conditional random fields), importance of each sentences is evaluated by CRF when the documents provide sequence of sentences as input.

Mei and Chen (2012) it is fuzzy based medoid clustering , it is instance of soft clustering for query focused multi-document summarization, it produces summary based on the subtopic it estimate the subset of sentences from main topic. It is based on subtopic so it can't conclude the applicable summary.

(Alguliev, Aliguliyev, & Hajirahimova, 2012) make use of Maximum Coverage and Less Redundancy (MCLR) model, in this they have used the quadratic Boolean based multi-document summarization. Used function contains the weighted combination of content and redundant documents by using quadratic programming of integer.

Lee et al. (2009) has experimented the query focused summarization method using non-negative factorization methods.
This is algebraic method and most successful, it produces summary based on similarity between semantic characteristics and query. This method uses the non negative matrix factorization method with term sentence to produce most relevant sentences from the document. Initially sentences are clustered based on their similarity then chooses most relevant sentence from the cluster and produce as summary. Former research used Latent Semantic Analysis for sentence selection as summary but it produces less meaningful sentence than Non-negative matrix factorization method.NMF uses cosine similarity between query and sentences also uses semantic characteristics to produce summary because of this it produce comparatively deficient summary.

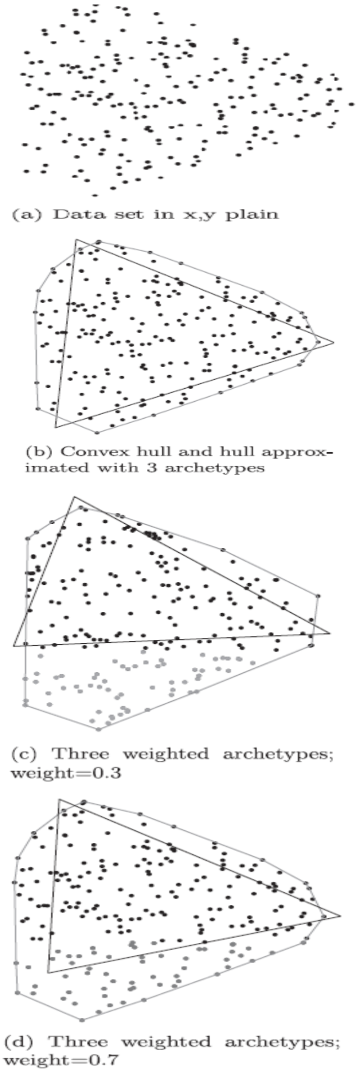
Symmetric Non-negative Matrix Factorization proposed by Lee et al. (2009), it is sentence level semantic analysis this get the sentences semantically based on the relationship between sentences using this it divides matrix and produces relevant sentences, concluding this it can't shows the nearness of cluster based summary for sub-topics.

3. ARCHETYPAL ANALYSIS

“the primary pattern or model of which all similar things are representations or copies."

Cutler and Breiman (1994) has produced the Archetypal Analysis, it means that it contain the data points from the dataset. Main goal of this analysis is to search some data points, not essentially observed, observed data points from various datasets in such way that each and every observation are close enough to combination of convex of archetypes. Complete archetypes are convex combination of observations.

Following figure1 describes the Archetypal Analysis:
Archetypal Analysis model is clustering Approach and Low-Rank Approximation approach; it is useful in data mining.
Archetypal Analysis is useful in probabilistic ranking, soft clustering. Application of Archetypal Analysis is in neuro imaging, chemistry, text mining, etc. Eugster and Leisch (2011) he discovered first Weighted Archetypal Analysis Algorithm.



(a) Data set in x,y plain

(b) Convex hull and hull approximated with 3 archetypes

(c) Three weighted archetypes; weight=0.3

(d) Three weighted archetypes; weight=0.7

Fig1: Set of data is represented by the convex hull to approximate Archetypal Analysis:

In figure (c, d) gray data point represents weight 03 and 0.7 respectively. In all figures black data points represents weight as 1. Based on point weights corresponding data points change their place.
Suppose an n x m matrix X. For a given k, the archetypal problem is to find the matrix Z of k m-dimensional archetypes.

$$X \approx SY \quad \text{with } Y = X^T C \qquad (1)$$

$$RSS(k) = \|X - SY^T\|^2 \quad \text{with } Y = X^T C$$

$$s.t. \sum_{j=1}^{z} S_{ij} = 1, S_{ij} \geq 0, i = 1, \ldots, n$$

$$\sum_{i=1}^{n} C_{ij} = 1, C_{ij} \geq 0, j = 1, \ldots, z \qquad (2)$$

DOI 10.17148/IJARCCE.2015.41151

More precisely, to find the two n x k coefficient matrices C and S which minimize the residual sum of squares.

Here, in above formula $\sum_{i=1}^{n} C_{ij} = 1$ with $C_{ij} \geq 0$ apply constraint matrix $Y = X^T C$ as convex combination and formula $\sum_{j=1}^{k} S_{ij} = 1 \text{ and } S_{ij} \geqslant 0,$ for meaningful combination of data points of archetypal component that is

$$[X]_{m \times k} = [[W_M]_{m \times m} \odot [A]_{m \times m}]_{m \times m} \otimes [G]_{m \times k} \qquad (5)$$

$\bar{X} = SY^T.$ here data points considered as mixer of archetypes $\|.\|^2$ this represents the Euclidean matrix norm.

Weighted Archetypal Analysis:

Same weight data points and remaining data points are reduced in equation (2), weighted Archetypal problems are formulated as below,

$$RSS(k) = \|W(X - SY^T)\|^2 \quad \text{with } Y = X^T C \qquad (3)$$

Assigning weight to remaining data points are similar to assigning weights to data set,

$$W(X - SY^T) = W(X - S(X^T C)^T) = W(X - SC^T X) = WX - W(SC^T X)$$
$$= WX - (WS)|(C^T W^{-1})(WX) = \hat{X} - \hat{S}\hat{C}\hat{X} = \hat{X} - \hat{S}\hat{Y}^T$$

Minimized formula is written as below

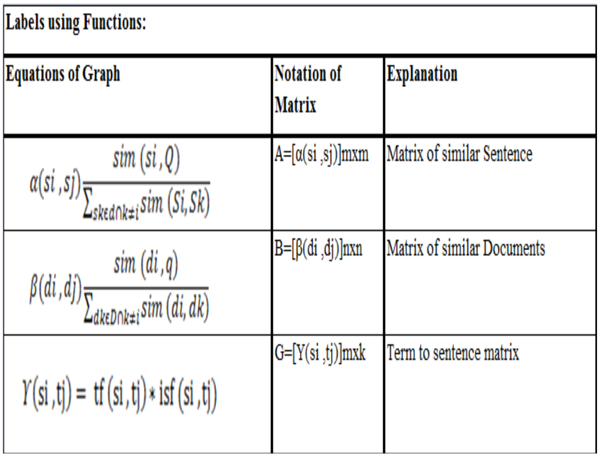
$$RSS(k) = \|\hat{X} - \hat{S}\hat{Y}\|^2 \text{ with } \hat{Y} = \hat{C}\hat{X} \quad \text{and } \hat{X} = WX \qquad (4)$$

Weighted Archetypal Analysis for generic document Summarization:

Following table contain the formulae, there notations and explanation;

Table contains the various formulae:

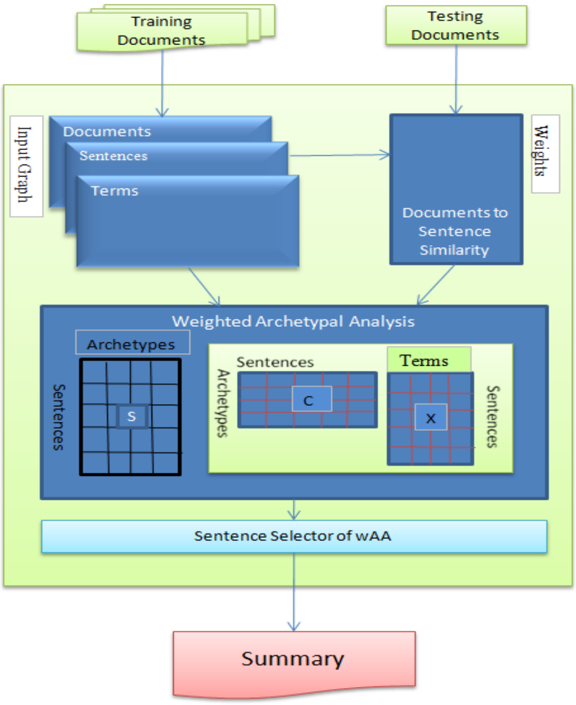
First formula is for calculating similarity between the sentences by using this formula redundancy of sentences has eliminated. Second formula is for identifying the similar documents from the given documents and third formula is for extracting the word i.e. terms from the given documents.



Labels using Functions:		
Equations of Graph	**Notation of Matrix**	**Explanation**
$\alpha(si, sj) \frac{sim(si, Q)}{\sum_{sk \in d \cap k \neq i} sim(Si, Sk)}$	A=[α(si ,sj)]mxm	Matrix of similar Sentence
$\beta(di, dj) \frac{sim(di, q)}{\sum_{dk \in D \cap k \neq i} sim(di, dk)}$	B=[β(di ,dj)]nxn	Matrix of similar Documents
$Y(si, tj) = tf(si, tj) * isf(si, tj)$	G=[Y(si ,tj)]mxk	Term to sentence matrix

4. ARCHITECTURE

Architecture of Archetypal Analysis is as shown below:



Framework of Archetypal Analysis is as given below:

1. Matrix M is to be constructed using equation(5)
2. Input of weight matrix W to be generated by ($S_{i,}$ q) labeling function it shown in last column of table.
3. For given W execute Weighted Archetypal Analysis Algorithm on matrix M:

I. Factorize the matrix M into S and C as explained above.
II. Amount of values given in related row of the matrix CX, $Sa_i = \sum_{j=1}^{m} CX_{i,j}.$ that is it should calculate important S_{ai} for each archetype i.
III. Resultant archetypes from above procedure need to be sort in descending order according to their significance that is matrix C generates the value of $S_{ai.}$
IV. Lowest weight archetypes i.e. insignificant archetypes ε removed from the result.

4. Select highest weight Archetypal sentences from the matrix:

I. Begin with most important archetype and distil highest weight sentences from the matrix C.
Start with significant sentence and retrieve important sentences from matrix C and if the length of summary is not matching with prerequisite value then continue with next important sentence.
II. Every selected sentence compare with former sentence if both are same the newly selected sentence need to be removed from final summary.

In above algorithm we have used the ranking for selecting important sentences from matrix C, for this step 3 and step 4 are most important steps in algorithm. Algorithm selects

DOI 10.17148/IJARCCE.2015.41151

the highest weight sentences i.e. it meets both the aspects like covering all points from matrix c and selecting important sentence from matrix as well. Above procedure is similar to Archetypal Analysis definition.

5. EXPERIMENTAL RESULTS

Results of Weighted Archetypal Analysis based in the sentence similarity graph:

Document	Recall	Precision	F-Measure
Sport News	0.41369606	0.44144144	0.42711864
International News	0.25960637	0.29343220	0.27548483
World News	0.29530201	0.33748801	0.31498881
Politics	0.39302694	0.38183217	0.38734869
Entertainment	0.44889267	0.52647352	0.48459770

Table 1 Comparison Result of Our Summarizer

The above table shows the comparison of summaries generated by our Summarizer with human-generated summaries. Here the human generated summary is taken as reference summary and summary generated by our Summarizer is taken as candidate summary. The numerical data shows how automated generated summaries are closer to human-generated summaries with the help of ROUGE-N technique.

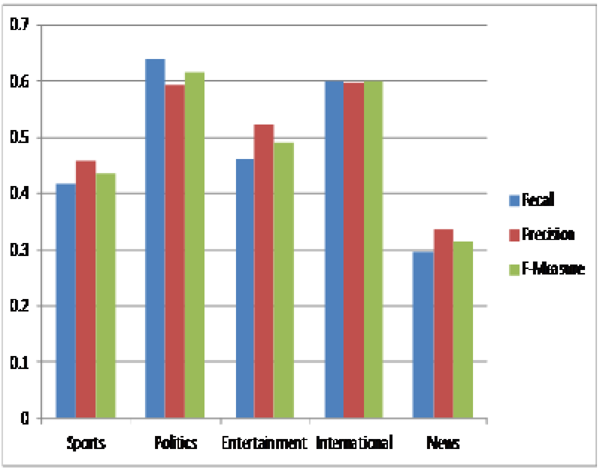


Fig 1 Comparison Graph of Our Summarizer

The above chart shows the representation of numerical statistics of Table 1 into graphical form. In above chart horizontal line represents event names and vertical line numerical values.

Document	Recall	Precision	F-Measure
Sports	0.41674876	0.4582881	0.4365325
Politics	0.64076690	0.5940130	0.6165048
Entertainment	0.46194690	0.5230460	0.4906015
International	0.60037523	0.5986903	0.5995316
News	0.29530201	0.3374880	0.3149888

Table 2 Comparison Result of Previous Summarizer

The above table shows the comparison of summaries generated by precious Summarizer with human-generated summaries. Here the human generated summary is taken as reference summary and summary generated by previous Summarizer is taken as candidate summary. The numerical data shows how automated generated summaries are closer to human-generated summaries with the help of ROUGE-N technique.

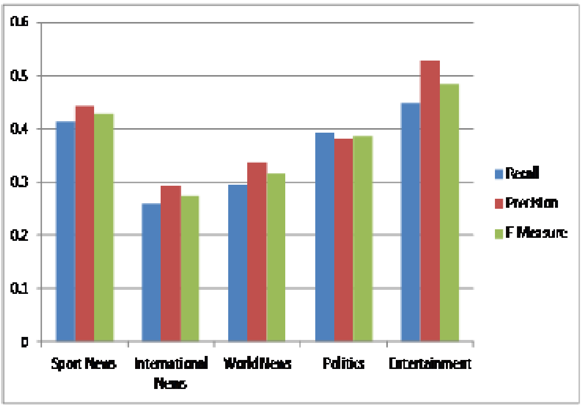


Fig 2 Comparison Graph of Previous Summarizer

The above chart shows the representation of numerical statistics of Table 2 into graphical form. In above chart horizontal line represents event names and vertical line numerical values.

With the rapid growth of the internet, the information available online become huge. As the time being a constraint, user needs to get the gist of the document in a short span of time. Document summarization serves as a tool to get the theme of the document effectively. Our proposed approach presents a summarization method that uses the neighbourhood knowledge and two statistical measures, sentence similarity and term frequency, for document summarization. Considering the term frequency besides with sentence similarity will give the summary more effectively.

Weighted Archetypal Analysis Results:

Above table contain the results of Weighted Archetypal Analysis.

We used different datasets to determine the significance of Weighted Archetypal Analysis algorithm, each and every dataset contains 5 documents related to specific domain with condition that generate summary according to users choice, i.e. we have provided the range(5%-40%) for generating summary. If dataset contains total 100 sentences then if we want only 20% of sentences from set of 100 sentences then we need to select 20% to generate summary. To evaluate these summaries we have used the Recall Oriented Understudy for Gisting Evaluation (ROUGE) evaluation package, this is one of the best evaluation tools to determine the significance of generated summary, ROUGE compares the various resultant summaries with reference summary or human generated summaries. It is very useful to measuring summaries.

We have used the Weighted Archetypal Analysis for multi-document summarization, various other algorithms uses the matrix factorization method for generic multi-document summarization, and this method uses the property of Non-negative matrix factorization method to extract significant sentences from the provided documents. Our used approach gives the best results than the specified methods, we have used the weighted archetype method to extract the archetypes from the given document set, according to that it clusters and rank the sentences of given documents this approach overcome the limitations of matrix factorization method by using the both clustering matrix factorization method due to this it gives the best results than compared methods.

6. CONCLUSION AND FUTURE WORK

The paper has presented the problem of generic document summarization as weighted archetypal analysis problem, with this paper we have examined how to interpret the selected sentences information with own nature of Archetypal Analysis and how to use weighted Archetypal Analysis with ranking and clustering. We have demonstrated and examined method with various other multi element graph based models. The weighted archetypal analysis is efficient summarization method. This project is compared with various closely related methods; it gives best result in comparison.

In future the efficiency of project could be improved; there are various possible ways for improvement of weighted archetypal analysis:

(1) At present we have used the simple statistical and semantic characteristics for calculating the similarity between the sentences semantically, in advance we can use WordNet for calculating the semantic similarity between the sentences by using synonyms set of their component term.

(2)We can apply advanced method for query processing/expansion techniques.

(3) In future we can apply this method to other summarization work like comparative and update summarization.

REFERENCES

[1**]** Elena Lloret, María Teresa Romá-Ferri, Manuel Palomar; A text summarization system for generating abstracts of research papers, Department of Software and Computing Systems, University of Alicante,Spain, 2013.

[2] Ferreira, R. ; Freitas, F. ; De Souza Cabral, L. ; Dueire Lins, R. A Context Based Text Summarization System, Document Analysis Systems (DAS), 2014

[3] Erwin, A. ; Eng, K.I. ; Muliady, W. Automatic text summarization based on semantic analysis approach Information Technology and Electrical Engineering (ICITEE), 2013

[4] Manuel J. A. Eugster, Archetypal Analysis Mining the Extreme, Institute fur Statistic, Ludwig-Maximiliams-Universiy at Munchen , HIIT seminar, Helsinki Institute for Information Technology, 2012.

[5] Y. Surendranadha Reddy*, Dr. A.P. Siva Kumar; An Efficient Approach for Web document summarization by Sentence Ranking, M.Tech, Department of CSE JNTUA College of Engineering Anantapur, India, 2012.

[6] Zhang Pei-ying, Automatic text summarization using sentences extraction and clustering, College of Computer & Communication Engineering., China University of Pet., Dongying, China, 2009

[7] Multi-document Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation, Future Computer and Communication, ICFCC, 2009.

[8] Peng Zhao ; Enhong Chen ; Qingsheng Cai; Huantong Geng, A Novel Automatic Text Summarization Study Based on Term Co-Occurrence, Cognitive Informatics, 2006

[9] Udo Hahn, Inderjeet Mani; The Challenges of Automatic Summarization, Albert Ludwigs University, Mitre Corp.2000

[10] Vaishali Shakhapure[1], A. R. Kulkarani**;** Text Summarization using weighted Archetypal Analysis, Walchand Institute of Technology, Solapur, India(2015).

[11] Rasim M. Alguliev, Ramiz M. Aliguliyev, Nijat R. Isazade; Multiple documents summarization based on evolutionary optimization algorithm, Institute of Information Technology and National Academy of Sciences, Azerbaijan(2012).

[12] Rasim M. Alguliev, Ramiz M. Aliguliyev, Nijat R. Isazade; CDDS: Constraint-driven document summarization models, Institute of Information Technology of Azerbaijan National Academy of Sciences, Azerbaijan(2013).

[13] You Ouyang Wenjie Li Qin Lu Renxian Zhang**;** A Study on Position Information in Document Summarization**,** Department of Computing, the Hong Kong Polytechnic University(2010).

[14] Rasim M. Alguliev, Ramiz M. Aliguliyev, Nijat R. Isazade; DESAMC+DocSum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization, Institute of Information Technology, Azerbaijan National Academy of Sciences, Azerbaijan(2012)

DOI 10.17148/IJARCCE.2015.41151