International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

CLUSTERING COMPARISON FOR SENTENCE EXTRACTION-A PROPOSED APPROACH

Pooja Zalte, Radha Ghodake, Prof.Rokade P.P.

SND College of Engineering and RC,Yeola,DistNashik,MaharashtraState, India

ABSTRACT:

World Wide Web is the largest source of information. Huge amount of data is present on the Web. We propose a method to create query specific sentences extraction by identifying the most query-relevant fragments and combining them using the semantic associations within the document is discussed. Extraction is the process of automatically creating a compressed version of a given text that provides useful information for the user.

Keywords: agglomerative hierarchical clustering, clustered graph extraction, nearest neighbor, threshold summarization

Introduction:

Extraction is the process of automatically creating a compressed version of a given text that provides useful information for the user. In particular, first a structure is added to the documents in the preprocessing stage and converts them to document graphs. Then, the best extracted sentences will be computed by calculating the top spanning trees on the document graphs. Efficient [1,2,3].Current sentence extraction methods usually represent documents as a term document matrix and perform clustering algorithm on it. Although these clustering methods can group the sentences satisfactorily, it is still hard for people to capture the meanings of the documents since there is no satisfactory interpretation for each sentence cluster.The main aim of this work is to combine both approaches of document clustering and query dependent sentence extraction. This mainly includes applying different clustering algorithms on a text document. Create a weighted document graph of the resulting graph based on the keywords. And obtain the optimal tree to get the extracted sentences. The performance of the result using different clustering techniques will be analyzed and the optimal approach will be suggested.[4]

METHODOLOGY

Need

World Wide Web is a huge collection of data of different file formats. With the coming of the information revolution, electronic documents are becoming a principle media of business and academic information. In order to fully utilize these on-line documents effectively, it is crucial to be able to extract the gist of these documents. It is not the case that a particular clustering algorithm is best suited for clustering of documents of different file formats.Having a Text Summarization system would thus be immensely useful in serving this need. In order to generate a summary, we have to identify the most important pieces of information from the document, omitting irrelevant information and minimizing details, and assembling them into a compact Coherent report. A particular algorithm is best suited for query dependent text document summarization. As every document we can convert into text, this strategy is much needful for the end users[3,5,6]

Problem Definition

“To find best suited query dependent clustering algorithm for text document summarization “

The purpose of this project is to suggest better query dependent clustering algorithm for text document summarization. Our present task aims at developing a query dependant single-document summarizer using Nearest Neighbour clustering and Agglomerative Hierarchical clustering techniques. We hope it will add another dimension towards solving the seemingly complex task of document summarization and presentation of the gist of documents.

Goal

The goal is to use nearest neighbor, agglomerative hierarchical clustering algorithm for query dependent clustering of nodes in text document, and finding query dependent summary. The summary will be compared and best algorithm will be suggested for query dependent clustering using different clustering techniques. This technique will help end users to prepare query dependent summary of text document s in which they are interested.

- The proposed work will be mainly focused on summarization of text files (i.e. .txt)
- The proposed work will be limited to clustering of text files of standard files related to the topic popular amongst researchers will be used.
- Only hierarchical clustering and nearest neighbor method clustering are considered for generating cluster based graph.
- Standard performance evaluation metrics will be used to validate performance

Literature Review

'A System for Query-Specific Document Summarization', by Ramakrishna Varadarajan, Vangelis Hristidis.

In this work a structure-based technique is presented to create query-specific summaries for text documents. In particular, the document graph of a document is created; to represent the hidden semantic structure of the document and then perform keyword proximity search on this graph. It is shown in the paper that with a user survey that our approach performs better than other state of the art approaches. Furthermore, feasibility of the approach with a performance evaluation is shown at last.

'An Incremental Summary Generation System', by C RavindranathChowdary P Sreenivasa Kumar

This paper deals with updating the available extractive summary in the scenario where the initial documents used for summarization are not accessible. The proposed algorithm updates the available summary as and when a new document is made available to the system. It will replace the sentences from the present summary with the sentences from the new document. Before replacement, scoring model will check for the best pair of sentences (one from the present summary and another from the newly available document) to be swapped..

'Using Lexical Chains for Text Summarization', by Regina Barzilay and Michael Elhadad

In this paper empirical results on the identification of strong chains and of significant sentences are presented. Preliminary results indicate that quality indicative summaries are produced. Pending problems are identified. Plans to address these short-comings are briefly presented.

'Automatic Text Summarization', by Mohamed Abdel Fattah, and Fuji Ren

In this paper, investigates the use of genetic algorithm GA), mathematical regression (MR), for automatic text summarization task. This new approach is applied on a sample of 100 English religious articles. The approach results outperform the baseline approach results. The approaches have been used the feature extraction criteria which gives researchers opportunity to use many varieties of these features based on the used language and the text type. Some text features are language dependent like positive and negative keywords while some other features are language independent.

'Analyzing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes ', by ParulAgarwal, M. AfsharAlam, RanjitBiswas

Clustering ,an important technique of data mining groups similar objects together and identifies the cluster to which each object of the domain being studied belongs to. The paper provides in depth explanation of implementation adopted for k-pragna, an agglomerative hierarchical clustering technique for categorical attributes.

'Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection', by Jackie CK Cheung

The paper proposes a summarization content selection framework, based on clustering, and applied this framework to evaluative domain. In this framework, clusters represent groupings of related information that are represented or covered by a unit of text in the output summary. Using the p-median problem from facility location theory as the clustering paradigm; detailed several methods of reducing the content selection problem to a p-median problem. These methods can select not only the features to be realized in the summary, but also the strategy with which to express them. More work is needed to specify the parameters to the resultant p-median problems, especially in the more general versions of the reduction which allow any combination of features and strategies to express the content.

'Multi-topic based Query-oriented Summarization', by Jie Tang,Limin Yao, and Dewei Chen

This paper, tries to break limitations of the existing methods and study a new setup of the problem of multi-topic based query-oriented summarization.

More specifically, this paper proposed two strategies to incorporate the query information into a probabilistic model. Experimental results on two different genres of data show that our proposed approach can effectively extract a multi-topic summary from a document collection and the summarization performance is better than baseline methods. The approach is quite general and can be applied to many other mining tasks, for example product opinion analysis and question answering.

'Automatic Document Summarization by Sentence Extraction', by R.M.Aliguliyev

This paper represents generic summarization method that generates documents summaries by clustering and extracting silent sentences from the source document.

The generic summarization method that extracts the most relevance sentences from the source document to form a summary is proposed. The proposed method is based on clustering of sentences. The specificity of this approach is that the generated summary can obtain the main contents of different topics as many as possible and reduce its redundancy at the same time. By adopting a new cluster analysis algorithm, he determined the different topics in the document.

'Sentence Compression as a Component of a Multi-Document Summarization System', by D.M. Zajic, B. Dorr, J. Lin, R. Schwartz

They incorporated a multi-document sentence trimmer into a feature based summarization system. In this paper trimming is used to preprocess documents and create multiple partially trimmed sentences as alternatives for the original sentence. The count of trimming operations done is then used as a feature in the sentence ranker.

'The Proceedings of the 2006 document understanding conference at HLT/NAACL , New York, NY, by Conroy, J., Schlesinger, J., O'Leary, D., & Goldstein, J.

In the submission to DUC , modified approach for sentence splitting and sentence trimming is mentioned. Use of POS tagger is removed for the sentence splitting and instead chooses a conservative sentence trimming strategy which relies on a list of function words. The trimming is very conservative and the error analysis showed an error rate of less than 3% that is less than three percent of the input sentences were made ungrammatical by this trimming task.

SYSTEM ANALYSIS

Scope of the Project

This project can compare the query dependent summary generated by two clustering algorithms. Here we have considered nearest neighbouring clustering algorithm and agglomerative hierarchical clustering algorithm for summary comparison. As every file format can be converted into text file. These two algorithms are applied on text file. Nodes in text file i.e. contents in every newline are clustered and query dependent summary can be generated. This summary is then compared by using several comparative factors such as time complexity, space complexity, length of summary, quality of summary etc.[7]

Requirement Specifications

Requirements are the desired characteristics of the software being developed. The first activity in most projects is the identification and documentation of the requirements. Requirements cover both requirements engineering (identification, analysis and capture) and requirements management (managing change, creating and maintaining agreement with customers, trace ability and metrics).

The development of large, complex systems presents many challenges to systems engineers. Foremost among these is the ability to ensure that the final system satisfies the needs of users and provide for easy maintenance and enhancement of these systems during their deployed lifetime. These systems often change and evolve throughout their life cycle. This makes it difficult to track the implemented system against the original and evolving user requirements.

Requirements establish an understanding of user's need and also provide the final yardstick against which implementation success is measured. Various studies have shown that roughly half of the application errors can be traced to requirement errors and deficiencies. Thorough documentation and properly managing requirements are the keys to developing quality applications. By allowing project teams to define and document requirement data including user defined attributes, priority, status, acceptance criteria and traceability, detection and correction of missing, contradictory or inadequately defined requirements can be done.

The following requirements and constraints were considered during the requirement analysis phase.

To find the better algorithm for clustering we have to take text as input file. If any other file format is there, it is firstly converted into text format. Then clustering of nodes in text document and query dependent summarization is done.

Product performance requirements

Input file must be text file. As the size of the text document changes, performance of the algorithms also changes .So if file is larger, we should have better hardware facilities.

Hardware Requirements

Processor	: Pentium IV or higher.
Ram	: Minimum 256 MB.
Hard Disk	: 40 GB.
Input device	: Standard Keyboard and Mouse.
Output device	: VGA and High Resolution Monitor.

Software Requirements

Software components required for building the Project are:

Operating System : WINDOWS XP or above

Techniques : Microsoft Visual Studio 2008 (.NET Framework 3.5
Internet Explorer 6.0

Feasibility Study

"Not everything imaginable is feasible!"

Therefore it is necessary to evaluate feasibility of project at the earliest stage.

The software feasibility has 3 solid dimensions:

- **Technology:**

Technical feasibility is study of functions, performance, and constraints that may affect the ability to achieve an acceptable system. This project is technically feasible to implement. The user does not require any extra hardware or any higher-end technology. The software can execute on a single client machine operating on a WINDOWS XP or a higher version of Operating System.

- **Finance:**

Financial feasibility is the evaluation of the development cost weighed against the ultimate income or benefits derived from the developed system. The resources that are required for the system can be available easily. The system is developed basically for study purpose so economical feasibility is not a major issue.

This project is financially feasible because the software does not require any extra hardware or any additional supporting technology which in turn adds no extra cost to the software. Thus the cost is only for the development. Thus the project is financially feasible.

- **Resources:**

The organization that wishes to implement this system requires only a single or multiple machines. Thus no additional resources are required to implement the system. Thus the software is also resource feasible.

SYSTEM MODELING

Unified Modeling Language is a general purpose visual modeling language that is used to specify, visualize, construct and document the artifacts of the software system. It captures decision and understanding about the system that must be constructed. It is used to understand, design, browse, configure, maintain, and control information about such systems.

UML gives standard way to write systems blue prints covering conceptual thins, such as business processes and system functions, as well as concrete things, such as classes written in other programming languages, database schemas, and reusable software components.

Conceptual Model of UML consists of:

1. UML's basic building blocks.
2. Rules that dictate how these blocks may be put together.
3. Some common mechanisms that apply throughout the language.

Use Case Diagrams

Use Case Diagrams are useful for modeling the dynamic aspects of the systems. Use case diagrams are central to modeling the behavior of the system, a subsystem or a class. They show a set of use cases and actors and their relationships.

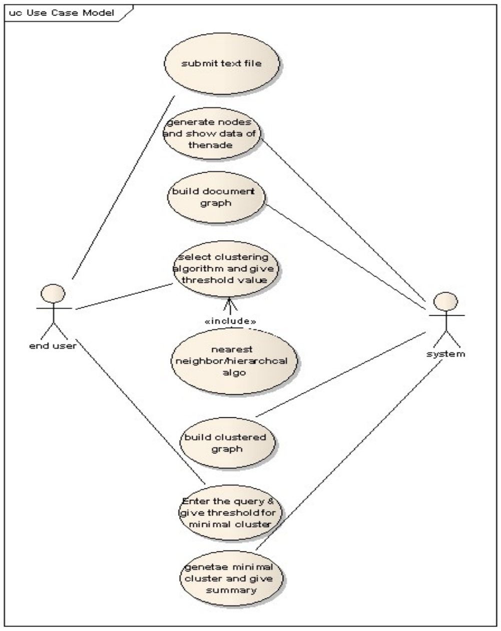
uc Use Case Model
submit text file
generate nodes and show data of thenade
build document graph
select clustering algorithm and give threshold value
«include»
nearest neighbor/hierarchcal algo
build clustered graph
Enter the query & give threshold for minimal cluster
genetae minimal cluster and give summary
end user
system

Figure.1 Use Case Diagram

Activity Diagram

Activity diagrams show the flow from activity to activity within a system. Activity diagrams address the dynamic view of the system.

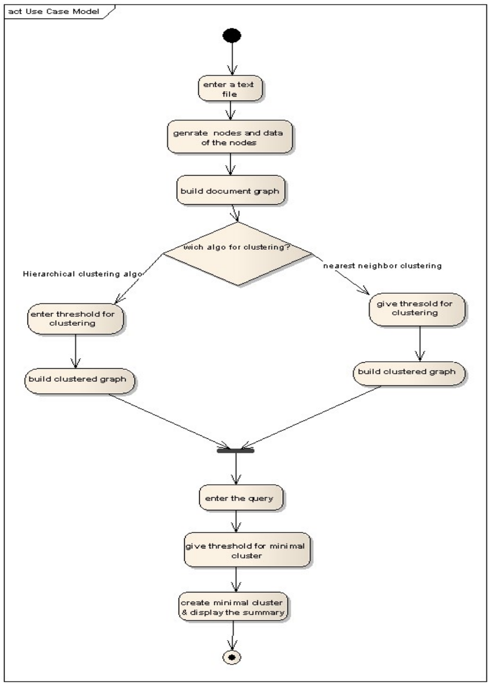
act Use Case Model
enter a text file
genrate nodes and data of the nodes
build document graph
wich algo for clustering?
Hierarchical clustering algo
nearest neighbor clustering
enter threshold for clustering
give thresold for clustering
build clustered graph
build clustered graph
enter the query
give threshold for minimal cluster
create minimal cluster & display the summary

Figure 2 Activity Diagram

Software Architecture

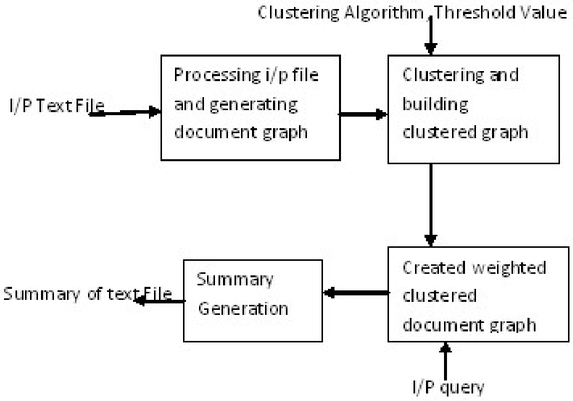
Clustering Algorithm ,Threshold Value
I/P Text File
Processing i/p file and generating document graph
Clustering and building clustered graph
Created weighted clustered document graph
I/P query
Summary Generation
Summary of text File

Figure 3. Architecture Diagram

Figure 3 shows the architecture diagram of the system. As shown in figure there are four main blocks : a block for uploading and processing text file and making document graph, a block for clustering and making clustered graph, a block for making weighted clustered document graph., the last block for generating summary for fired query.

Block 1: Processing input file and generating document graph:

This block is needed to accept the text file only. It is responsible to upload text file, to process the file i.e. to form nodes for every newline contents. It is also responsible for generating weight from each node to very other node

Block 2: Clustering node and building clustered graph:

This block is responsible for choosing a clustering algorithm out of two. It also accepts the threshold, so that can check the similarity between the clusters up to that level. It is responsible for making clusters.[11,13]

Block 3: Creating weighted document clustered graph:

This block is responsible to accept the fired query. It is responsible to check the similarities between the query a contents and the contents in the clusters. It then build weighted clustered document graph.

Block 4.Summary generation:

This block is responsible for generating the summary of the clusters we formed, as a response for fired query. It generated the minimal clusters and after finding the weight of the node for fired query, it gives top most summaries.

CONCLUSION:

In this paper, we proposed a method for summarization comparison using Nearest neighbor clustering and agglomerative Hierarchical Clustering Algorithm.

REFERENCES:

1. Ramakrishna Varadarajan, Vangelis Hristidis,"A System for Query-Specific Document Summarization",2013.
2. C. RavindranathChowdary P Sreenivasa Kumar "An Incremental Summary Generation System",2015.
3. Regina Barzilay and Michael Elhadad ,"Using Lexical Chains for Text Summarization", 2017.
4. Mohamed Abdel Fattah, and Fuji Ren ,"utomatic Text Summarization",2019.
5. ParulAgarwal, M. AfsharAlam, RanjitBiswas ,"Analyzing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes ",2021.
6. Jackie CK Cheung ,"Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection" .