An algorithm of word indexing model for document summarization based on perspective of document

Meha Shah[1], Chetna Chand[2]

[1] *Gujarat Technological University , Kalol Institute of Engineering , Kalol, Gujarat , India*
[2]*Assistant Professor, Gujarat Technological University, Kalol Institute of Engineering , Kalol, Gujarat , India*

ABSTRACT

Natural language processing (NLP) is an area of computer science, artificial intelligence, and computational linguistics connected with the communications between computers and natural languages. There are many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation. Document summarization is a part of it. Many different classes of such process based on machine learning are developed. In researches earlier document summarization mostly use the similarity between sentences in the document to extract the most significant sentences. The documents as well as the sentences are indexed using traditional term indexing measures, which do not take the context into consideration. The resulting indexing weights are used to compute the sentence similarity matrix. The proposed sentence similarity measure has been used with the baseline graph-based ranking models for sentence extraction.

Keyword: *- Data mining, Document Summarization, Text mining, Stemming, Sentence Similarity, Context Similarity.*

1. Introduction

Existing models for archive outline generally utilize the closeness between sentences in the report to separate the most notable sentences. The reports and in addition the sentences are recorded utilizing customary term indexing measures, which don't contemplate the setting. Information retrieval has become eminent paradigm for people working in IT industry. It not only provides information in the required format but also analysis and provides only summarized data to the standards provided by its user. A summary of a document is valuable since it can give an idea of the original document in a shorter period of time. A person who reads will check whether or not to read the complete document after going through the summary. For example, readers first read the abstract of a scientific article before reading the complete paper.

1.1 Document Summarization

Because of the constraints in regular dialect preparing innovation, abstractive methodologies are limited to particular spaces. Interestingly, extractive methodologies usually select sentences that contain the most huge ideas in the records. These methodologies have a tendency to be more down to earth. As of late different viable sentence highlights have been proposed for extractive synopsis, Such as signature word, occasion and sentence importance.

Two Summary Construction Methods are connected initial one is Abstractive strategy where outlines produce created content from the vital parts of the archives and second is Extractive Method where synopses recognize vital segments of the content and utilize them in the synopsis as they seem to be.

1.2 SENTENCE SIMILARITY AND WORD INDEXING

Sentences are grouping of words and characters are grouped to form texts. The probability and statistics, a Bernoulli model of randomness generate a finite sequence of random numbers. so it is a discrete-time stochastic process that takes only two values, canonically 0 and 1. They all have the same Bernoulli distribution. Much of what can be said about the Bernoulli process can also be generalized to more than two outcomes (such as the process for a six-sided die); this generalization is known as the Bernoulli scheme. However, the significance level as well as the ratio of average lexical association between the target summary and original document is much higher for the Bernoulli measure as compared to the MI measure. Thus, the proposed Bernoulli measure is a better fit for H_2.

1.3 CONTEXT BASED WORD INDEXING

Given the lexical association measure between two terms in a document from hypothesis H_2, the next task is to calculate the context sensitive indexing weight of each term in a document using hypothesis H_3. A graph -based iterative algorithm is used to find the context sensitive indexing weight of each term. Given a document D_i, a document graph G is built. Let G = (V,E) be an undirected graph to reflect the relationships between the terms in the document Di. $V = \{V_j | 1 \leq j \leq\_ |V|\}$ denotes the set of vertices, where each vertex is a term appearing in the document. E is a matrix of dimensions $|V| \times |V|$. Each edge $e_{jk} \varepsilon E$ corresponds to the lexical association value between the terms corresponding to the vertices v_j and v_k. The lexical association between the same terms is set to 0.

1.4 MODEL ON CONCEPTBASED MINING

For the proposed idea based digging model for web archive's content grouping, a crude content record is given as the information. Record taken is only a standard doc that has been composed with characters, extraordinary images and numbers as plain information in English. Every archive has unmistakable sentence limitations. Every sentence in the archive is stamped dully and might have one or more checked verb contention development. The measure of named data is absolutely reliant on the data present in the sentence. The sentence contained numerous stamped verb contention arrangement involves numerous verbs associated with their contentions. The named verb contention structures are analyzed by the idea construct mining model in light of sentence and web record levels.

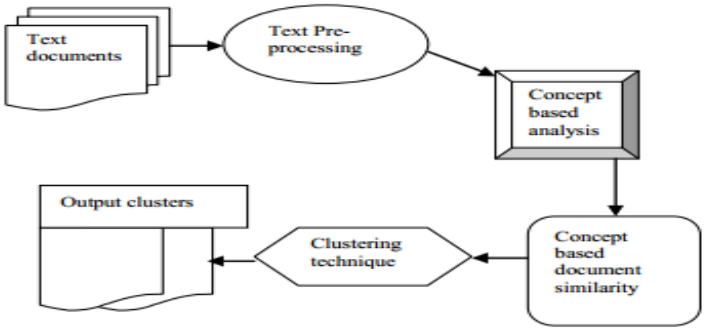


FIG: 1 MODEL FOR CONCEPT BASED ANALYSIS OF DATA

2. Literature Review and Motivation

The main goal of a summary is to present the main ideas in a document/set of documents in a short and readable paragraph. Summaries can be produced either from a single document or many documents. Summarization can also be specific to the information needs of the user, thus called "query-biased" summarization.

2.1 Context-Based Word Indexing Model for Document Summarization

The main goal of a summary is to present the main ideas in a document/set of documents in a short and readable paragraph. Summaries can be produced either from a single document or many documents. Summarization can also be specific to the information needs of the user, thus called "query-biased" summarization. For instance, the QCS system (query, cluster, and summarize) retrieves relevant documents in response to a query, clusters these documents by topic and produces a summary for each cluster.

2.1.1 CONCEPTS AND ALGORITHMS

- Exploring lexical association for text Summarization
- Bernoulli Model of Randomness: Derivation of the Term Association Metric.
- Context-Based Word Indexing
- Sentence Similarity Using the Context-Based Indexing

2.2 Context-Based Similarity Analysis for Document Summarization

The previous studies on the sentence extraction is mainly based on the text summarization task in which it uses a graph-based algorithm to calculate the saliency of each sentence in a document and the most salient sentences are extracted to build the document summary.

2.2.1 BERNOULLI MODEL OF RANDOMNESS

A Bernoulli process is a sequence of independent identically distributed Bernoulli trials. It is verified by two statistical test and thus holds trust for its measures used for further process.

2.2.2 CONTEXT BASED WORD INDEXING

It's a process that indexes words processed and matched after random selection from the group. The documents specified by the user are taken into consideration.

2.2.3 SENTENCE SIMILARITY USING THE CONTEXT-BASED INDEXING

This process is used to count weighted index. This indexing weight is used to count the similarity between any two sentences.

2.3 A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model

To make the text clustering more consistent, in our work, we plan to present a Conceptual Rule Mining On Text clusters to evaluate the more related and influential sentences contributing to the document topic.

2.3.1 Models

- Web Document Based Text Clustering using the Concept Based Mining Model
- Concept Based Similarity Measure for Web Document Text Clustering

The performance of the proposed web documents based text clustering using the concept based mining model is measured in terms of

i) Sentence similarity ratio

ii) Clustering efficiency

iii) Sentence Contributory rate

2.3.1.1 Result

No. of Sentences	*Sentence Similarity Ratio*	
	Proposed Web Document Text Clustering	*Existing doc. Clustering*
10	12	8
20	15	11
30	20	14
40	28	17
50	36	21

FIG: 1 NO OF SENTENCES VS. SENTENCE SIMILARITY RATIO

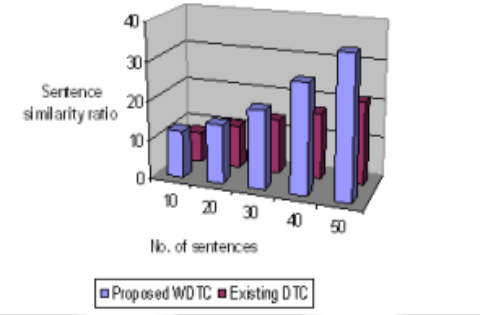


FIG: 2 GRAPH PRESENTATION OF ABOVE STATED TABLE

2.4 System for document summarization using graphs in text mining

The summarization concept is mainly started on the principle of index of books. In books when person want to search particular topic he or she will refer index of book and then that point will be retrieved in the less time. Their approach is related to NLP in which both the query and sentences are represented with word vectors. [3], this approach suffers from the shortcoming that it merely considers lexical elements (words) in the documents, and ignores semantic relations among sentences.

2.4.1 DOCUMENT SUMMARIZATION PROCESS

- Query Summarization Characteristic
- Document Summarization

These systems first rank all the sentences in the original document set and then select the most salient sentences to compose summaries for a good coverage of the concepts. For the purpose of creating more concise and fluent summaries, some intensive post-processing approaches are also appended on the extracted sentences.

2.4.2 Results

Query Keywords	No of file contains keyword	Output time In second	Score
Network	11	7	11.5044016
Soft	1	6	100.744885
Software	11	8	3.6043922
Computer	22	9	3.9982438
System	24	6	3.3116584

Table: 1 Score of input query

2.5 LexPageRank: Prestige in Multi-Document Text Summarization

Multi document extractive summarization relies on the concept of sentence centrality to identify the most important sentences in a document. We are now considering an approach for computing sentence importance based on the concept of eigenvector centrality (prestige) that we call LexPageRank. a sentence connectivity matrix is constructed based on cosine similarity.

2.5.1 Concept of centrality

The comparison is performed based on 2 centrality concepts for Prestige-based sentence centrality

- Degree centrality
- Eigenvector centrality and LexPageRank

The corresponding similarity graph is generated to find the centroid. Comparison between centroid generate the given results

2.5.1.1 RESULTS:

Policy Code	ROUGE-1 (unigram)	ROUGE-2 (bigram)	ROUGE-W (LCS)
degree0.5T0.1	0.38304	0.09204	0.13275
degree1T0.1	0.38188	0.09430	0.13284
lpr2T0.1	0.38079	0.08971	0.12984
lpr1.5T0.1	0.37873	0.09068	0.13032
lpr0.5T0.1	0.37842	0.08972	0.13121
lpr1T0.1	0.37700	0.09174	0.13096
C0.5	0.37672	0.09233	0.13230
lpr1T0.2	0.37667	0.09115	0.13234
lpr0.5T0.2	0.37482	0.09160	0.13220
C1	0.37464	0.09210	0.13071
lpr1T0.3	0.37448	0.08767	0.13302
degree0.5T0.2	0.37432	0.09124	0.13185
lpr0.5T0.3	0.37362	0.08981	0.13173
degree2T0.1	0.37338	0.08799	0.12980
degree1.5T0.1	0.37324	0.08803	0.12983
degree0.5T0.3	0.37096	0.09197	0.13236
lpr1.5T0.2	0.37058	0.08658	0.12965
C1.5	0.36885	0.08765	0.12747
lead-based	0.36859	0.08669	0.13196
lpr1.5T0.3	0.36849	0.08455	0.13111
lpr2T0.3	0.36737	0.08182	0.13040
lpr2T0.2	0.36737	0.08264	0.12891
C2	0.36710	0.08696	0.12682
degree1T0.2	0.36653	0.08572	0.13011
degree1T0.3	0.36517	0.08870	0.13046
degree1.5T0.3	0.35500	0.08014	0.12828
degree1.5T0.2	0.35200	0.07572	0.12484
degree2T0.3	0.34337	0.07576	0.12523
degree2T0.2	0.34333	0.07167	0.12302
random	0.32381	0.05285	0.11623

Table 2: Results for Task 2

Task 4b			
lpr1.5T0.1	0.40639	0.12419	0.13445
degree2T0.1	0.40572	0.12421	0.13293
lpr2T0.1	0.40529	0.12530	0.13346
C1.5	0.40344	0.12824	0.13023
degree1.5T0.1	0.40190	0.12407	0.13314
C2	0.39997	0.12367	0.12873
degree2T0.3	0.39911	0.11913	0.12998
lpr2T0.3	0.39859	0.11744	0.12924
lpr1.5T0.3	0.39858	0.11737	0.13044
lpr1.5T0.2	0.39819	0.12228	0.12989
lpr2T0.2	0.39763	0.12114	0.12924
degree2T0.2	0.39752	0.12352	0.12958
lpr1T0.1	0.39552	0.12045	0.13304
degree1.5T0.3	0.39538	0.11515	0.12879
lpr1T0.2	0.39492	0.12056	0.13061
C1	0.39388	0.12301	0.12805
degree1.5T0.2	0.39386	0.12018	0.12945
lpr1T0.3	0.39053	0.11500	0.13044
degree1T0.1	0.39039	0.11918	0.13113
degree1T0.2	0.38973	0.11722	0.12793
degree1T0.3	0.38658	0.11452	0.12780
lpr0.5T0.1	0.38374	0.11331	0.12954
lpr0.5T0.2	0.38201	0.11201	0.12757
degree0.5T0.2	0.38029	0.11335	0.12780
degree0.5T0.1	0.38011	0.11320	0.12921
C0.5	0.37601	0.11123	0.12605
lpr0.5T0.3	0.37525	0.11115	0.12898
degree0.5T0.3	0.37455	0.11307	0.12857
random	0.37339	0.09225	0.12205
lead-based	0.35872	0.10241	0.12496

Table 3: Results for Task 4

2.6 Weighted consensus multi-document summarization

Multi-document summarization is a fundamental tool for document understanding and has received much attention recently. Given a collection of documents, a variety of summarization methods based on different strategies have been proposed to extract the most important sentences from the original documents. Experimental results on DUC2002 and DUC2004 data sets demonstrate the performance improvement by aggregating multiple summarization systems, and our proposed weighted consensus summarization method outperforms other combination methods.

2.6.1 WEIGHTED CONSENSUS SUMMARIZATION (WCS)

- Notations
- Optimization-based weighted consensus summarization.

3. Proposed Work

Document summarization for text mining depending upon the similarity between documents. This approach also incurs a unique approach to process salient features of documents to process and extract related words based on effective summarization. Chart based positioning calculations are basically a method for choosing the significance of a vertex inside a diagram, in light of data drawn from the diagram structure. Therefore, we present new formulae for chart based positioning that consider edge weights when figuring the score connected with a vertex in the diagram.

3.1 Sentence Based Concept Analysis:

To examine every concept at the sentence level, a novel concept-based frequency assess, called the conceptual term frequency ctf (Conceptual term frequency) is computed. The ctf is the number of concept c happened in verb

argument structures of sentence S. The concept c, which normally materializes in diverse verb argument structures of the similar sentence S, has the prime job of contributing to the significance of S.

3.2 Concept based similarity Measure:

The similarity measure of the sentence and the terms are identified in a sentence and document level. The experimental evaluation tests aimed at comparing the existing efficient concept based mining model for enhancing text clustering with the proposed web document based text clustering using the concept based mining model.

3.3 Document preprocess

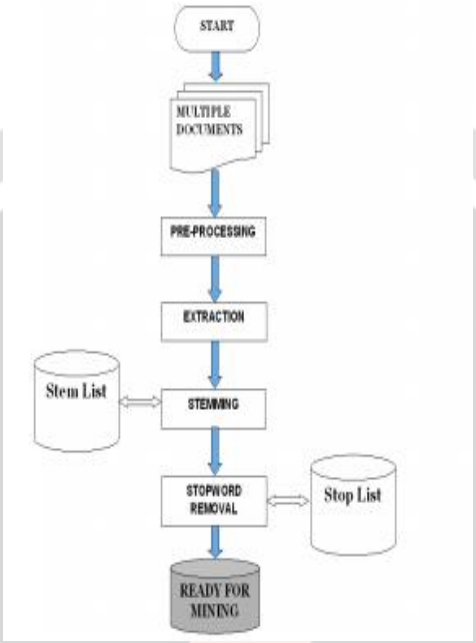


Fig: 3 Document preprocessing

This process consists of the following stages. First the data should pass the process of lexical analysis. It is a process of comparing all the terms with its lexical equivalents and whichever is found match, will be processed for the next stage. Elimination of stop words is the next stage that removes all the text and characters like dot, comma and question marks etc. Most frequently used words in English are useless in Text mining. Such words are called Stop words. Stop words are language specific functional words which carry no information. It may be of the following types such as pronouns, prepositions, conjunctions. Our system uses the SMART stop word list. Stemming is the next process towards accurate result processing. Stemming means elimination of synonyms and related words. The basic function of both the methods – stemming and lemmatizing is similar.

3.4 Term index process

Statistical weight estimation process is applied with term and its count values. Term weight estimation is performed with Term Frequency (TF) and Inverse Document Frequency (IDF) values. Context sensitive index model uses the term weights for term index process. Latent semantic analysis is applied to estimate relationship values.

3.5 Semantic index process

Ontology is a repository that maintains the concept term relationships. Semantic weights are estimated using concept relations. Synonym, hypernym and meronym relationships are used in the concept analysis. Context sensitive index model uses the semantic weight values for index process.

3.6 Document summarization

Lexical association between terms is used to produce context sensitive weight. Weights are used to compute the sentence similarity matrix. The sentence similarity measure is used with the baseline graph -based ranking models for sentence extraction. Document summary is prepared with sentence similarity values.

3.7 Document classification

Document classification is carried out to assign document category values. Term weight and semantic weights are used for the classification process. Context sensitive index is used for the document classification process. Sentence similarity is used in classification process.

3.8 PROPOSED ALGORITHM

Find documents to summarize

Input: copy and paste the data that has to be processed

Expected Output: Summary of documents

Step 1: Start

Step 2: For each Document

Step 3: Stopping method to remove additional symbols

Step 4: Stemming method to group similar meaning words

Step 5: Using Stemming algorithm removes blank space and extract keywords sentences using wordnet dictionary

End For each

Step 6: For each generated output using stemming algorithm

Sentences index generated by Bernoulli model of randomness.

Context based sentence similarity indexing

Now, Use Context based word indexing on the generated output to create the summarization of text;

$$\textbf{do}\begin{cases} E \leftarrow 0 \\ \textbf{for} j \leftarrow 1 \textbf{to} |S| \\ \quad \textbf{do}\begin{cases} memoWt[v_j] \leftarrow indexWt[v_j] \\ indexWt[v_j] \leftarrow \mu \cdot \sum_{\forall k \neq j} indexWt[v_k] \cdot \tilde{E_{kj}} \\ + \frac{1-\mu}{|V|} \\ E \leftarrow E + (indexWt[v_j] - memoWt[v_j])^2 \end{cases} \\ E \leftarrow \sqrt{E} \end{cases}$$

return $indexWt$

End for each

Go to next file and repeat above algorithm;

End

4. Implementation and Results

The overall summary generated here is depending upon the gist of the document as well as the importance of words filtered.

- The results are as follows:

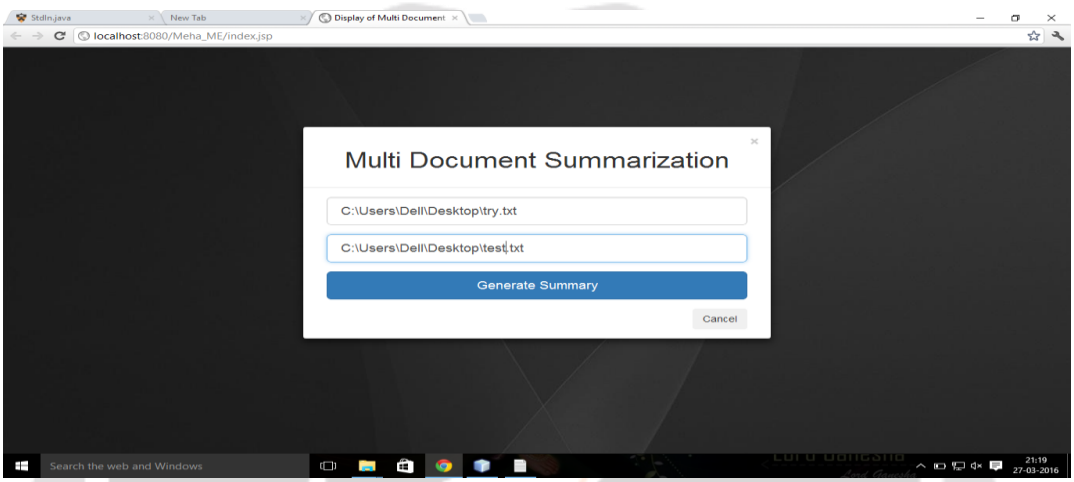


Fig:4 Input of docs name with path screen

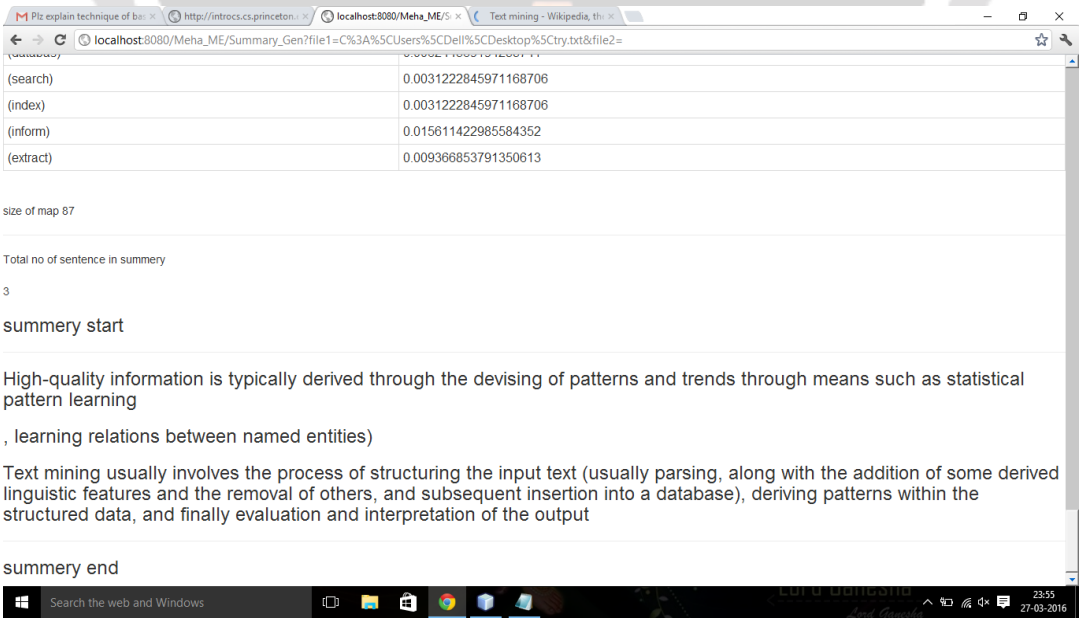
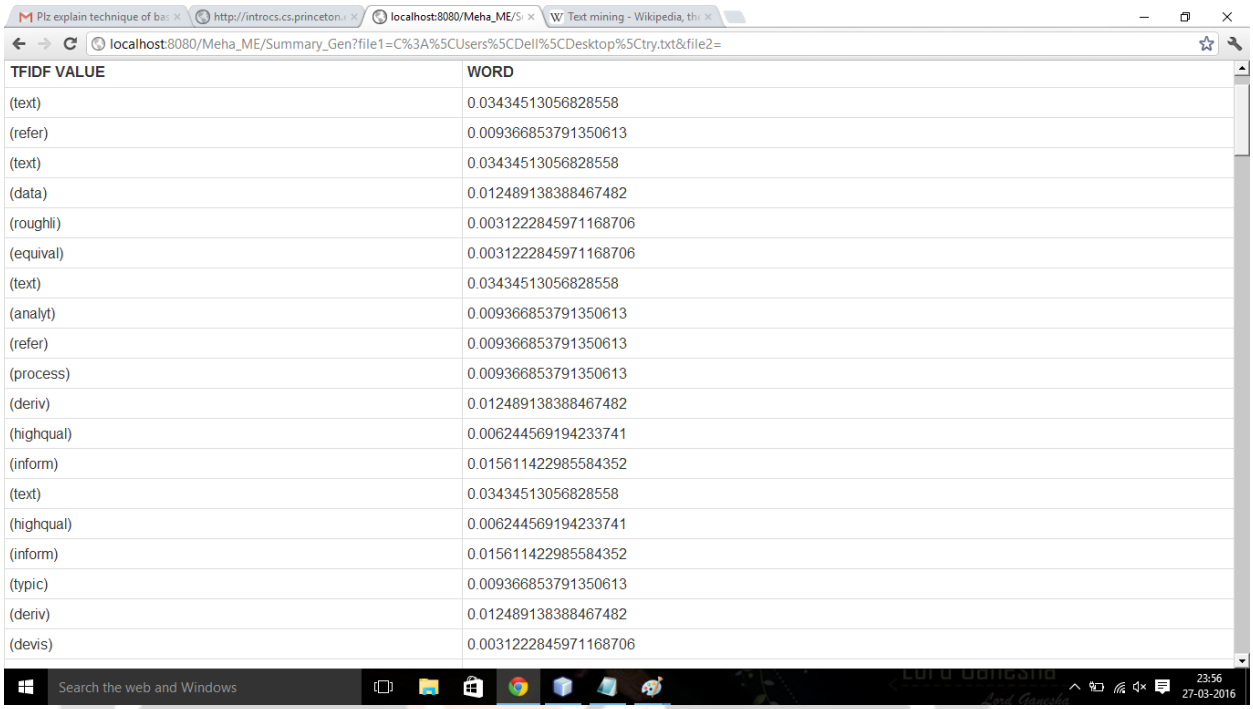
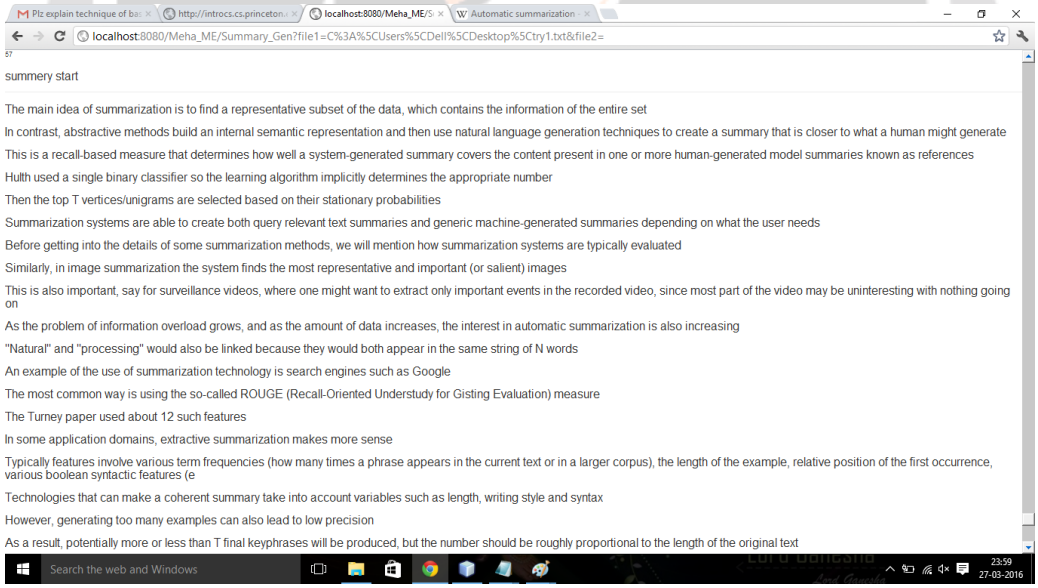


Fig:5 summary generated after process for a small text file



TFIDF VALUE	WORD
(text)	0.03434513056828558
(refer)	0.009366853791350613
(text)	0.03434513056828558
(data)	0.012489138388467482
(roughli)	0.0031222845971168706
(equival)	0.0031222845971168706
(text)	0.03434513056828558
(analyt)	0.009366853791350613
(refer)	0.009366853791350613
(process)	0.009366853791350613
(deriv)	0.012489138388467482
(highqual)	0.006244569194233741
(inform)	0.015611422985584352
(text)	0.03434513056828558
(highqual)	0.006244569194233741
(inform)	0.015611422985584352
(typic)	0.009366853791350613
(deriv)	0.012489138388467482
(devis)	0.0031222845971168706

Fig:6 Words list finalized after process



summery start

The main idea of summarization is to find a representative subset of the data, which contains the information of the entire set

In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate

This is a recall-based measure that determines how well a system-generated summary covers the content present in one or more human-generated model summaries known as references

Hulth used a single binary classifier so the learning algorithm implicitly determines the appropriate number

Then the top T vertices/unigrams are selected based on their stationary probabilities

Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs

Before getting into the details of some summarization methods, we will mention how summarization systems are typically evaluated

Similarly, in image summarization the system finds the most representative and important (or salient) images

This is also important, say for surveillance videos, where one might want to extract only important events in the recorded video, since most part of the video may be uninteresting with nothing going on

As the problem of information overload grows, and as the amount of data increases, the interest in automatic summarization is also increasing

"Natural" and "processing" would also be linked because they would both appear in the same string of N words

An example of the use of summarization technology is search engines such as Google

The most common way is using the so-called ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure

The Turney paper used about 12 such features

In some application domains, extractive summarization makes more sense

Typically features involve various term frequencies (how many times a phrase appears in the current text or in a larger corpus), the length of the example, relative position of the first occurrence, various boolean syntactic features (e

Technologies that can make a coherent summary take into account variables such as length, writing style and syntax

However, generating too many examples can also lead to low precision

As a result, potentially more or less than T final keyphrases will be produced, but the number should be roughly proportional to the length of the original text

Fig:7 summary generated after processing a large text file

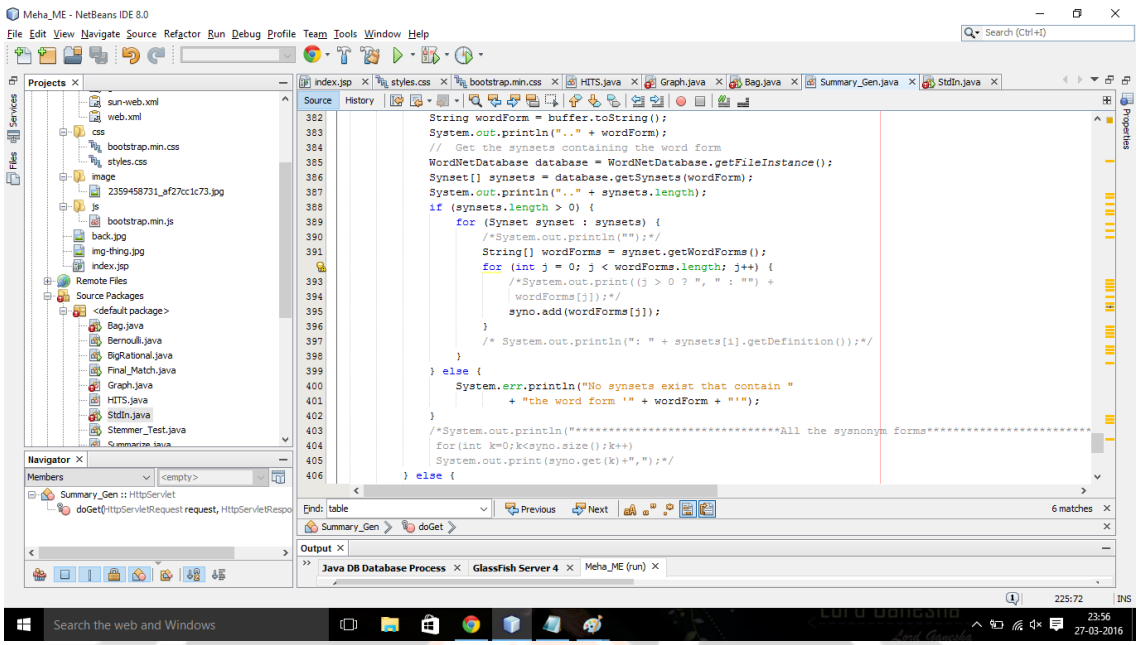


Fig:8 Sample code of proposed algorithm

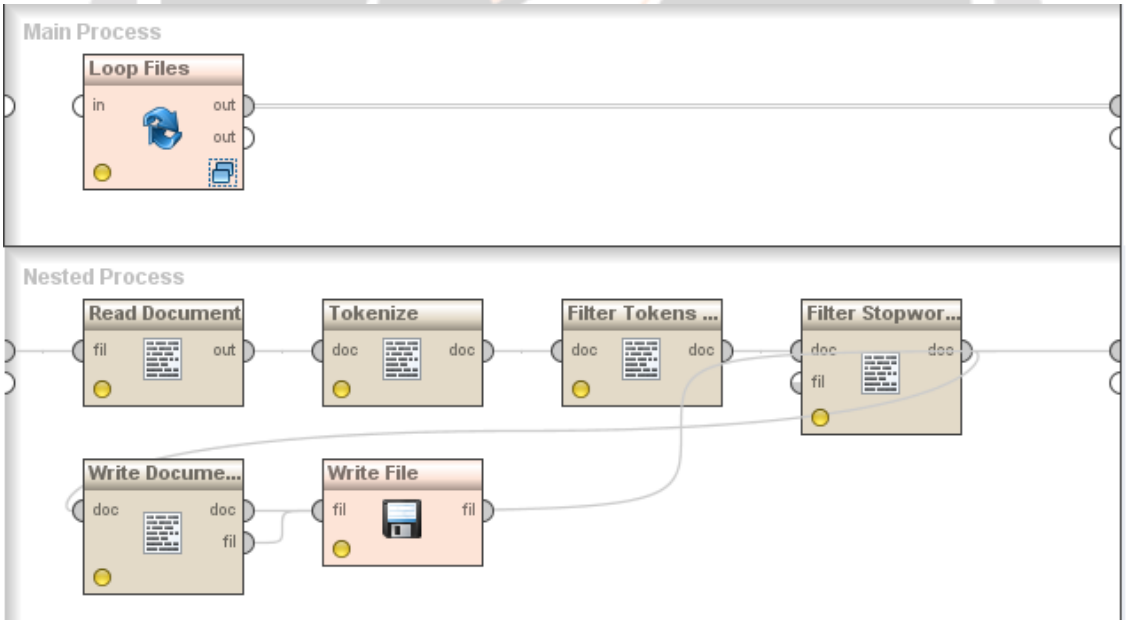
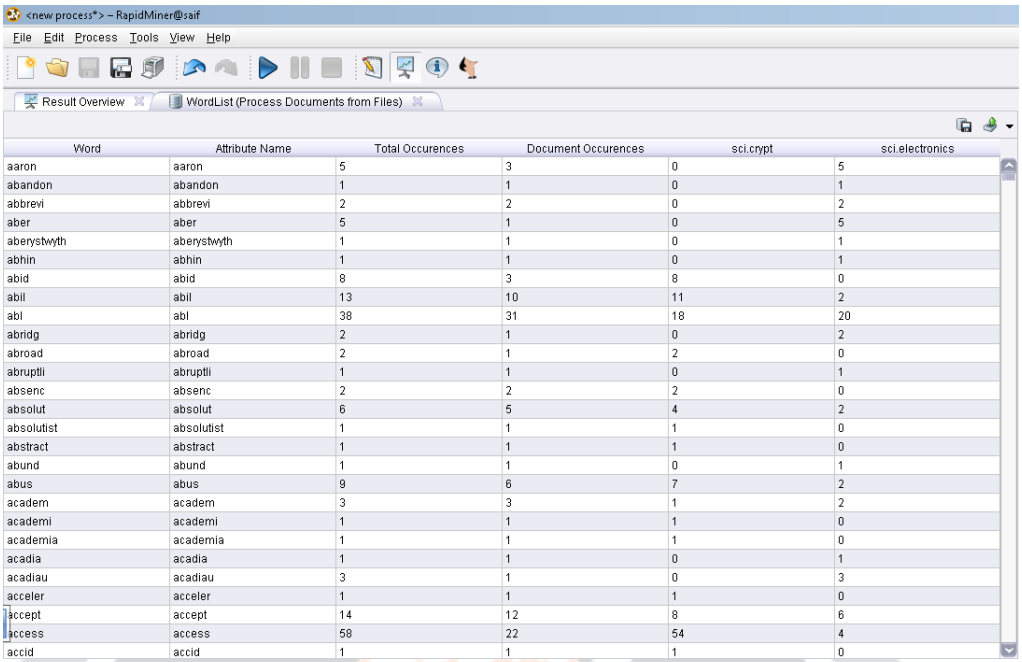


Fig:9 Implementation with Rapid Miner



Word	Attribute Name	Total Occurences	Document Occurences	sci.crypt	sci.electronics
aaron	aaron	5	3	0	5
abandon	abandon	1	1	0	1
abbrevi	abbrevi	2	2	0	2
aber	aber	5	1	0	5
aberystwyth	aberystwyth	1	1	0	1
abhin	abhin	1	1	0	1
abid	abid	8	3	8	0
abil	abil	13	10	11	2
abl	abl	38	31	18	20
abridg	abridg	2	1	0	2
abroad	abroad	2	1	2	0
abruptli	abruptli	1	1	0	1
absenc	absenc	2	2	2	0
absolut	absolut	6	5	4	2
absolutist	absolutist	1	1	1	0
abstract	abstract	1	1	1	0
abund	abund	1	1	0	1
abus	abus	9	6	7	2
academ	academ	3	3	1	2
academi	academi	1	1	1	0
academia	academia	1	1	1	0
acadia	acadia	1	1	0	1
acadiau	acadiau	3	1	0	3
acceler	acceler	1	1	1	0
accept	accept	14	12	8	6
access	access	58	22	54	4
accid	accid	1	1	1	0

Fig:10 Context based similarity analysis with co-occurrence matrix

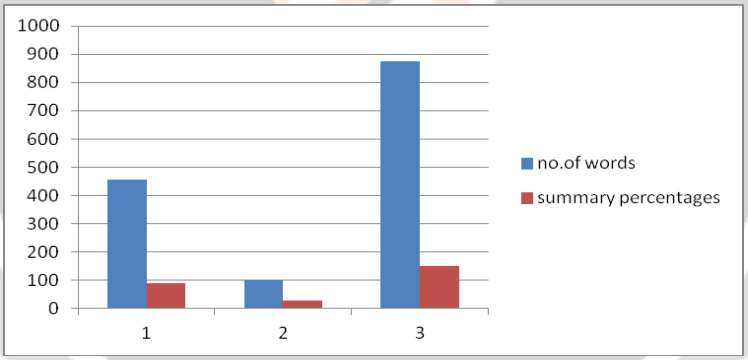


Fig:11 Comparison of different docs and summary in words

6. REFERENCES

[1] A Context-Based Word Indexing Model for Document Summarization Pawan Goyal, Laxmidhar Behera,Senior Member, IEEE, and Thomas Martin McGinnity,Senior Member, IEEE

[2] Context-Based Similarity Analysis for Document Summarization 1 S.Prabha, 2 Dr.K.Duraiswamy, 3 B.Priyanga 1 Associate Professor, Department of Information technology K.S.Rangasamy College of Technology, Tiruchengode – 637215, Tamil Nadu, India 2 Dean Academic

[3] Document Summarization and Classification using Concept and Context Similarity Analysis J.Arun 1, C. Gunavathi M.E 2 PG scholar, K.S.Rangasamy College of technology, Tiruchengode, Tamil Nadu, India

[4]A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model V.M.Navaneethakumar 1, Dr.C.Chandrasekar 2 1 Assistant Professor, Department of Computer Applications, K.S.R College of Engineering, Tiruchengode, Tamilnadu, India

[5]SYSTEM FOR DOCUMENT SUMMARIZATION USING GRAPHS IN TEXT MINING Prashant D. Joshi 1, M. S. Bewoor 2, S. H. Patil 3 1 Deptt. of Computer Engineering, Researcher, BharatiVidyapeeth University, Pune, India.

[6]L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," technical report, Stanford Digital Library Technologies Project, http://citeseer.ist.psu.edu/page98pagerank.html, 1998.

[7] H.P. Luhn, "The Automatic Creation of Literature Abstracts,"IBM J. Research and Development,vol. 2, pp. 159-165, http://dx.doi.org/10.1147/rd.22.0159, Apr. 1958.

[8] C.-Y. Lin and E. Hovy, "Identifying Topics by Position,"Proc. Fifth Conf. Applied Natural Language Processing,pp. 283-290, http://dx. doi.org/10.3115/974557.974599, 1997.

[9] E. Hovy and C.-Y. Lin, "Automated Text Summarization and the Summarist system,"Proc. Workshop Held at Baltimore, Maryland (TIPSTER '98), pp. 197-214, http://dx.doi.org/10.3115/1119089. 1119121, 1998.

[10] R. Katragadda, P. Pingali, and V. Varma, "Sentence Position Revisited: A Robust Light-Weight Update Summarization 'Base-line' Algorithm," Proc. Third Int'l Workshop Cross Lingual Informa-tion Access: Addressing the Information Need of Multilingual Societies, pp. 46-52, http://portal.acm.org/citation.cfm?id=1572433.1572440, 2009.

[11] Y. Ouyang, W. Li, Q. Lu, and R. Zhang, "A Study on Position Information in Document Summarization," Proc. 23rd Int'l Conf.Computational Linguistics: Posters,pp. 919-927, http://portal.acm.org/citation.cfm?id=1944566.1944672, 2010.

[12] H.P. Edmundson, "New Methods in Automatic Extracting,"J. ACM, vol. 16, pp. 264-285, http://doi.acm.org/10.1145/321510.321519, Apr. 1969.