Ms. Meghana. N.Ingole, Mrs.M.S.Bewoor, Mr.S.H.Patil / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 4, July-August 2012, pp.168-171

Text Summarization using Expectation Maximization Clustering Algorithm

Ms. Meghana. N.Ingole, Mrs.M.S.Bewoor, Mr.S.H.Patil
Department of Computer Engineering,
Bharati Vidyapeeth Deemed University, College of Engineering, Pune, India.

ABSTRACT

Large amount of data is available on internet due to increase in growth of World Wide Web. It is very difficult for human beings to manually find out useful and significant data. This problem can be resolved by using text summerization.Text Summarization is the process of condensing the input text file into shorter version by preserving its overall content and meaning. This paper is about called text summarization using natural language processing. This paper describes a system where in first module we are implementing the phases of natural language processing that is splitting, tokenization, part of speech tagging, chunking and parsing. In second module we are implementing Expectation Maximization Clustering Algorithm to find out sentence similarity. Based on the value of sentences similarity, we can easily summarize text.

Keywords: Document Graph, splitting, similarity ,tokenization,.

I.INTRODUCTION

Overview There is abundant amount of information available on the web. It will be easy for user to search particular information by entering keyword but it is difficult to find out useful and significant information. Text summarization is an important tool for assisting and interpreting the data when vast amount of data is available. Text summarization not only means to condense the information but also to maintain the original meaning. Abstractive summarization means a method which consists of understanding the meaning of text and then giving it in fewer words where extractive summarization means picking only important text from original document and concatenating them into shorter version. Interpretation of the text can be one by using natural language processing. In natural language Processing, the input text has to go through different phases where sentence can be split, parsing can be done, and then tokenization and finally chunking has to be done.

These pre-processing can be done before generating summery. The main aim of research work is to twofold. One is to implement Expectation Maximization Clustering algorithm. Second is query dependent summarization by removing ambiguity. For interpreting the text used word net dictionary is used. The input text is processed where each sentence is considered as a single node and each node will be compared with all other node. This comparison can be done by calling word net dictionary where it will calculate weight. Then it is represented in the form of document graph. Expectation Maximization Clustering Algorithm is used before summarization to generate effective summary.

II. RELATED WORK

The Expectation Maximization clustering algorithm is used for query dependent text summerization.This technique will avoid cumbersome work of user, the user can easily get query dependent summary of text document. The proposed work is context sensitive text summarization using Expectation Maximization Clustering algorithm. The different clustering algorithm can also be implemented. The main focus of summarization is using natural Language processing which will remove sentence ambiguity, and will get accurate summary maintaining the originality. The phases of NLP are an important task where it will chunk the text in the form of graph by understanding meaning of text.

However after going through some papers and the findings are somewhat different from my analysis in the algorithm. The paper [1] considers how to achieve high accuracy and how to represent count data where they have considered different data clustering model. Finally the mixture model made up from different distribution is optimized by expectation-maximization and minimum description length..In this paper, [2] unsupervised document has been considered where document is text file and they have used probabilistic

model. The model is of word count and different component which corresponds to different theme. Here Expectation Maximization is used as a tool which is used to calculate maximum a posteriori (MAP) estimates of the parameters and in obtaining meaningful document clusters. Finally MAP has estimated by Gibbs Sampling. Paper [3] is about sentence clustering in a document. Generally sentence similarity does not represent sentence in metric form, this paper used fuzzy clustering algorithm operating on data which is in the form of square matrix. Here for graph representation, Expectation maximization algorithm has object in graph is interpreted as a likelihood. Finally the paper result is the algorithm is capable of been used where central identifying overlapping clusters of semantically related sentences. The drawback is time complexity because page rank is applied to each EM cycle which takes large convergence time if there are large number of clusters.[4] Expectation Maximization algorithm is used for interaction between image segmentation and object recgnition.Image segmentation is done at E step and reconstruction of object is done at M step. For capturing shape and appearance of image, Active Appearance Model (AAMs) has been used. The [5] technique is used Expectation Maximization algorithm for estimation of statically parameters of classes in Bayesian decision rule. It is composed of context sensitive initialization and iterative procedure for statically parameters. They have used different algorithms for unsupervised classification in multistage clustering. [6] Focused on co-clustering and constrained clustering to improve the performance and for increasing effectiveness supervised and unsupervised constraints has been considered. For optimization of model, Expectation Maximization algorithm is used. They have used NE extractor for constructing document constraints based on overlapping entities and word net for word constraints based on semantic distance.

III.SYSTEM DESCRIPTION

1. Input:

Input text file containing set of sentences probably from same context is fed to the system. The stop words like “the”, “him” and “had” which do not contribute to understanding the main ideas present in the text can be removed.This long text can be pre-processed through NLP phases. Distance matrix can be calculated and Expectation Maximization clustering algorithm is implemented which builds document graph and will generate specific summary. Every node in the cluster maintains association with every other node.

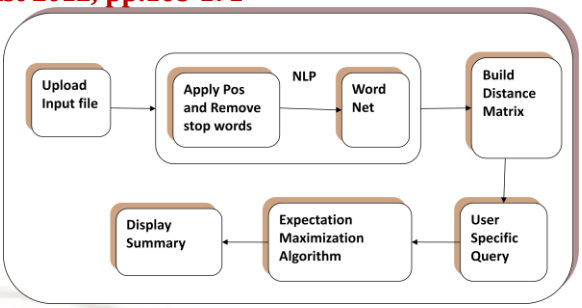

Figure1. Phases of text analysis

This association (edge weight) is greater than or equal to the cluster threshold value which is taken as input from user.

2. NLP Parser Engine

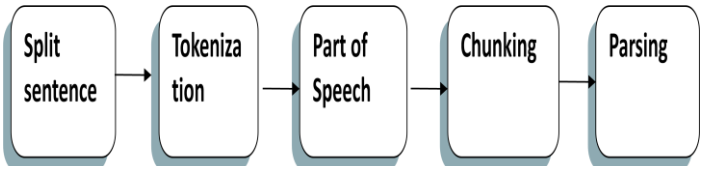

Figure 2. Phases of NLP Parser Engine

2.1 Split Sentence

Recognizing the end of a sentence is not an easy task for a computer. The input data is split into separate sentences by the new line character and converted into the array of paragraphs by using split method. This can be done by treating each of the characters '.', '!', '?' as separator rather than definite end-of-sentence markers.

2.2 Tokenization

It will separate the input text into separate tokens. The text can be separated into tokens. Punctuation marks, spaces and word terminators are the word breaking characters.

2.3 Part Of Speech tagger

POS tagger is applied for grammatical semantics. Part-of-speech tagging is the process which is applied after tokenization. The input to a tagging algorithm is a string of words of a natural language sentence. The output is a single best POS tag for each word. Part-of-speech tags are

NN: Noun DT: Determiner
VBN: Verb, past participle IN: Preposition
CC- coordinating conjunction :
NNP: Proper noun
DD: Common determinant
VBD: Verb, past tense
:

2.4 Chunker

Text chunking is dividing text into parts of words and forming groups like verb group, noun group. Text chunking divides the input text into such phrases and assigns a type such as NP for noun phrase, VP for verb

phrase, PP for prepositional phrase where the chunk borders are indicated by square brackets. Each word is assigned only one unique tag.

2.5 Parser

It generates the parse tree for given sentence. Parsing is converting a input sentence into a hierarchical structure that corresponds to the units of meaning in the sentence.

3. Built Distance matrix:

Lexical semantics begins with recognition that a word is a conventional association between lexicalized concepts.

4. User fires a query:

We have distance matrix as an input to the clustering algorithm. Once the distance matrix giving associatively of different sentences is ready, user can fire a query.

5. Clustering

5.1 Expectation Maximization Clustering Algorithm

The EM algorithm is an iterative procedure that consists of two alternating steps: an expectation step, followed by a maximization step. The expectation is with respect to the unknown underlying variables, using the current estimates of the parameters and conditioned upon the observations. The

Maximization step provides new estimates of the parameters. At each iteration, the estimated parameters provide an increase in

the maximum-likelihood (ML) function until a local maximum is achieved. Initialize k distribution parameters; each distribution parameter corresponds to a cluster center. Iterate between two steps global maximum of the likelihood function. This depends on the initial starting points.

Expectation step: assign points to clusters

$$w_k = \frac{\sum_i \Pr(x_i \in C_k)}{n}$$

$$\Pr(x_i \in C_k) = \Pr(x_i \mid C_k) \Big/ \sum_j \Pr(x_i \mid C_j)$$

Maximization step: estimate model parameters that maximize the likelihood for the given assignment of point

6. Build Document Graph

Depending upon weight, sentence similarity can be found out. Sentence with maximum weight can be considered in summarization

IV. RESULT

$$r_k = \frac{1}{n} \sum_{i=1}^{n} \frac{\Pr(x_i \in C_k)}{\sum_k \Pr(x_i \in C_j)}$$

Figure 3. Document Graph

Figure 3. Shows document graph(Similarity Matrix), where clustercomputing.txt is uploaded as input text file and showing node table data and document graph. Node table data contains node number and node (sentence), document graph shows similarity between each node with each other node.

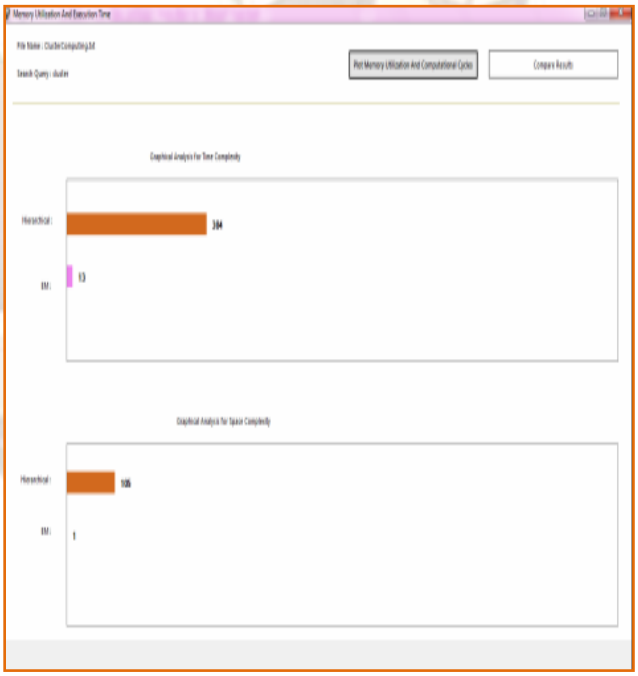

Figure 4.Graphical analysis with respect to time and space complexity

Graphical Analysis for Time Complexity in terms of Computational cycles
Hierarchical: 304, EM: 13
Graphical Analysis for Space Complexity in terms of Computational cycles
Hierarchical: 105, EM: 1

V. CONCLUSION

In this work we presented a structure-based technique to create query-specific summaries for text documents. The clustering algorithm plays an important role in result space and time consumption. Different clustering algorithms can be used on the same framework and their performance can be measured in term of quality of result and execution time. We show with a user survey that our approach performs better than other state of the art approaches.

VI. FUTURE WORK

In the future, we plan to extend our work to account for links between documents of the dataset. Also we will try to implement same algorithm in different applications.

VII.REFERENCES

1. "Count Data Modeling and Classification Using Finite Mixtures of Distributions", IEEE Transaction on Neural Networks.Vol.22, No.2, February 2011.
2." Inference for Probilistic Unsupervised Text clustering", 2005 IEEE.
3." Clustering Sentence-Level Text using a Novel Fuzzy Relational Clustering Algorithm", IEEE Transactions on Knowledge and Data Engineering 2011.
4." Synergy between Object Recognition and Image Segmentation Using
the Expectation-Maximization Algorithm", IEEE Transaction on Pattern Analysis and Machine Intelligence ,Vol. 31, No. 8, August 2009
5." A Context-Sensitive Clustering Technique Algorithm Based on Graph-Cut Initialization and Expectation-Maximization", IEEE Geosciences and Remote Sensing Letters", Vol. 5, No. 1, January 2008.
6." Constrained Text Co-clustering with Supervisedand Unsupervised Constraints", IEEE Transaction on Knowledge and Data Engineering2012.