A NOVEL APPROACH FOR NOVELTY DETECTION USING EXTRACTIVE TEXT SUMMARIZATION

Ishuka, Komal Kumar Bhatia, Sushil Kumar
M.Tech Scholar, Pofessor, Assistant Professor
(Department of Computer Engineering),
J.C Bose University of Science and Technology, YMCA, Faridabad, India

ABSTRACT:

As the information available over the internet is increasing day by day which result in information overload. Internet contains a huge number of pages related to the query. It extracts the irrelevant pages also and therefore it becomes cumbersome for a user to select the desired document. To resolve this problem we proposed a Novelty detection technique with text summarization. This work will provide the relevant and novel results to the user query.

Automatic Summarization is one of the most attractive, and interesting topics for researchers to find out the relevant text from the huge bunch of documents. The key aim of this Automatic Text Summarization is to provide the users to get their data in the minimum period of time. Novelty Detection is a much needed requirement for good classification system. Here, an idea has been proposed to generate the novel document from the given sets of documents. Here we are merging two techniques extractive text summarization and novelty detection. The proposed idea provides enhanced results when compared with already existing work.

Keywords: Extractive summary, Single document, Novelty and Novelty Detection.

1.INTRODUCTION:

WWW is a vast source of hyperlinked and heterogeneous information including text,audio, video and metadata. Nowadays, the use of Internet is massive and the electronic data is growing exponentially which needs to be process and manage. Sometimes, it is very strenuous to find the relevant information from the huge data repository.

Search engine are defined as the “programs” which are used to look for the documents related to specific keywords and response is generated in terms of list consisting of the related keywords. Search engines are referred to as software which provides the high end quality results or information.

As Internet users do not have that much of required time to read all the relevant and irrelevant documents and they are in search of compact and exact information. So Summarization technique is considered as one of the most important technique as it compresses the original text into a short version and let the user to quickly understand the integral volume of information they are looking for.

Summarization has drawn a substantial interest from researchers and software developers as they provide a solution to the information overload problem for the users in the digital era of the World Wide Web.

Automatic Summarization is one of the most attractive, and interesting topics for researchers to find out the relevant text from the huge bunch of documents. The key aim of this Automatic Text Summarization is to provide the users to get their data in the minimum period of time.

The two fundamental techniques for Automatic Text Summarization are : Abstractive and Extractive Summarization.

Extractive Text Summarization v/s Abstractive Text Summarization:

Extractive system produces the summary by obtaining the main sentences from the original documents. They just provide the replica of the original sentences by gathering the highest ranked sentences. The very common problem associated with extractive techniques is that they suffer from inconsistencies, lack of cohesion and redundancy.

Abstractive system produces the summary by constructing the new sentences which are short and concise. Summary might have words and phrases that are not explicitly available in the main text.

Mono-lingual v/s Multi-lingual

Based on the different types of languages, summarization can be classified into two categories: Mono-lingual system and multi-lingual system. As the name suggest, mono-lingual system work only on one particular language, such as English. Whereas, multi-lingual system works on one or more than one language such as English, Spanish, Japanese etc.

Single-document v/s Multi-document

Depending on the number of documents classification can be categorized into: single document and multi document. Multi document means having multiple document of same or relevant topic for generating the summary.

Generic v/s Query-based:

Sometimes one summary can be used for solving the purpose of many different users. Hence these types summary are very important and independent to the subject of document. They are termed are generic based summary. In contrast when the users want some specific information from the document is query based summary.

Indicative v/s Informative:

They are considered as the next classification of the summarization system. As the name suggest indicative summary only indicates the user with the main idea of the text, whereas informative system provides the concise information of the original and can be used as a substitute in place of original document

2. RELATED WORK:-

The study of automated summarization started 40 years ago. In that era, simple document features such as word frequency and word positions were harnessed to create document summary. Since then many single documents approaches have been created [1]

.Harsha Dave and Shree Jaiswal presented a novel approach to generate abstractive summary from extractive summary using WordNet Ontology[2].The experimental result shows the generated summary in a well-compressed , grammatically correct and human readable-format.

Anusha Bagalkotkar and Shivam Pandey presented a novel technique for generating the summarization of domain-specific text from a single web document by using statistical NLP techniques on the text in a reference corpus and on the web documents [3]. The summarizer proposed contributed a summary based on the Calculated Sentence Weight, the rank of a sentence in the document's content, the number of terms and the number of words in a sentence, and using the term frequency in the input corpus.

GlorianYapinus and Alva Erwin discussed the development of multi-document summarization for Indonesian documents by using hybrid abstractive-extractive summarization approach[4]. Multi-document summarization is a technique that is able to summarize multi-documents and present them in a one summary. Abstractive-Extractive summarization technique which is a combination of WordNet based text summarization and title word based text summarization. After the experiment, the research methodology successfully produced the well-compressed and readable summary with fast processing time.

Yungang Ma, Ji Wu presented the method for extractive multi-document summarization based on combining features of n-grams co-occurrences and dependency word pairs co-occurrences where Unigram is considered as the basic text unit, bigram and skip-bigram reflect the word sequential relationships between words[5].The co-occurrences of each feature reflect the common topics of multiple documents in different perspective.

Dahae Kim and Jee-Hyoung Lee contributed to a new multi-document summarization system by creating a synthetic document vector covering the whole documents based on Language Model[6].They experimented with DUC 2004 dataset provided by Document Understanding Conference and the output generated stated that the method summarizes the multiple documents effectively based on their core contents.

Prof .R. Nedunchelian proposed a multiple-document summarization system with user-interaction. He introduced the system that would extract the summary from multiple documents based on the document cluster-centroids, which is effectively the distribution of terms in the multiple documents in the cluster. This summarization technique is a cluster based, extractive summarization method, where passages are first clustered based on similarity, prior to the selection of passages that form the extractive summary of documents. The implementation is based on the MEAD extraction and redundancy based algorithm. MEAD extraction requires three features to compute the salience of the sentence. They are Centroid Value, Positional-Value and First –Sentence overlap. Timestamp are issued to maintain the chronological order of the sentences and hence a coherent and free-flowing summary can be generated[7].

3. PROPOSED DESIGN:

This segment talks about the suggested framework for the single text summarization document.

A. Input Documents:

Single document as an input will be used to draw the text from the article on any subject such as politics, economy, entertainment and sports. These documents have texts of many different sizes.

B. Text pre-processing:

Pre-processing is a very widely used method nowadays, in the field of computational linguistic as the quality of the obtained summary rely upon how efficient the text is represented. This phase further constitutes to tokenization, stop word removal, stemming.

Tokenization: In this section the task is to split streaming of text into words, phrases, symbols.

Stop word removal: Process of discarding the words that have no significance in a text. The stop words in the concatenated title are demolished based on the suggested stop words list.

Stemming:It is defined as the process of diminish the derived words to their word stem or root stem.

C. Term Frequency Text Summarization:

Tf-idf is used for term frequency-inverse document frequency, is a numerical statistic which intends to reflect how important a word is[8] to a document in a corpus. It is often used as weighting measure in information retrieval and text mining. The Tf-Idf value increases proportionally to the number of times a word appears in a document.

The term frequency is very important characteristic. TF(Term Frequency) results how many times the term appears in the document (usually a compression function such as square root or logarithm is applied) to calculate the term frequency. The term that identifies the sentence boundaries in a document is split into sentences and these sentences are nothing but the tokens.

Keyword Frequency:

These are the top high frequency words [8] in term sentence frequency. After removing the unwanted or noisy data calculate the frequency of each word. And the words which hold the highest frequency are termed as Keywords. The words score are chosen as keywords, based on this feature, any sentence in the document is scored by number of keywords it contains, where the sentence is given 0.1 score for each keyword.

IDF(Inverse Document Frequency):

The inverse document frequency determines how much information the word deliver, that is, whether the term is common or rare across all[9] the documents.

It is the logarithmically scaled inverse fraction of the documents which includes the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

Calculation of TF-IDF:

$(1+\log_{10}N)*IDF$

Where N=number of terms

We always take value of IDF=1

If value of N=0 then TF-IDF=0

ALGORITHM:

Step1: input a text document

Step2: perform the pre-processing on the original text document.

Step3: generate the term frequency list of the pre-processed document.

Step4: for all terms

4.1 calculate the TFIDF value.

Step5: for all TFIDF values

5.1 generate the sentences from the original text document having highest TF-IDF value terms.

Step6: generate a summary of the original text document having sentence generated by step 5.1.

Step7: Finally a summarized document is generated.

D. Novelty Detection:

A burdensome for sentence categorization and novelty detection ,reveal not only when the text is applicable to the user's particulars or facts, but also when [10] it contains something unique which the user has not seen before. It incorporates two quest that needs to be resolved. The very first is to recognize the relevant sentences (categorization) and the second is to identify the novel or new information from those relevant sentences (novelty detection).

As we know that the information is growing exponentially, sentence categorization and novelty detection has become one of the potential techniques for handling and organizing the data. The main intention of sentence categorization is to categorize the sentences into a fixed number of pre-defined groups. Sentence level novelty detection targets at detecting the novel information from a chronologically organized list of relevant sentences. Novel information means the sentences which contains latest content, and is typically defined in literature as the opposite of redundancy.

Novelty Detection has been recommended to reclaim the novel and yet new and relevant information, based on the specific topic defined by the user. It has been done at three different levels: event level, document level, sentence level. Identifying the unique sentences is more helpful for the user as there is a lot of unwanted information in one document, especially when the area is new. Therefore, we here cornerstone on sentence-level novelty detection.

Cosine Similarity:

Cosine similarity is the cosine of the angle between the two sentences which are represented as vectors which contain the frequency of words of the input sentences[11]. It is computed as follows:

$$Cosine\_similarity = \frac{\text{Sa. Sb}}{||\text{Sa}||||\text{Sb}||}$$

Cosine similarity, is thus a judgement of orientation not magnitude: two vectors with same orientation have a cosine similarity of 1, two vectors at 90° have a cosine similarity of 0, and two vectors diametrically opposed have a similarity of 1, independent of their magnitude [12]. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

4. EXPERIMENTAL AND RESULT ANALYSIS:

Example 1:

The peacock is the national bird of India. They have colourful feathers, two legs and a small beak. They are famous for their dance. When a peacock dances it spreads its feathers like a fan. It has a long shiny dark blue neck. Peacocks are mostly found in the fields they are very beautiful birds. The females are known as 'Peahen[1]. Their feathers are used for making jackets, purses etc. We can see them in a zoo.

a - The peacock is the national bird of India.
b - They have colourful feathers, two legs and a small beak.
c - They are famous for their dance.
d - When a peacock dances it spreads its feathers like a fan.
e - It has a long shiny dark blue neck.
f - Peacocks are mostly found in the fields.
g - They are very beautiful birds.
h - The females are known as peahen.
i - Their feathers are used for making jackets, purses etc.
j - We can see them in a zoo.

Terms	Term frequency for documents									
	A	B	C	D	E	F	G	H	I	J
peacock	1	0	0	1	0	1	0	0	0	0
nation	1	0	0	0	0	0	0	0	0	0
bird	1	0	0	0	0	0	1	0	0	0
india	1	0	0	0	0	0	0	0	0	0
colour	0	1	0	0	0	0	0	0	0	0
feather	0	1	0	1	0	0	0	0	1	0
two	0	1	0	0	0	0	0	0	0	0
leg	0	1	0	0	0	0	0	0	0	0
small	0	1	0	0	0	0	0	0	0	0
beak	0	1	0	0	0	0	0	0	0	0
famous	0	0	1	0	0	0	0	0	0	0
dance	0	0	1	1	0	0	0	0	0	0
spread	0	0	0	1	0	0	0	0	0	0
like	0	0	0	1	0	0	0	0	0	0
fan	0	0	0	1	0	0	0	0	0	0
long	0	0	0	0	1	0	0	0	0	0
shiny	0	0	0	0	1	0	0	0	0	0
dark	0	0	0	0	1	0	0	0	0	0
blue	0	0	0	0	1	0	0	0	0	0
neck	0	0	0	0	1	0	0	0	0	0
mostly	0	0	0	0	0	1	0	0	0	0
found	0	0	0	0	0	1	0	0	0	0
field	0	0	0	0	0	1	0	0	0	0
beauty	0	0	0	0	0	0	1	0	0	0
female	0	0	0	0	0	0	0	1	0	0
known	0	0	0	0	0	0	0	1	0	0
peahen	0	0	0	0	0	0	0	1	0	0
us	0	0	0	0	0	0	0	0	1	0
make	0	0	0	0	0	0	0	0	1	0
jacket	0	0	0	0	0	0	0	0	1	0
purse	0	0	0	0	0	0	0	0	1	0
etc	0	0	0	0	0	0	0	0	1	0
can	0	0	0	0	0	0	0	0	0	1
see	0	0	0	0	0	0	0	0	0	1
zoo	0	0	0	0	0	0	0	0	0	1

Combination pair	cosine value	Combination pair	cosine value
A – B	0	C - J	0
A – C	0	D - E	0
A – D	0.204	D - F	0.204
A – E	0	D - G	0
A – F	0.25	D - H	0
A – G	0.354	D - I	0.167
A – H	0	D - J	0
A – I	0	E - F	0
A – J	0	E - G	0
B – C	0	E - H	0
B – D	0.167	E - I	0
B – E	0	E - J	0
B – F	0	F - G	0
B – G	0	F - H	0
B – H	0	F - I	0
B – I	0.167	F - J	0
B – J	0	G - H	0
C – D	0.289	G - I	0
C – E	0	G - J	0
C – F	0	H - I	0
C – G	0	H - J	0
C – H	0	I - J	0
C – I	0		

Threshold = 0.04

Text Summary after applying novelty detection

When a peacock dances it spreads its feathers like a fan.
Peacocks are mostly found in the fields.
They are very beautiful birds.

Applying Proposed Technique

a - The peacock is the national bird of India.
b - They have colourful feathers, two legs and a small beak.
c - They are famous for their dance.
d - When a peacock dances it spreads its feathers like a fan.
e - It has a long shiny dark blue neck.
f - Peacocks are mostly found in the fields.
g - They are very beautiful birds.
h - The females are known as peahen.
i - Their feathers are used for making jackets, purses etc.
j - We can see them in a zoo.

term	term frequency	Inverse document frequency
peacock	3.0	2.09
nation	1.0	1.0
bird	2.0	1.69
india	1.0	1.0
colour	1.0	1.0
feather	3.0	2.09
two	1.0	1.0
leg	1.0	1.0
small	1.0	1.0
beak	1.0	1.0
famous	1.0	1.0
dance	2.0	1.69
spread	1.0	1.0
like	1.0	1.0
fan	1.0	1.0
long	1.0	1.0
shiny	1.0	1.0
dark	1.0	1.0
blue	1.0	1.0
neck	1.0	1.0
mostly	1.0	1.0
found	1.0	1.0
field	1.0	1.0
beauty	1.0	1.0
female	1.0	1.0
known	1.0	1.0
peahen	1.0	1.0
us	1.0	1.0
make	1.0	1.0
jacket	1.0	1.0
purse	1.0	1.0
etc	1.0	1.0
can	1.0	1.0
see	1.0	1.0
zoo	1.0	1.0

Term With Highest Tf-Idf weight - peacock

Sentences that contain this term
The peacock is the national bird of India.
They have colourful feathers, two legs and a small beak.
They are famous for their dance.
When a peacock dances it spreads its feathers like a fan.
It has a long shiny dark blue neck.
Peacocks are mostly found in the fields.
They are very beautiful birds.
The females are known as peahen.
Their feathers are used for making jackets, purses etc.
We can see them in a zoo.

Combination pair	Cosinevalue	Combination pair	cosine value
A – B	0	C - H	0
A – C	0	D - B	0
A – D	0.2	D - C	0.2
A – E	0	D - D	0
A – F	0.25	D - E	0
A – G	0.35	D - F	0.17
A – H	0	D - G	0
A – I	0	E - B	0
A – J	0	E - C	0
B – B	0	E - D	0
B – C	0.17	E - E	0
B – D	0	E - F	0
B – E	0	F - B	0
B – F	0	F - C	0
B – G	0	F - D	0
B – H	0.17	F - E	0
B – I	0	G - B	0
C – B	0.29	G - C	0
C – C	0	G - D	0
C – D	0	H - B	0
C – E	0	H - C	0
C – F	0	I - B	0
C – G	0		

Threshold = 0.04001059590337666

Text Summary after applying novelty detection

When a peacock dances it spreads its feathers like a fan.
Peacocks are mostly found in the fields.
They are very beautiful birds.

Example 2:

An elephant is the biggest living animal on land. It is quite huge in size. It is usually black or grey in colour. Elephants have four legs, a long trunk and two white tusks near their trunk. Apart from this, they have two big ears and a short tail. Elephants are vegetarian. They eat all kinds of plants especially bananas. They are quite social, intelligent and useful animals. They are used to carry logs of wood from one place to another. They are good swimmers.

a - An elephant is the biggest living animal on land.
b - It is quite huge in size.
c - It is usually black or grey in colour.
d - Elephants have four legs, a long trunk and two white tusks near their trunk.
e - Apart from this, they have two big ears and a short tail.
f - Elephants are vegetarian.
g - They eat all kinds of plants especially bananas.
h - They are quite social, intelligent and useful animals.
i - They are used to carry logs of wood from one place to another.
j - They are good swimmers.

Terms	Term frequency for documents									
	A	B	C	D	E	F	G	H	I	J
elephant	1	0	0	1	0	1	0	0	0	0
biggest	1	0	0	0	0	0	0	0	0	0
live	1	0	0	0	0	0	0	0	0	0
animal	1	0	0	0	0	0	0	1	0	0
land	1	0	0	0	0	0	0	0	0	0
quite	0	1	0	0	0	0	0	1	0	0
huge	0	1	0	0	0	0	0	0	0	0
size	0	1	0	0	0	0	0	0	0	0
usual	0	0	1	0	0	0	0	0	0	0
black	0	0	1	0	0	0	0	0	0	0
grei	0	0	1	0	0	0	0	0	0	0
colour	0	0	1	0	0	0	0	0	0	0
four	0	0	0	1	0	0	0	0	0	0
leg	0	0	0	1	0	0	0	0	0	0
long	0	0	0	1	0	0	0	0	0	0
trunk	0	0	0	2	0	0	0	0	0	0
two	0	0	0	1	1	0	0	0	0	0
white	0	0	0	1	0	0	0	0	0	0
tusk	0	0	0	1	0	0	0	0	0	0
near	0	0	0	1	0	0	0	0	0	0
apart	0	0	0	0	1	0	0	0	0	0
big	0	0	0	0	1	0	0	0	0	0
ear	0	0	0	0	1	0	0	0	0	0
short	0	0	0	0	1	0	0	0	0	0
tail	0	0	0	0	1	0	0	0	0	0
vegetarian	0	0	0	0	0	1	0	0	0	0
eat	0	0	0	0	0	0	1	0	0	0
kind	0	0	0	0	0	0	1	0	0	0
plant	0	0	0	0	0	0	1	0	0	0
especi	0	0	0	0	0	0	1	0	0	0
banana	0	0	0	0	0	0	1	0	0	0
social	0	0	0	0	0	0	0	1	0	0
intellig	0	0	0	0	0	0	0	1	0	0
use	0	0	0	0	0	0	0	1	0	0
us	0	0	0	0	0	0	0	0	1	0
carri	0	0	0	0	0	0	0	0	1	0
log	0	0	0	0	0	0	0	0	1	0
wood	0	0	0	0	0	0	0	0	1	0
one	0	0	0	0	0	0	0	0	1	0
place	0	0	0	0	0	0	0	0	1	0
anoth	0	0	0	0	0	0	0	0	1	0
good	0	0	0	0	0	0	0	0	0	1
swimmer	0	0	0	0	0	0	0	0	0	1

Combination pair	cosine value	Combination pair	cosine value
A – B	0	C - J	0
A – C	0	D - E	0.136
A – D	0.149	D - F	0.236
A – E	0	D - G	0
A – F	0.316	D - H	0
A – G	0	D - I	0
A – H	0.2	D - J	0
A – I	0	E - F	0
A – J	0	E - G	0
B – C	0	E - H	0
B – D	0	E - I	0
B – E	0	E - J	0
B – F	0	F - G	0
B – G	0	F - H	0
B – H	0.258	F - I	0
B – I	0	F - J	0
B – J	0	G - H	0
C – D	0	G - I	0
C – E	0	G - J	0
C – F	0	H - I	0
C – G	0	H - J	0
C – H	0	I - J	0
C – I	0		

Threshold = 0.029

Text Summary after applying novelty detection

Elephants have four legs, a long trunk and two white tusks near their trunk.
Elephants are vegetarian.
They are quite social, intelligent and useful animals.

Applying proposed technique

a - An elephant is the biggest living animal on land.
b - It is quite huge in size.
c - It is usually black or grey in colour.
d - Elephants have four legs, a long trunk and two white tusks near their trunk.
e - Apart from this, they have two big ears and a short tail.
f - Elephants are vegetarian.
g - They eat all kinds of plants especially bananas.
h - They are quite social, intelligent and useful animals.
i - They are used to carry logs of wood from one place to another.
j - They are good swimmers.

term	term frequency	Inverse document frequency
elephant	3.0	2.09
biggest	1.0	1.0
live	1.0	1.0
animal	2.0	1.69
land	1.0	1.0
quite	2.0	1.69
huge	1.0	1.0
size	1.0	1.0
usual	1.0	1.0
black	1.0	1.0
grey	1.0	1.0
colour	1.0	1.0
four	1.0	1.0
leg	1.0	1.0
long	1.0	1.0
trunk	2.0	1.69
two	2.0	1.69
white	1.0	1.0
tusk	1.0	1.0
near	1.0	1.0
apart	1.0	1.0
big	1.0	1.0
ear	1.0	1.0
short	1.0	1.0
tail	1.0	1.0
vegetarian	1.0	1.0
eat	1.0	1.0
kind	1.0	1.0
plant	1.0	1.0
especi	1.0	1.0
banana	1.0	1.0
social	1.0	1.0
intellig	1.0	1.0
use	1.0	1.0
us	1.0	1.0
carry	1.0	1.0
log	1.0	1.0
wood	1.0	1.0
one	1.0	1.0
place	1.0	1.0
another	1.0	1.0
good	1.0	1.0
swimmer	1.0	1.0

Term with Highest Tf-Idf weight - elephant

Text Summary after applying novelty detection
An elephant is the biggest living animal on land.
It is quite huge in size.
It is usually black or grey in colour.
Elephants have four legs, a long trunk and two white tusks near their trunk.
Apart from this, they have two big ears and a short tail.
Elephants are vegetarian.
They eat all kinds of plants especially bananas.
They are quite social, intelligent and useful animals.
They are used to carry logs of wood from one place to another.
They are good swimmers.

Combination pair	cosine value	Combination pair	cosine value
A – B	0	C - H	0
A – C	0	D - B	0.14
A – D	0.15	D - C	0.24
A – E	0	D - D	0
A – F	0.32	D - E	0
A – G	0	D - F	0
A – H	0.2	D - G	0
A – I	0	E - B	0
A – J	0	E - C	0
B – B	0	E - D	0
B – C	0	E - E	0
B – D	0	E - F	0
B – E	0	F - B	0
B – F	0	F - C	0
B – G	0.26	F - D	0
B – H	0	F - E	0
B – I	0	G - B	0
C – B	0	G - C	0
C – C	0	G - D	0
C – D	0	H - B	0
C – E	0	H - C	0
C – F	0	I - B	0
C – G	0		

Threshold = 0.02878406395883234

Text Summary after applying novelty detection

Elephants have four legs, a long trunk and two white tusks near their trunk.
Elephants are vegetarian.
They are quite social, intelligent and useful animals.

5.1 CONCLUSION:

Novelty Detection is the most essential activity to extricate the unique or novel sentences. It lessens the time to explore the query for the document. As Internet retrieves the ‘n’ number of pages associated to query it becomes clumsy for the user to get the desired results.

Text summarization also plays an important part in serving the user’s query. It précise the content of the document.

Therefore both the cosine similarity and text summarization had their own issues in imparting their conclusion.

Keeping in mind all the problems, we have proposed the new idea to rectify the issues. The proposed idea includes amalgamation of the two techniques: extractive text summarization and novelty detection.

We have presented the proposed technique with the better and the summarized outcomes then the already existed one’s.

So, the final conclusion is that the proposed design generates better results with less time and space complexity.

5.2 FUTURE WORK:

Text Summarization and Novelty Detection domain is huge. There are numerous techniques available for Text Summarization and Novelty Detection.

For the Future consideration we can perform with Cluster-based expertise and can evaluate the multiple documents. On the resultant clustered documents we will seek the Text Summarization.

References:

1. ”GlorianYapinus, Alva Erwin”, Automatic Multi-Documents Summarization for Indonesian Documents Using Hybrid Abstractive-Extractive Summarization Technique, ” 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE) Yogyakarta, Indonesia.

2. “ Harsha Dave and Shree Jaiswal “, Multiple Text Document Summarization System using Hybrid Summarization Technique, “ 2015 1st International Conference on Next Generation Computing Technologies (NGCT-2015), Dehradun, India, 4-5 September, 2015.

3. “Anusha Bagalkotkar and ShivamPande”, A Novel Technique for Efficient Text Document Summarization As a Service, “ 2013 Third International Conference on Advances in Computing and Communications”.

4. ”GlorianYapinus, Alva Erwin”, Automatic Multi-Document Summarization for Indonesian Documents Using Hybrid Abstractive-Extractive Summarization Technique , ” 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE) Yogyakarta, Indonesia.

5. Yungang Ma, Ji Wu”, Combining N-gram and Dependency Word Pair for Multi-Document Summarization, “ 2014 IEEE 17TH International Conference On Computational Science and Engineering.

6. “ Dahae Kim, Jee-Hyoung Lee”, Multi-Documents Summarization by Creating Synthetic Document Vector Based on Language Model, “ 2016 Joint 8th International Conference on Soft Computing and Intelligent Systems and 2016 17th International Symposium on Advanced Intelligent Systems.

7. “Prof. R. Nedunchelian”, Centroid Based Summarization of Multiple Documents Implemented using Timestamps, “ 2008 1st International Conference on Emerging Trends in Engineering and Technology (ICETET).

8. “Spark Jones, K” , Automatic Summarizing: factors and directions.” In Inderjeet Mani and Mark Marbury, editors “, Advances in Automatic Text Summarization.

9. Wikipedia.

10. “Makoto Hirohata, YousukeShinnaka, Koji Iwano and SadaokiFurui“, Sentences extraction-based presentation summarization techniques and evaluation metrices”, Department of Computer-Science, Tokyo Institute of Technology.

11. “Yi Zhang Flora S. Tsai, AgusTrisnajayaKwee”, Information Processing and Management:- Multilingual sentence categorization and Novelty Mining.

12. “ Ronald. T. Fernandez, David E. Losada”, Effective sentence retrieval based on Query-Independent Evidence.