International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Text Summarization using weighted Archetypal Analysis

Vaishali Shakhapure[1], A. R. Kulkarani[2]

[1 2]Walchand Institute of Technology, Solapur, India

Abstract: ***Present research held on the Matrix Factorization approaches to Query-Focused Multi-Document Summarization searched by hard/soft clustering or Low rank approximation methods. We utilised distinct method called matrix Factorization to Query focused multi-document Summarization. Query Focused Summarization means set of sentences are integrated to graph and weighted by similarity between query and sentences in the document. Data set boundaries are weighted by values on positive or negative sentences. This project mainly focused on using Archetypal Analysis algorithm in existing matrix factorization method to show the important sentences of the document in the summary also studied model called multi-element graph and its effects on query focused summarization through weighted archetypal analysis.***

Keywords: Archetypal analysis, Convex hull, Matrix Factorization, Multi-element graph.

1. Introduction

Document Summarization means producing compressed document for better understanding within short time. Summarization can be performed by two ways first one is Generic and second is Query-focused Summarization. Generic summarization it defines main properties of the document i.e. semantic of document in contrast query focused summarization generate the summary from all the documents to which queries are interpreting. Query focused summarization is suitable for multi document summarization. These two summarization technique applies on Single document or multi-document. Single document summarization means generating the summary of one document in other hand multi-document summarization is like data mining i.e. integration of information from multiple documents into single document Summarization means compact model of a wide document. Summary gives the brief information of large document to understand speedily.

Researchers are drawing towards generating the summary of document because use of internet becomes increasing day by day and to get the required document from vast availability of documents is challenging task. To fulfil the user requirements queries plays an important role but queries are practically complicated (e.g. The Apparent Incompatibility of the Law of Propagation of Light with the Principle of Relativity). Such interconnected queries are difficult to express practically.

Present research on soft or hard clustering and low rank approximation of matrix factorization method and query focused multi document summarization. Earlier is more flexible but to be difficult to represent features. Recent distil the features same as existent data, this is easy to understand but use of binary assignment makes it invariable. Analysing all the above techniques generates a summary means contains list of document interpreted by query. Document summarization using query is only information related t topic it means generated summary should hide the multifariousness up to the feasible extent.

The main focus of this project is successfully deals with 1. How to include query information into main characteristics of archetypal analysis based summarizer. 2. How to enhance the quality of summarization and diverseness of created summary based on query –focused technique. We examined and noted our result with other summarization method; we have used data set of DUC2002 and other progressive approaches.

2. Related Work

In the earlier research, different techniques were proposed and presented for Text Summarization. Summarization of document based on query focused method:(Christophe Brun, Jean-François Dufourd, Nicolas Magaud 2012). Designed and provided a correct convex hull algorithm with hyper map paper contain the information for correctness of algorithms are ensured and programming techniques progressed by computational geometric. This includes case study of classical problems geometrical computation which contain object of geometric element by collecting planer points of incremental convex hull [1]

Non-negative Matrix Factorization for document clustering (Wei Xu, Xin Liu, Yihong Gong). In this paper relations are find by partition clustering method using Open Relation Extraction (ORE) which was not used in former project. Earlier they used manually defined relation to distil the patterns. Subsequently the pattern can extract automatically.[2]

Multi-objective evolutionary algorithms using Convex hull ranking algorithm (M. Davoodi Monfareda,∗, A. Mohadesa, J. Rezaei). In this project they used convex hull method to interpret the ranking procedure for algorithms of evolutionary multi-objective this algorithm is most suitable for proposed problem of multi-objective evolutionary algorithms.[3]

Non-negative matrix factorization is used for Automatic generic document summarization Inha University, Incheon, Republic of Korea, Department of Computer Science and Information Engineering. In proposed project they used non

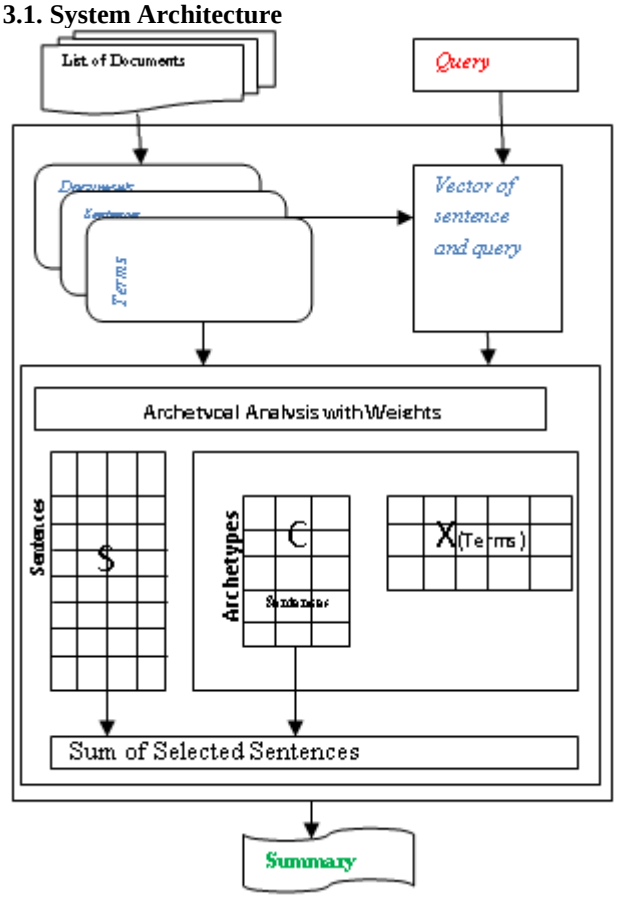
negative values suitable for process of human knowledge and used matrix factorization method to select important sentences for automatic generic document summarization. Selected sentences are significant for automatic generic document summarization than the Latent Semantic Analysis (LSA).[4]

Functions of Archetypal:
Economics and Law Department , Italy (Giovanni C. Porzio, Paola Costantini, Juan Romo, Giancarlo Ragozini). Archetypal analysis means using some underlying ideas discovers the datasets by some mathematical procedure. Main objective of this function is detect some point not an essentially determined but combine all points by approximation of convex combination of determined data.[5]

LexRank (Erkan & Radev, 2004) andTextRank (Mihalcea. & Tarau, 2004) this are the Graph-based methods. It takes sentences as vertices and compute the weight based on similar sentences. This method consider information altogether to extract important sentence to produce summary [6]

3. Proposed Systems

3.1. System Architecture



1. Formula to create matrix X:
$[X]mx\,k = [[W_M]mx\,m\,O\,[A]\,mx\,m]\,mx\,m\,[G]\,mx\,k$
$\otimes$
2. This is diagonal matrix of query to sentence similarity this shows W weighted matrix

$$\partial(si\ ,q) = \frac{sim\ (si\ ,q)}{\sum_{sk\ \epsilon D} sim\ (sk\ ,q)}\ W$$
$$= [\partial(si\ ,q)]mxm$$

3. Calculate weighted archetype AA of matrix X:
a) Calculate C and S as given in equation (2)
b) Calculate importance of archetype by using
c) $Sa_i = \sum_{j=1}^{m} CXij$. CX is sum of values.
d) Arrange the archetypes based on the C and Sa_i values.
e) Removes the archetypes those are less important i.e. lowest weight archetype.

4. Select archetypes *l* which has the highest weight than other archetypes
a) Select important archetype i.e. having maximum weight calculated by selected sentence C and keep on selecting archetypes in decreasing order (having weight less) until best level summary not generate.
b) Then selected sentence is compare with earlier sentence if it similar then newly selected sentence should not include as summary.

3.2. Implementation Details:

3.2.1. Analysis of Weighted Archetypal
Important form of Archetypal Analysis is shown in first section of summarization (Cutler & Breiman, 1994), and then we present inherited form of archetypal analysis, weighted version from (Eugster & Leisch, 2011)paper.

3.2.2. Archetypal Analysis:
Take an m x n matrix, X stands for variable data set with n observation and m attributes then factorize matrix into random matrices S ϵ R^{nxz} and C ϵ R^{nxz} as follows:

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

S and C initialise to calculate starting archetype XC based on constraints. Initially it chooses the value randomly then using equation (2) continues updating S and C until it reach to maximum iteration.
$RSS(k) = \| X - SY^T \|^2$ with $Y = X^T C$

$\sum_{j=1}^{n} Sij = 1\ S_{ij} \geq 0$, i=1, ..,n

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

Both $\sum_{j=1}^{n} Sij = 1$ and $S_{ij} \geq 0$ used for feature matrix and $\sum_{i=1}^{n} Cij = 1$, $Cij \geq 0$ used for bell-shaped(convex) combination, $X = SY^T$ shows meaningful combination of archetypes. ‖.‖ This represents the Euclidean matrix for archetype mixture.

3.2.3. Weighted archetypal analysis
$RSS(k) = \|W(X - SY^T)\|^2$ with $Y = X^T C$
X= n x m matrix and W= n x n Square matrix, W is weight matrix shows similarity between sentences and query.

3.2.5. Summary of document using Query and Weighted Archetypal Analysis:

Query focused Multi-document summarization selects Sentences and represents using Weighted Archetypal Analysis by following methods:

Here sentences are grouped into weighted archetypes and n, m,k represents documents, sentences and terms respectively

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

3.3. Expected Results

The weighted Archetypal based summarization technique compare with other existing summarization techniques to estimate effectiveness. We compared results with DUC 2004 and DUC 2005 and other proposed techniques, it generate the summary less than 250 words.

4. Conclusion

The paper has validated with query based document summarization and weighted archetypal analysis problem in addition with this we have used AA with query information also used weighted methods of AA for clustering and ranking simultaneously. Compared our results with several existing summarization techniques.

We can improve this algorithm by using WorldNet for interpreting the similarity between sentences and set of terms and many other techniques can be used.

In the comparison of other technique this project gives the best result because use of the weighted archetypal analysis includes query information into its own form and weighted edition for ordering and combining the most significant sentences for summarization.

References

[1] Machine learning and data mining by Archetypal Analysis Technical University of Denmark, Richard Petersens Plads, Denmark (2011)
[2] Functions of Archetypal Department of Economics and Law, University of Cassino, Italy (2010)
[3] Multi-document summarization using Weighted archetypal analysis and query-focused method of the multi-element graph Faculty of Computer and Information Science, Ljubljana University of Slovenia (2013)
[4] Discovering Relations using Matrix Factorization Methods Max-Planck-Institute of Informatic , Germany
[5] Generic Document Summarization settled on Non-negative Matrix Factorization A Department of Computer Science and Information Engineering, Inha University, Republic of Korea (2013)
[6] Archetyping for Courseware Valsamidis Kavala' s Institute of Technology, Kavala, Greece (2013)
[7] Scientists using Archetypal Munich University, Economic Research Institute (2013)
[8] Convex hull ranking algorithm for multi-objective evolutionary algorithms Laboratory of Algorithms and Computational Geometry, and Amirkabir ,Department of Mathematics and Computer Science, University of Tech,Iran(2011)
[9] The Study for summarization of multi-document Automatically New York University and Communications Research Laboratory,Hikaridai, Japan .
[10] The Summaries are automatically evaluated by package called ROUGE University of Southern California, Information Sciences Institute.