Automatic Condensation of Electronic Document

Madhusudan M. Oza[1] Pratibha P. Dhundale[2] Vishal S. Baviskar[3] Falgun R. Gahane[4] Amardip R Sabale[5]

[1,2,3,4,5]Department of Computer Engineering

[1,2,3,4,5]SSBT's COET, Bambhori, Jalgaon. North Maharashtra University, Jalgaon (M.S.), India

Abstract— In today's world there are huge amount of data is available on Internet. Everyone is used to with retrieval of this data/info from different sites, applications in direct or indirect way. Many times it is necessary to summarize the articles or documents in short but also important format. Summarization via manual method is an older technique. Each human may have different summary however, some important aspect of text is always present in everyone's summary. In practical, humans have limitations of speed and time. It reduces the performance as per time of work and time of input articles are increasing. It is pretty good if same is done by Computer which has intelligence! Text Summarization is a way to generate a text, which contains the important portion of information of the original text or texts. Several techniques are generated depending upon many parameters to find the summary as the type, position and format of the sentences in an input text, formats of different words occurrence of a particular word in a text etc. The solution to be discussed is about to use Automatic Method for summarization.

Key words: TRM, Integer Array for Processing

I. Introduction

Summary is a text that is formed from one or more text, that contains a important portion of the information from the original text, and that is not more than half of the original text.[3]

This trouble-free definition derives three important properties that characterize automatic summarization [3]:

1) Summaries may be formed from a solo document or several documents
2) Summaries should protect vital information
3) Summaries should be short

'Text' here includes multimedia documents, on-line documents, hypertexts etc. of the several types of summary that have been famous, analytic summaries (that give an idea of what the text is about, with lack of actual content) and informative summery (that give some reduced version of the content) are frequently referenced. Extracts are summaries formed by reusing portions (sentences, words, etc.) of the key text, while abstracts are shaped by re-producing the extracted content.[1]

Need to note that, this solution is based on abstraction technique.

II. Proposed work

A. Objective:

Proposed system should work on Paragraphs for the Automatic Condensation of Electronics Document's Text. It should take all words as input. Then it will create weight for each word. This weight is based on occurrence of a word in whole text document. By using similarity function system should create value in between 0 to 1 for all possible pairs of words. Now consider each paragraph as node. Based on numerical similarity joined them to create task relationship map. Select a typical threshold value (min value) and select the paragraphs above the threshold value.

Now perform the text traversal using convenient techniques from paragraph to paragraph to select the important sentences. These sentences will logically combine for an output file.

B. Workflow:

Salton et al. employs the technique to produce intra document links between paragraphs of a document, and get a text relationship map (TRM) according to those links. Each node on the TRM indicates a paragraph and it is represented by a vector of weighted terms. If two nodes have high similarity then, a link is created between them, which is typically computed as the inner product between the vectors of the corresponding paragraphs [2].

Simply, two nodes, they are said to be semantically related if there is a link in them. to measure paragraphs importance, bushiness of a paragraph is defined. The bushiness of a paragraph is the number of links connecting it to other paragraphs. For example, the bushiness of P2 in Fig. 1 (finding bushy node) is 5 because it linked with P1, P3, P4, p5 and P6.

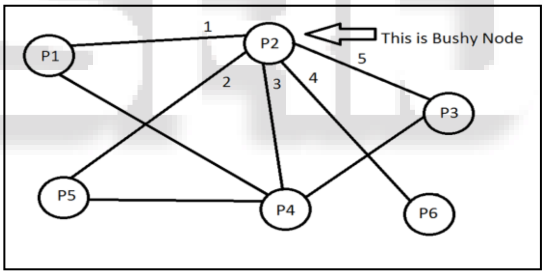


Fig. 1: Text Relationship Map (TRM)

A highly bushy node linked with many other nodes; simply, it has a lot of overlapping vocabulary with other nodes; thus, a highly bushy node is likely to discuss main topics that are covered in many other paragraphs.

As for text summarization, we proposed method to generate the summary according to the bushiness of paragraphs: global bushy path. It identifies paragraphs with high bushiness but traverse them in different text order. System focuses on Bushy Method.

C. Algorithm:

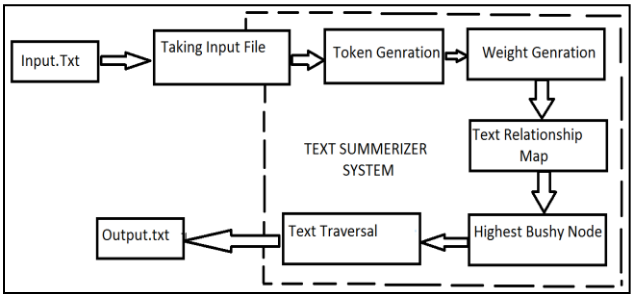


Fig. 2: Detailed Workflow

Above fig. 2, shows detailed workflow of text summarizer system. Following algorithm is a backbone of workflow of the system:

1) Take Input in txt format
2) Generate tokens for each paragraph i.e. separate each valid word from remaining document. This separate word is called as token.
3) Find occurrence of each word .i.e. how many time the 'word' occurred in whole text. (fig3)

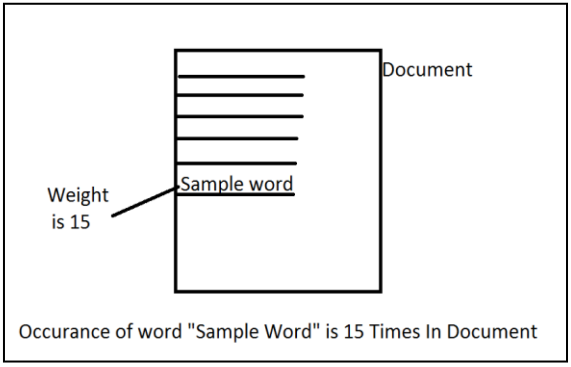


Fig. 3: Occurrence of Word

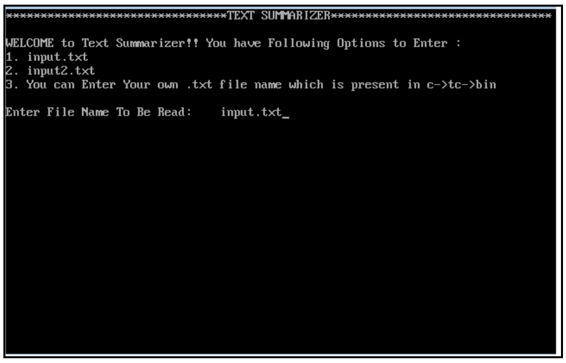
4) Generate Links between Nodes based on similarity of words in each node. This graphical presentation is called Text Relationship Map(TRM) (fig 1)
5) Find the highest bushy node by checking which node is having maximum links in TRM (fig 1)
6) Create a similarity function that checks total occurrence of a 'word' between nodes and compare this occurrence with threshold value. If, it is greater than threshold value then word is valid for consideration. Store this word in separate file with its occurrence. ex. If word "Develop" is have total occurrences in whole document is 57% and threshold value is 40% then "Develop" is a valid word to take in consideration.

   In this proposed solution 40% threshold value is used.
7) After storing all valid words as per above discussion, extract sentences from actual document which contains this valid words. Here first priority is to a highest bushy node then, remaining less bushy nodes i.e. priority to extract sentences from nodes is from highest bushy node to lowest bushy node. This step is called as Text Traversal. Out of the final summary 40% summary is taken from highest bushy node. Cause it contains most overlapping vocabulary. Remaining 60% summary is taken from all less bushy nodes.
8) After extracting sentences by using above step, rearrange sentences in proper order. This re-arrangement is required because in step 7 we are focusing on priority of bushy nodes to extract sentences but, the generated output must be in original sequence as provided in input text.
9) Put generated summary in output file.

III. RESULT AND DISCUSSION

This section contains sample snapshots of systems important processing and its explanation. These snapshots are from system developed by our team.

A. *Welcome Screen:*

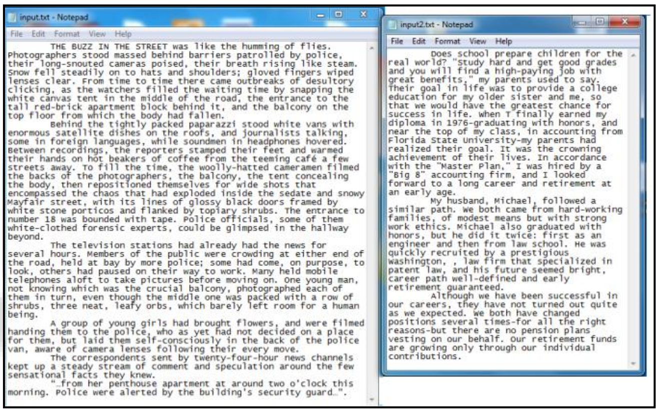


```
**********************************TEXT SUMMARIZER**********************************

WELCOME to Text Summarizer!! You have Following Options to Enter :
1. input.txt
2. input2.txt
3. You can Enter Your own .txt file name which is present in c->tc->bin

Enter File Name To Be Read:    input.txt_
```

Fig. 4: Welcome Screen

Above Fig. 4, is snapshot of Welcome screen. This screen is providing direction to user that how to use the software. Provided two options are input files named input.txt and input2.txt respectively. These two files are present in bin folder of TurboC++ directory. There is third option that user can add its own input file by providing name of that file. The file must be present in bin folder of TurboC++ directory.

B. *Input files:*



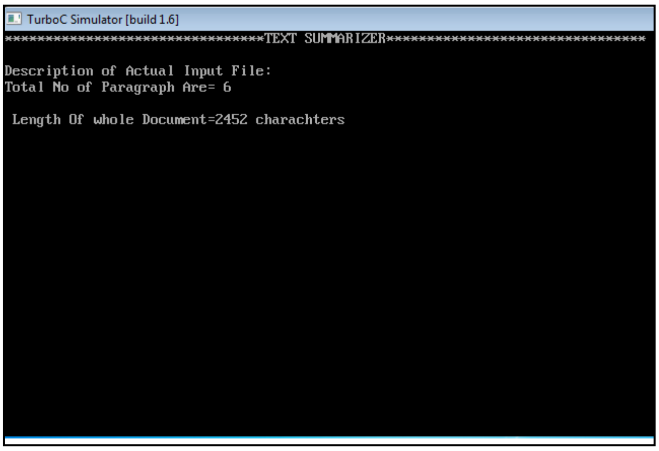
```
        THE BUZZ IN THE STREET was like the humming of flies.
Photographers stood massed behind barriers patrolled by police,
their long-snouted cameras poised, their breath rising like steam.
Snow fell steadily on to hats and shoulders; gloved fingers wiped
lenses clear. From time to time there came outbreaks of desultory
clicking, as the watchers filled the waiting time by snapping the
white canvas tent in the middle of the road, the entrance to the
tall red-brick apartment block behind it, and the balcony on the
top floor from which the body had fallen.
        Behind the tightly packed paparazzi stood white vans with
enormous satellite dishes on the roofs, and journalists talking,
some in foreign languages, while soundmen in headphones hovered.
Between recordings, the reporters stamped their feet and warmed
their hands on hot beakers of coffee from the teeming café a few
streets away. To fill the time, the woolly-hatted cameramen filmed
the backs of the photographers, the balcony, the tent concealing
the body, then repositioned themselves for wide shots that
encompassed the chaos that had exploded inside the sedate and snowy
Mayfair street, with its lines of glossy black doors framed by
white stone porticos and flanked by topiary shrubs. The entrance to
number 18 was bounded with tape. Police officials, some of them
white-clothed forensic experts, could be glimpsed in the hallway
beyond.
        The television stations had already had the news for
several hours. Members of the public were crowding at either end of
the road, held at bay by more police; some had come, on purpose, to
look, others had paused on their way to work. Many held mobile
telephones aloft to take pictures before moving on. One young man,
not knowing which was the crucial balcony, photographed each of
them in turn, even though the middle one was packed with a row of
shrubs, three neat, leafy orbs, which barely left room for a human
being.
        A group of young girls had brought flowers, and were filmed
handing them to the police, who as yet had not decided on a place
for them, but laid them self-consciously in the back of the police
van, aware of camera lenses following their every move.
        The correspondents sent by twenty-four-hour news channels
kept up a steady stream of comment and speculation around the few
sensational facts they knew.
        "…from her penthouse apartment at around two o'clock this
morning. Police were alerted by the building's security guard…".
```

```
        Does school prepare children for the
real world? "Study hard and get good grades
and you will find a high-paying job with
great benefits," my parents used to say.
Their goal in life was to provide a college
education for my older sister and me, so
that we would have the greatest chance for
success in life. When T finally earned my
diploma in 1976-graduating with honors, and
near the top of my class, in accounting from
Florida State University-my parents had
realized their goal. It was the crowning
achievement of their lives. In accordance
with the "Master Plan," I was hired by a
"Big 8" accounting firm, and I looked
forward to a long career and retirement at
an early age.
        My husband, Michael, followed a
similar path. We both came from hard-working
families, of modest means but with strong
work ethics. Michael also graduated with
honors, but he did it twice: first as an
engineer and then from law school. He was
quickly recruited by a prestigious
Washington, , law firm that specialized in
patent law, and his future seemed bright,
career path well-defined and early
retirement guaranteed.
        Although we have been successful in
our careers, they have not turned out quite
as we expected. We both have changed
positions several times-for all the right
reasons-but there are no pension plans
vesting on our behalf. Our retirement funds
are growing only through our individual
contributions.
```

Fig. 5: Two Sample Input Files

Above Fig. 5 is showing two input files respectively named input.txt and input2.txt.

C. *Input Files Information:*

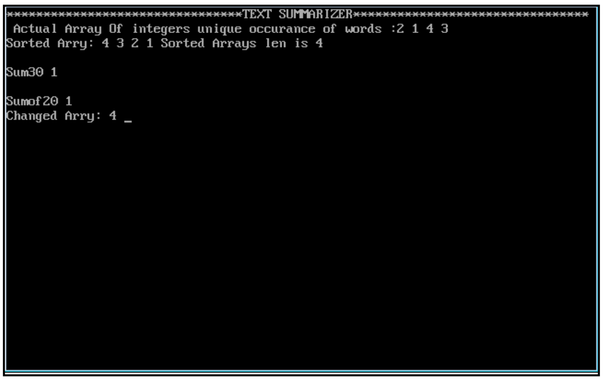


```
**********************************TEXT SUMMARIZER**********************************

Description of Actual Input File:
Total No of Paragraph Are= 6

 Length Of whole Document=2452 charachters
```

Fig. 6: Description of Input.Txt File

Above fig. 6, is snapshot after providing input.txt file name. This screen indicates description of input.txt file. It shows total number of paragraphs and total number of characters in input.txt file including spaces and special symbols.

D. Integer Array for Processing:

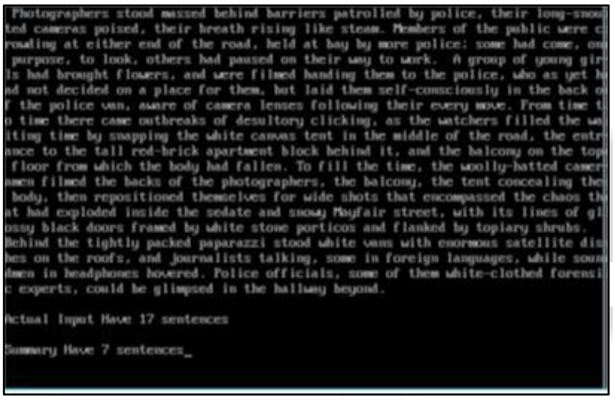


```
**********************************TEXT SUMMARIZER**********************************
 Actual Array Of integers unique occurance of words :2 1 4 3
Sorted Arry: 4 3 2 1 Sorted Arrays len is 4

Sum30 1

Sumof20 1
Changed Arry: 4 _
```

Fig. 7: Sorted Array of Integer Numbers

Above fig 7 is snapshot which is showing sorted array of integer numbers. This array includes integer occurrence of all words in input.txt file. Next line in snapshot shows 20% and 30% length of array. This length is used in text traversing i.e. out of all array 20% top occurrence words are taken in consideration from highest bushy node and 30% top occurrence words are taken in consideration from non-bushy node.

E. Generated Summary:



```
 Photographers stood massed behind barriers patrolled by police, their long-snou
ted cameras poised, their breath rising like steam. Members of the public were c
rowding at either end of the road, held at bay by more police: some had come, on
 purpose, to look, others had paused on their way to work.  A group of young gir
ls had brought flowers, and were filmed handing them to the police, who as yet h
ad not decided on a place for them, but laid them self-consciously in the back o
f the police van, aware of camera lenses following their every move. From time t
o time there came outbreaks of desultory clicking, as the watchers filled the wa
iting time by snapping the white canvas tent in the middle of the road, the entr
ance to the tall red-brick apartment block behind it, and the balcony on the top
 floor from which the body had fallen. To fill the time, the woolly-hatted camer
amen filmed the backs of the photographers, the balcony, the tent concealing the
 body, then repositioned themselves for wide shots that encompassed the chaos th
at had exploded inside the sedate and snowy Mayfair street, with its lines of gl
ossy black doors framed by white stone porticos and flanked by topiary shrubs.
Behind the tightly packed paparazzi stood white vans with enormous satellite dis
hes on the roofs, and journalists talking, some in foreign languages, while soun
dmen in headphones hovered. Police officials, some of them white-clothed forensi
c experts, could be glimpsed in the hallway beyond.

Actual Input Have 17 sentences

Summary Have 7 sentences_
```

Fig. 8: Resulted Summary

Above fig 8 is a snapshot of Resulted summary. After summary, it includes total number of sentences of actual input file and total number of sentences in summary.

IV. CONCLUSION AND FUTURE SCOPE

The proposed system focus on paragraph extraction mechanism. Early mentioned bushy method is with little bit complex but powerful algorithm. System requirement is economical for both development and deployment side. All the risks from different point of view are to taken in consideration. Hence, this system is feasible to use. Result of generated summary contains near about all important sentences from actual input file. Result is also short as compare with actual input file.

Hence, from above discussion we can say that this system is useful and reliable.

After implementation of this proposed system there are huge future scopes for summarizer.

Some of them listed below:

1) Text Summarizer can be developed for all kind of document such as- .docx, .doc, .PDF...etc.
2) In future Text Summarizer can be a very important module of Robotics field. like, A robot have to learn new culture of society by reading different books in less time period.
3) Text summarizer needs to develop with consideration of similarity of words. Like word "use" and "utilize" have same meaning hence, occurrence of word "use" need to increase with 2.

V. ACKNOWLEDGMENT

We would like to express our deep gratitude and sincere thanks to all who helped us to complete this work successfully. Our sincere thanks to the principal Dr. Prof. K.S.Wani, SSBT's COET for providing us facilities and resources to complete our work. Our deep gratitude goes to Dr. Prof. G.K.Patnaik, head of the department, for providing us opportunity to conduct this work. We are sincerely thankful to Miss. Prachi Chaudhari, our guide, for his valuable suggestions and guidance. Great thanks to our project associates and all those who helped directly or indirectly for completion of this project. Above all we would like to thank our parents without whose blessings, we unable to accomplish this work.

REFERENCES

[1] Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. Information Processing & Management, 33(2), 193207.

[2] Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text summarization by Paragraph Extraction

[3] Alok Ranjan Pal, Projjwal Kumar Maiti and Diganta Saha, An Approach To Automatic Text Summarization Using Simplified Lesk Algorithm And Wordnet, International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.3, No.4/5, September 2013.

[4] Young, S. R., & Hayes, P. J. (1985). Automatic classification and summarization of banking telexes. In proceedings of the 2nd Conference on Artificial Intelligence Applications (pp. 402408).

[5] Maheedhar Kolla(2002). Automatic Text Summarization Using Lexical Chains: Algorithms And Experiments, Department of Mathematics and Computer Science University of Lethbridge, Canada.

[6] Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, I-Heng Meng , Text summarization using a trainable summarizer and latent semantic analysis, Information Processing and management 41 (2005) 7595.

[7] Banko, M., Mittal, V., Kantrowitz, M., and Goldstein, J.1999. Generating extraction based summaries from handwritten summaries by aligning text spans. In Proceedings of PACLING99.