Document Summarization and Evaluation using Knowledge based Super Set Features

Sneh Garg
C-DAC Mohali
India

Sunil Chhillar
C-DAC Mohali
India

ABSTRACT
Document summarization is an important step while clustering the large no. of digital documents data base. Documents are clustered in accordance with their contents using the document text summary. The document summarization involves the knowledge corpus scheme comprising of corpus coverage, sentence coverage and term coverage weight. Further, three new weights are introduced as super sentence coverage weight, super corpus coverage weight and super term coverage weight. Super coverage weight is based on synonyms of the key words. The quality of document summary improves and diversified when synonyms of key words are also given due weightage in the process of text processing. The evaluation for the document summary quality is based on inner content metrics precision, recall, F-measure method.

Keywords
Super Sentence Coverage Weight, Super Corpus Coverage Weight, Super Term Coverage Weight, Document Summarization, Knowledge Corpus, Synonyms

1. INTRODUCTION
Document summarization is an important activity in digital documents processing. The application includes clustering, segmentation, document retrieval, quick reference or flash view of whole document etc. The document summary is primarily based on primary keywords. Normally, all sentences that contain the keywords are the part of document summary. Therefore, the quality of document summary primarily depends upon the accuracy or exactness of the key words vector set. Once the key words vector set is determined, the different documents may be clustered and segmented based on key words vector set comparison.

The keywords form the basis of the document summary. Further, A documents may contain a word that has potential to be keyword and another document contain the synonyms of that word and does not tend to be the keyword from that document properties point of view. In that case, this issue is resolved based on synonyms coverage weightage consideration.

A key word is attributed by its frequency in the document, its synonyms frequency and weightage in knowledge corpus. It is important to note here that the document summarization is independent of the grammar of the language. And is therefore content based keywords search and then summarization and not based on the context based.

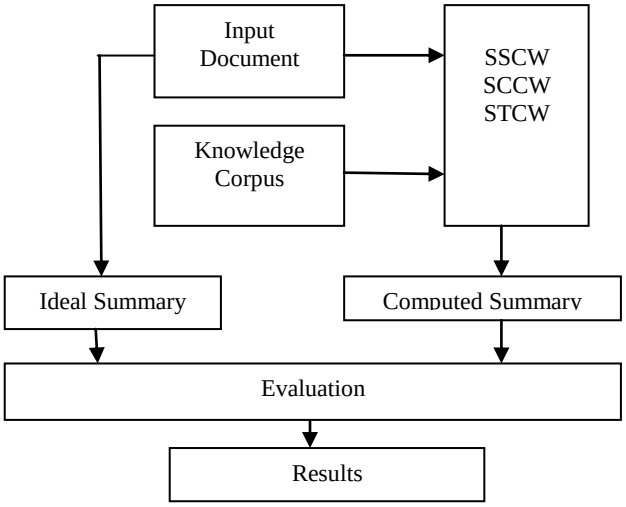

Fig. 1 System Block Diagram

SSCW-Super Sentence Coverage Weight

SCCW-Super Corpus Coverage Weight

STCW-Super Term Coverage Weight

2. RELATED WORK
A knowledge based feature set is extracted using the Sentence coverage weight (SCW), corpus coverage weight (CCW) and term coverage weight (TCW). The text summary is computed using the above features. The performance evaluation is analyzed using precision and recall parameters. [1]

Rhetorical Structure Theory (RST) is proposed for text summarization using an analytic frame work. This framework considers text structure at the clause level. Extraction of text's rhetorical structure and relations among sentences are calculated. Less important segments are removed. [2]

Sentences are prioritized using their connective strength (CS) values. K-mixture probabilistic model is used to establish term weights in a statistical sense, and further identifies the term relationships to derive the connective strength (CS) of nouns. [3]

The sentences are assigned some feature for the summary called ranking sentences and then select the best ones. The first step in summarization is by extraction is the identification of important features. To improve the quality and extract important feature for each sentence HMM tagger [4] is used. Ranked sentences are collected by identifying important features and summary is generated. [4]

The ATS is the identification of most important sentences from the given text by identifying the prime features of the sentences properly. A Conditional Random Field (CRF) based ATS can be used to identify and extract the correct features. [5]

Existing summarization approaches inherently assign more weights to the important sentences, whose extractive summary approach predicts the summary sentences that are important as well as readable to the target audience with good accuracy. Neural network technique is used for summary extraction of science and social subjects in the educational text. [6]

Synonyms based approach is used when the text summary is not target oriented or is very less i.e. less than 5% of the whole document. It is imperative to note that the length of text summary may be in between 30-40% of the whole document. [7].

3. ALGORITHM

Concept learning algorithm is applied on knowledge corpus of specific domain to extract the concept from documents. In the base paper, text summarization is done by considering three metrics: -Sentence Coverage Weight, Term Coverage Weight, and Corpus Coverage Weight.

However, based on above mentioned metrics, some important part of the document may be missed on account of synonyms of the terms used in the document. This synonyms based approach provides another three sets of the above mentioned metrics as given below:

- Super Sentence Coverage Weight (SSCW)
- Super Term Coverage Weight (STCW)
- Super Corpus Coverage Weight (SCCW)

The effects of synonyms terms as well as repetitions of synonyms in sentence and in corpus are taken into account while designing the algorithm for text summarization. This generates a super set of six number of features that may be statistically analyzed and evaluated for text summarization. Further, a histogram based approach may also provide information about a key word that may left even after applying the super set concept on the document under scanner. A brief view of the proposed system is depicted in fig. 1

The proposed work is divided into following stages:

3.1 Computation of Super Sentence Coverage Weight and Sentence Coverage Weight

- Extract SentenceSet from input document.
- Make TokenSet from SentenceSet.
- Eliminate stop words, stem TokenSet.
- Make FilterTokenSet by removing repetitive tokens.
- Synonyms Dictionary insertion of Key words and makes SynonymTokenSet.
- Synonyms words Array Generation
- Sentence Coverage Weight and Super Sentence Coverage Weight are

$$\text{SenCovWt(t)} = 1 - 1/(\text{occurance of term(t)in doc} * \text{TermFreq(t)/SenFreq(t)})$$

$$\text{SUPSenCovWt(t)} = 1 - 1/(\text{occurance of SYNterm(t)in doc} * \text{SYNTermFreq(t)/SYNSenFreq(t)}$$

where synonym term frequency, term frequency, sentence frequency and synonym sentence frequency are given below

$$\mathbf{TermFreq(t)} = \frac{occurance\ of\ term(t)\ in\ doc}{total\ no\ of\ elements\ in\ FilterTokenSest}$$

$$\boldsymbol{SYNTermFreq(t)} = \frac{occurance\ of\ Synterm(t)\ in\ doc}{total\ no\ of\ elements\ in\ SynoymTokenSest}$$

$$\boldsymbol{SenFreq(t)} = \frac{occurance\ of\ sentence\ having\ term(t)\ in\ SentenceSet}{total\ no\ of sentences\ in\ SentenceSet}$$

$$\boldsymbol{SYNSenFreq(t)} = \frac{occurance\ of\ sentences\ having\ SYnterm(t)\ in\ SentenceSet}{total\ no\ of sentences\ in\ SentenceSet}$$

3.2 Computation of Super Corpus Coverage Weight and Corpus Coverage Weight

- Make Knowledge Set from corpus and extract sentences from it and make token set from it.
- Calculate knowledge corpus term frequency and synonym knowledge corpus term frequency

$$\boldsymbol{KCTermFrq(t)} = \frac{occurance\ \ of\ \ term\ (t) in\ corpus}{no\ \ of\ \ elements\ \ in\ KCTokenSet}$$

$$\boldsymbol{SYNKCTermFrq(t)} = \frac{occurance\ \ of\ \ term\ (t) in\ corpus}{no\ \ of\ \ elements\ \ in\ SynKCTokenSet}$$

- Calculate Knowledge Document Frequency and synonym knowledge document frequency

$$\mathbf{KCDocFrq(t)} = \frac{No.of\ \ documents\ \ \ having\ \ term\ (t)\ from\ \ KC\ token\ \ set}{no\ \ of\ \ documets\ \ \ in\ KC}$$

$$\mathbf{SYNKCDocFrq(t)} = \frac{No.of\ \ documents\ \ \ having\ \ term\ (t)\ from\ \ synonym\ \ \ kc\ token\ \ set}{no\ \ of\ \ documets\ \ \ in\ KC}$$

- Calculate Corpus Coverage Weight and Super Corpus Coverage Weight

$$\mathbf{KCCovWt(t)} = 1 - 1/(occurance\ of\ \ term(t) in\ KC * KCTDocFrq(t)/KCTermFre(t))$$

$$\mathbf{SUPKCCovWt(t)} = 1 - 1/(occurance\ of\ term(t) in\ KC * SYNKCTDocFrq(t)/SYNKCTermFre(t))$$

3.3 Computation of Super Term Coverage Weight and Term Coverage Weight

$$\mathbf{TermCovWt(t)} = 1 - \frac{1}{\text{SenCovWt (t)} + \text{KCCovWt (t)}}$$

$$\boldsymbol{SUPTermCovWt(t)} = 1 - \frac{1}{\text{SUPSenCovWt (t)} + \text{SUPKCCovWt (t)}}$$

- Arrange TokenSet and Synonym TokenSet in descending order of their term coverage weight and super term coverage weight respectively.
- Make Sentence set SS and Super Sentence Set by moving sentences if it contains term (t) at least once.
- Computed summary is generated from super sentence set whereas ideal summary is generated by domain expert and are evaluated using precision, recall F-measure method.

4. FACTUAL SUMMARY

Due weightage is given to the factual information that is covered within double quoted comas in each document irrespective of keywords attribute and is made integral part of the text summary. This introduces the conservative approach flavor in preserving or retaining the factual information in the text summary. Similarly document title is also made part of the text summary as the document title is the prima-facie identifier of the document. The text summary length is controlled by providing the keywords extraction attributes control parameters like frequency of keyword occurrence, synonyms occurrence and term frequency etc. This way, the length of the text summary may be controlled according the application requirement.

5. PERFORMANCE EVALUATION

The performance of the algorithm is computed by computing the precision, recall and F-measure.

$$\text{Precision (P)} = \frac{IS \cap CS}{IS} \quad \dots \text{eq}^{n}.\ 1$$

$$\text{Recall (R)} = \frac{IS \cap CS}{CS} \quad \dots \text{eq}^{n}.\ 2$$

$$\text{F-Measure (F)} = \frac{2.P.R}{(P+R)} \quad \dots \text{eq}^{n}.\ 3$$

Where IS and CS are the length of ideal and computed summary respectively.

6. DATASET

For testing purposes, 100 documents from five different domains are used as follows:

Domain	Text Domain	No. of Documents
Domain-1	"Swatch Bharat Abhiyan" News Clips from Different News papers	20
Domain-2	"NAMO US Visit' News Clips from Different News papers	20
Domain-3	Nobel Prize - 2014 News Clips	20
Domain-4	Text Summarization Related IEEE papers	20
Domain-5	India LS Election 2014 News Clips	20

7. RESULT

The presented algorithm is implemented in matlab version 7.5.

Computed Text Summary Result for Domain-1

Domain	Computed Text Summary
Domain-1	"Modi also asked everyone to be a part of the 'Swatch Bharat' campaign and to make it a public movement rather than just a government mission" The Union Minister of Rural Development, Drinking water & Sanitation Nitin Gadkari said that scientifically proven Solid and Liquid Waste Management activities be launched in each Gram Panchayat of the countryClean India Mission - A public movement rather than a government mission The Prime Minister had also urged everyone to be a part of the 'Swatch Bharat' campaign and to make it a public movement rather than just a government mission Cleanliness of a country is directly linked to tourism and economy and thus Modi has asked everyone to make the 'Clean India Mission' a economic activity Government offices up to panchayat level will be involved in a cleanliness drive that shall be carried out from September 25 till Diwali which falls on October 23He said if the tourist destinations in the country are clean, it will bring more people and will also bring a change in India's global perception How is Swatch Bharat Mission and Indian economy connected

Computational Results for Different Domain for IS,CS

Domain	IS	CS	CS∩ *IS*	P	R	F
Domain-1	17	20	16	0.94	0.80	0.86
Domain-2	21	22	20	0.96	0.90	0.93
Domain-3	25	31	23	0.92	0.74	0.82
Domain-4	32	27	25	0.86	0.92	0.89
Domain-5	30	28	26	0.86	0.92	0.89

8. CONCLUSION

The documents summary from the test documents as discussed in result section is observed to very near to the summary when analyzed manually for testing and validation purposes. By virtue of inclusion of factual texts covered in text summary, no important information remains uncovered in the computed text summary. The presented algorithm is tested however on text documents of approximately containing 1000 words and that too in note pad file format. The main emphasis is given on accuracy of the keywords and in turn the text summary. However, the time constraint factor is not taken care off. If the documents size or no. of documents under scanner increases, there may be large processing time while computing the text summary.

9. REFERENCES

[1] Durga Bhavani Dasari, Dr. Venu gopala Rao. K., "Single Document Text Summarization by Knowledge-Corpus", 978-1-4799-1626-9/ 2013 IEEE.

[2] Li Chengcheng," Automatic Text Summarization Based On Rhetorical Structure Theory", 978-1-4244-7237-62010 IEEE.

[3] Te-Min Chang, Wen-Feng Hsiao," A hybrid approach to automatic text summarization", 978-1-4244-2358-3/2008 IEEE.

[4] Suneetha Manne, S. Sameen Fatima," A Feature Terms based Method for Improving Text Summarization with Supervised POS Tagging", International Journal of Computer Applications (0975 – 8887) Volume 47– No.23, June 2012.

[5] Nowshath K. Batcha, Normaziah A. Aziz," CRF Based Feature Extraction Applied for Supervised AutomaticText Summarization", Procedia Technology 11 (2013) 426 – 436.

[6] K. Nandhini, S.R. Balasundaram," Improving readability through extractive summarization for learners with reading difficulties", Egyptian Informatics Journal (2013) 14, 195–204.

[7] Alexander Yates, Oren Etzioni," Unsupervised Methods for Determining Object and Relation Synonyms on the Web", Journal of Artificial Intelligence Research 34 (2009) 255-296.

[8] Vipul Dalal, Dr.Latesh Malik,"A Survey of Extractive and Abstractive Automatic Text summarization Techniques", 978-1-4799-2560-5/2013 IEEE DOI 10.1109/ICETET.2013.31.

[9] Egitim Fakültesi, Mehmet Akif Ersoy," Quality of written summary texts: An analysis in the context of gender and school variables*"*, 1877-0428 © 2010 Published by Elsevier Ltd.

[10] Donia Scott, Catalina Hallett, Rachel Fettiplace," Data-to-text summarisation of patient records: Using computer-generated summaries to access patient histories", D. Scott 156 et al. / Patient Education and Counseling 92 (2013) 153–159

[11] Kushal Bafna, Durga Toshniwal," Feature Based Summarization of Customers' Reviews of Online Products", 2013 The Authors. Published by Elsevier B.V.

[12] Tiedan Zhu, Kan Li," The Similarity Measure Based on LDA for Automatic Summarization", 2011 Published by Elsevier Ltd

10. AUTHOR'S PROFILE

The author is pursuing her M.Tech. (IT) thesis work in Text Mining from CDAC, Mohali, India.