2011 2nd International Conference on Networking and Information Technology
IPCSIT vol.17 (2011) © (2011) IACSIT Press, Singapore

A Fast and Compact Method for Document Summarization on Mobile Devices using Non-Negative Matrix Factorization

Hiroya Kitagawa, El-Sayed Atlam, Takuki Ogawa, Masao Fuketa, Kazuhiro Morita and Jun-ichi Aoe

Department of Information Science and Intelligent Systems, Faculty of Engineering, University of Tokushima, 2-1Minami josanjima, Tokushima, 770-8506, Japan

{kitagawa, atlam, fuketa, kam, aoe}@is.tokushima-u.ac.jp

Abstract— With a fast paced economy, organization need to make a decision as fast as possible, access to large text documents or other information sources is important during decision making. Unfortunately, there are many shortcomings of the handheld devices, such as limited resolution and narrow bandwidth. In order to overcome the shortcomings, this paper proposes fast and compact approaches that can summarize documents on mobile devices efficiently. The presented method introduces unsupervised schemes using non-negative matrix factorization (NMF) that can determine the sentence precedence without morphological and syntax analysis.
From simulation results for DUC2006 test data, our approach could reduce the matrix size by about 74% and could speed up the precision of summarization by 8.5 times faster than the original method.

Keywords- non-negative matrix factorization; mobile devise; Summarization.

1. Introduction

With a fast paced economy, organization need to make a decision as fast as possible, access to large text documents or other information sources is important during decision making. Unfortunately, there are many shortcomings of the handheld devices, such as limited resolution and narrow bandwidth that lead to impossible loading and visualizing large documents. Therefore, document summarization on mobile phones is one of the most convenient applications. The primary goal of this paper is to create fast and compact approaches that can summarize documents on mobile devices efficiently.

Automatic document summarization based on unsupervised schemes is very useful approaches because it does not require training data. Reference [1] and [2] have proposed summarization methods using Latent Semantic Analysis (LSA), however the LSA methods could not extract meaningful paragraphs because the meaning of the semantic features cannot be captured intuitively. Reference [3] has proposed an unsupervised generic document summarization method using a non-negative matrix factorization (NMF) method to solve this problem. However, this method has no discussions for mobile phones with low computing ability.

Many approaches for automatic text summarization system have been applied on mobile devices [4-7]. However, all these approaches based on irrelevant content (i.e., HTML tags, multimedia) for the web page or syntax analysis summarization which needs big dictionaries.

Therefore, the presented method proposes an efficient preprocessing and matrix reduction for the NMF method on mobile phones that have low computing ability and slow wireless connectivity using the NMF method. The new approach can determine the paragraph precedence without morphological and syntax analysis requiring dictionaries.

From simulation results for DUC2006 test data, our approach could reduce the matrix size by about 74% and could speed up the precision of summarization by 8.5 times faster than the original method without degrading the precision of extracted paragraphs.

2. System OutlineS

2.1 Presented Method Conditions

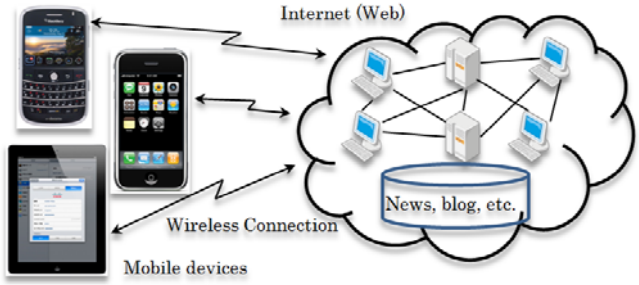

Figure 1 Relationships between mobile devices and Web documents.

Figure 1 shows relationships between mobile devices and Web documents. Generally, it is difficult to look the whole document because the display is very small in mobile devices, therefore text summarization for reducing the text size is very important. Two general schemes can be considered to reduce the text size. The first one is to obtain summarized texts by transferring generated texts on server systems into the mobile devices with no additional computation on those devices. The second one is to generate summary for original documents of Web on the devices directly without high cost servers. This paper focuses on the second one that can perform a fast summarization by means of speed up and compact storages on mobile devices as follows:

[Condition 1] Speed

Simple and fast summarization algorithms are required because the mobile device is lower computing ability than general personal computers. Therefore, the presented method introduces the NMF method requiring no heavy text analysis and computation.

[Condition 2] Storage

In the summarization algorithms, computation by less memory must be expected because some of mobile devices do not have enough storage to analyze whole documents. Therefore, the presented method proposes the compact approach based on the non-negative matrix factorization (NMF).

2.2 Summarization Process Outlines

Figure 2 represents the overall flow of the summarization system to be presented in this paper.

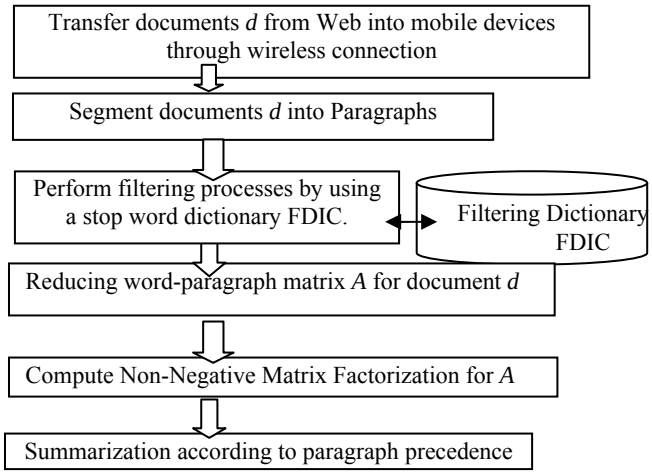

Figure 2. Overall flow of the summarization system

In Figure 2, documents are transferred from Web into mobile devices through wireless connection. In segmentation, documents are separated into collection of paragraphs including a few paragraphs according to [8] approach. The next step performs filtering processes by using stop-word dictionary FDIC[1] and word stemming by using Porter's algorithm [9]. A word-paragraph matrix *A* is constructed for paragraphs, and the matrix *A* is decomposed by two matrices. Finally, summarization can be performed by extracting precedence paragraphs from document *d*.

An efficient preprocessing and matrix reduction for the NMF method on mobile phones have low computing ability and slow wireless connectivity will be discussed in Section 3.

3. SUMMARIZATION ALGORITHMS

3.1 Stemming and Filtering Approaches

There are some preprocesses before building word-paragraph matrix *A* from document *d* in Fig. 2. Segmentation according to [8] approach is very fast and it does not affect the summarization results.

In order to reduce the total number of words for the matrix *A*, word stemming is performed by Porter's algorithm [9] and stop words is performed by using word list. Moreover, the complexity of NMF computation is related to the size of matrix A, so the reduction is useful approaches if the relevancy of summarization doesn't degrade. Therefore, this paper performs matrix reduction together with the above stop words and confirms the boundary of possible reduction.

3.2 NMF Document Summarization

NMF is introduced by basic notations [10]. NMF decomposes word-paragraph matrix *A* ($n \times m$) size into two matrices *W* ($n \times k$) size and *H* ($k \times m$) size for *k* such that $k < n$ and $k < m$ as follows:

$$A \approx WH \tag{1}$$

Where *W* is called a non-negative semantic feature matrix (NSFM) and *H* is called a non-negative semantic variable matrix (NSVM). Let *A*[*i,j*], W[i,j] and H[i,j] be elements of *A, W* and *H* for the *i-th* row and for the *j-th* column, respectively. *W* [*,*j*] and *H[*,j]* represents a semantic feature and semantic variable vectors for the *j*-the column, respectively.

By the recomposing process, the paragraph vector *A[*,j]* can be represented by a linear combination of the *h-th* semantic feature vectors *W[*,h]* and the semantic variable *H[h,j]* as follows:

$$A[*, j] = \sum_{h=1}^{k} H[h, j] W[*, h] \tag{2}$$

3.3 Presented Method Algorithm

For document *d*, *W (d)* represents a set of words in document *d* and *P(d)* represents a set of paragraphs in document *d*. Suppose that the total elements of *W(d)* and *P(d)* are *n* and *m*, respectively. The original matrix A becomes $n \times m$. Let *F(x)* be a frequency for word *x* and let α be the threshold of the low word frequency to be removed.

Step1. Compute *F(x) for all x in W(d).*

Step2. For all *x*, remove *x* from *W(d)* if $F(x) \leq \alpha$

Step3. For all paragraphs *p*, remove *p* from *P(d)* if *p* is the empty such that *p* has no *x*

Step4. Construct reduced matrix *B* of the size $n' \times m'$ such that *n'* and *m'* are the total elements of *W(d)* and *P(d)*, respectively.

Table 1 shows a paragraph matrix A for words and paragraph obtained by the preprocessing applied on a set of paragraphs by the original method. In Table 1, for matrix A, the rows representing 561 of terms and the columns representing 57 segmented paragraphs, while the *i-th* rows and the *j-th* columns representing term frequency (i.e. A[6,5] = 3).

[1] *http://www.webconfs.com/stop-words.php*

Table 2 shows reduced matrix B by using the presented algorithm for matrix A. Table 3 explains the semantic feature of paragraphs applying NMF to matrix B, where W_i *means W[*,i].*

Table 1. Paragraphs of Matrix A for Documents in Table 1 and their Frequencies

N_0	Stemming words	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	…	P_{57}
1	attent	1	0	1	0	0	0	0	0	0	0	…	0
2	countri	0	0	0	0	0	0	0	1	0	0	…	0
3	major	0	0	0	0	0	1	0	0	0	0	…	0
4	problem	0	0	0	0	0	0	1	0	0	1	…	0
5	percent	0	0	0	0	3	5	2	0	0	0	…	0
6	leader	0	1	0	0	0	0	0	0	0	0	…	0
7	poverti	1	0	1	0	0	0	0	1	1	0	…	0
8	poorest	1	0	0	0	0	0	0	1	0	0	…	0
9	job	0	0	0	0	0	0	1	0	0	0	…	0
10	suggest	0	0	1	0	0	0	0	0	0	0	…	0
…	……	…	…	…	…	…	…	…	…	…	…	…	…
561	help	0	0	0	0	0	0	0	0	0	2	…	0

Table 2. Reduced Paragraphs of Matrix B for Document using Presented Algorithm

N_0	Stemming words	P_1	P_2	P_3	P_5	P_6	P_7	P_8	P_9	…	P_{10}
1	attent	1	0	1	0	0	0	0	0	…	0
2	countri	0	0	0	0	0	0	1	0	…	0
3	major	0	0	0	0	1	0	0	0	…	0
4	problem	0	0	0	0	0	1	0	0	…	1
5	percent	0	0	0	3	5	2	0	0	…	0
6	leader	0	1	0	0	0	0	0	0	…	0
8	poorest	1	0	0	0	0	0	1	0	…	0
…	……	…	…	…	…	…	…	…	…	…	…
196	help	0	0	0	0	0	0	0	0	…	2

Table 3: Semantic Feature of Paragraph using NMF for Matrix B

Semantic variable	Paragraph											
	P_1	P_2	P_3	…	P_5	P_6	P_7	P_8	P_9	…	…	P_{10}
H_1	0	0	0	…	0	0	0	0	0	…	…	0.230
H_2	0	0	0.203	…	0	0	0	0.013	0	…	…	0
H_3	0	0	0	…	0	0	0	0.001	0	…	…	0
H_4	0	0	0	…	0	0	0	0	0	…	…	0
H_5	0	0	0	…	0.212	0.014	0	0.003	0	…	…	0
H_6	0.176	0	0.003	…	0	0	0	0.045	0	…	…	0
H_7	0	0.002	0	…	0	0	0	0.009	0.219	…	…	0
H_8	0	0	0	…	0	0	0	0	0	…	…	0.002
H_9	0	0.204	0	…	0	0.001	0	0.028	0.002	…	…	0
H_{10}	0	0	0	…	0	0.001	0	0.006	0	…	…	0

Table 4: Semantic Variable Vectors of Paragraphs using NMF

N_0	Stemming words	Semantic feature											
		W_1	W_2	W_3	…	W_5	W_6	W_7	W_8	W_9	…	…	W_{10}
1	attent	0	0	0	…	13.32	0	0	0	0	…	…	0.01
2	countri	0	0	8.12	…	0	0	0	0	0	…	…	0
3	major	3.86	0	0	…	0	0	0	0	0	…	…	0
4	problem	0	4.95	0	…	0.04	6.35	4.64	0	0.29	…	…	0.02
5	percent	0	4.89	0	…	0	0	0	0	0	…	…	0
7	poverti	0	0	0	…	0	5.32	0	0	0	…	…	0
8	poorest	0	0	4.11	…	0	0	0	0	0	…	…	0

…	……	……..	……..	……..	…	……..	……..	……	……	……	…	…	……
196	help	0	0	0	…	0	0	9.07	0	4.66	…	…	0

From Table 2, it is clear that the matrix A is reduced after using Step 2 of the presented algorithm and delete low frequency terms such as "leader", "job", "suggest", etc with F(x) = 1, and delete the empty paragraphs as P_4, P_{14}. The original matrix A size is corresponding to 561 terms of 57 paragraphs while the presented reduction matrix B is corresponding to 196 terms of 10 paragraphs.

From Table 3, the semantic feature vectors, $W_1, W_2,...,W_{14}$, obtained from NMF decomposition of matrix A. From Table 3, it is clear that the highest semantic feature values are $W[3,3]$, $W[3,4]$, $W[5,6]$, and $W[8,6]$ with the terms "country", "problem", and "poverty" respectively, which will be affected in extracting important paragraphs as the summary.

From Table 4, the semantic variable vectors $H_1, H_2,\ldots,H_{10}$ obtained from NMF decomposition of matrix B. From Table 4, it is clear that the semantic variable vector $H[10,1]$ is representing the highest value. This value will use in Generic Relevance of a Paragraph (*GRP)* for document summarization in the following section.

3.4 Generic Document Summarization

Reference [3] proposed a novel method to select paragraphs based on NMF and defined *GRP* for the *j-th* paragraphs as follows:

$$GRP = \sum_{i=1}^{k} \left(H\left[i, j\right] \times weight\left(H\left[i,*\right]\right)\right) \tag{5}$$

$$weight\left(H\left[i,*\right]\right) = \frac{\sum_{q-1}^{n} H\left[i,q\right]}{\sum_{q-1}^{k}\sum_{q-1}^{n} H\left[p,q\right]} \tag{6}$$

Where the weight *(H[i,*])* is the relative relevance among the *i-th* semantic feature. It is clear that the generic relevance could reflect the major topics of paragraphs using the representation of their semantic features.

Table 5 shows the paragraph extraction process from the semantic variable vector *H[i,*]* for paragraphs obtained by NMF in Table 4. By using Eq.(5), paragraph P_6 in Table 5 is extracted, which is corresponding to the highest *GRP* (0.194).

Table 5 Paragraph Extraction Processing using *GRP*

Paragraph	P_1	P_2	P_4	P_5	P_6	…	P_{10}
GRP	0.025	0.053	0.008	0.129	0.194	…	0.168

4. Experimental Results

4.1 Experimental Data

The evaluation system has the following conditions:

- Mobile computer

The proposed system has been developed on OS: Android2.2, Maker: Camangi, Model: Camangi-FM600, CPU: 0.6GHz, Main Memory: 500MB

- Test data

The proposed approach using test data documents from DUC2006 data set [11] that can compare manual summaries by experts with the automatic summaries. 50 documents of test data are selected randomly which includes 50 topics with 25 documents related to each topic [8]. Each document is segmented into paragraphs and manual summaries up to 250 words are used. ROUGE evaluation systems [12] can compare generated summaries by the presented method with manual summaries for DUC2006 data set.

4.2 Reduction and Speed Observations

Table 6 shows reduction average of matrix size using proposal method for 50 documents, where α is the threshold of the low word frequency, average word or average paragraph is representing the average of words or paragraphs in documents. While matrix represents the total elements from multiplying total words and total paragraphs.

Table 6 Reduction Average of Matrix Size using Presented Method

	Average word	Average paragraph	matrix
Original	172.2	12.8	2663.4
α=1	42.5	12.7	677.4(74%)
α=2	18.7	12.5	294.5(88%)
α=3	10.5	12.0	165.0(93%)
α=4	6.8	11.7	109.4(95%)

From Table 6, it turns out that the matrix size for the original method was 2663.4 while it becomes 677.42 by the presented method which means that the presented method can achieve maximum improvement with about 74% reduction of matrix A.

Table 7 shows the time speeding of the computation of the NMF method for the original and the presented methods in mobile phone for DUC2006.

Table 7 Speed Average for Each Stage of Summarization in Mobile Phone

	ConstructMatrix[ms]	NMF[ms]	Total[ms]
Original	185.7	2334.7	2520.4
α=1	145.5	632.9	778.5(3.2times)
α=2	135.8	317.8	453.7(5.5times)
α=3	152.8	198.8	351.7(7.1times)
α=4	145.4	150.4	295.9(8.5times)

From Table 7 results, it turns out that the maximum time speeding of the presented method using the construct matrix is about 8.5 times faster than the original method.

4.3 Evaluation System

In the previous section the improvement of the speeding time is verified. However, it is important to confirm that the relevancy of the system is not degrading. The presented method is using the traditional evaluation ROUGE system as follows:

1) ROUGE Evaluation System

ROUGE scores were computed by running ROUGE-1.5.5 [12] with stemming but no removal of stop words. The input file implemented so that scores of systems and humans could be compared.

ROUGE evaluation systems is used to compute recall, precision and *f*-measure by using ROUGE_N representing recall between generated summary of the proposed system and manual summary. Let *n* be the length of the *n-gram*, $gram_n$ is the maximum number of n-gram in the generated summary and $Count_n(gram_n)$ is a set of manual summary.

$$ROUGE\_N = \frac{\sum_{S \in \{manual\ summary\}} \sum_{gram_n \in S} Count_n(gram_n)}{\sum_{S \in \{manual\ summary\}} \sum_{gram_n \in S} Count\ (gram_n)}$$

In the system, five automatic evaluation methods are prepared in the ROUGE evaluation system ROUGE_N, ROUGE_L, ROUGE_W, ROUGE_S, and ROUGE_SU.

2) Relevancy of Generated Summary

Relevancy of generated summaries is computed by the number of extracting paragraphs. In the presented method, filtering dictionaries are used to reduce the total number of words representing the word-paragraph matrix. Therefore, the evaluation of relevancy is to compare the presented method with the original NMF approaches.

Table 8 Average Precision for Presented and Original Methods**.**

	ROUGE-N	ROUGE-L	ROUGE-W	ROUGE-SU
Original	0.40836	0.362191	0.225636	0.178623

α=1	0.411709	0.365855	0.22773	0.179655
α=2	0.412725	0.368108	0.228997	0.179897
α=3	0.412451	0.368029	0.229342	0.180873
α=4	0.419124	0.375551	0.235797	0.185717

From the results of Table 8, it is clear that the precision of the presented method higher than the original method among all ROUGE measures.

In order to confirm the precision rate for the threshold α of the low word frequency, the average of the best case and worst case are estimated in Table 9 and Table 10.

Table 9. Average of Precision for the best Case Results.

	ROUGE-N	ROUGE-L	ROUGE-W	ROUGE-SU
Original	0.477235	0.427044	0.266373	0.222752
α=1	0.475591	0.42503	0.264674	0.218573
α=2	0.473058	0.427048	0.265833	0.220121
α=3	0.475832	0.428203	0.266827	0.220045
α=4	0.470657	0.425698	0.266716	0.215884

Table 10 Average of the Precision for Worst Case Results

	ROUGE-N	ROUGE-L	ROUGE-W	ROUGE-SU
Original	0.336629	0.298122	0.189653	0.134487
α=1	0.33852	0.303671	0.192672	0.137196
α=2	0.350343	0.314221	0.200625	0.143667
α=3	0.355757	0.318796	0.203555	0.148952
α=4	0.361594	0.320937	0.202907	0.151131

From the simulation results of Tables 9 and 10, it is clear that the average of the best case is decreased in descending order among all ROUGE measures, while the average of the worst case is increasing. The reason for increasing the worst case is that the low frequency words have no information meaning. Therefore, removing these low frequency words is useful for the presented method.

5. Conclusion

This paper has proposed a fast and compact approach that can summarize documents on mobile devices efficiently. The presented method has introduced unsupervised schemes using non-negative matrix factorization (NMF) that can determine the paragraph precedence without morphological and syntax analysis. From simulation results for DUC2006 test data, our approach could reduce the matrix size by about 74% and could speed up the precision of summarization by 8.5 times faster than the original method without degrading the precision of extracted paragraphs

Future works could extend this work using large corpus to get more speeding up of the computation and increasing of the precision of extracted paragraphs.

6. References

[1] H. Zha, Generic summarization and key phrase extraction using mutual reinforcement principle and sentence clustering. In Proceedings of the 25th annual international ACM SIGIR conference, 2002, pp. 113–120). Tampere, Finland.

[2] J. -Y. Yeh and H. -R. Ke, Text summarization using a trainable summarizer and latent semantic analysis, Information Processes and Management, Vol. 41, 2005, pp. 75–95.

[3] J-H. Lee, S., Park, C-M Ahn and D., Kim, Automatic generic document summarization based on non-negative matrix factorization, Information Processing and Management, Vol. 45, 2009, pp. 20–34.

[4] H. Alam, R. Hartono, A. Kumar, F. Rahman, Y. Tarnikova and C. Wilcox, Web Page Summarization for Handheld Devices: A Natural Language Approach, in proceedings of the Seventh International Conference on

Document Analysis and Recognition, 2003.

[5] O. Buyukkokten, H. Garcia-Molina, A. Paepcke, Text summarization of web pages on handheld devices, Proceedings of Workshop on Automatic Summarization 2001, in conjunction with The Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001), Association for Computational Linguistics, Pittsburgh, PA, USA, 2001 (Jun.)

[6] C.C. Yang, F.L. Wang, Automatic summarization for financial news delivery on mobile devices, Proceedings of the Twelfth International Conference on World Wide Web (WWW 2003), ACM Press, Budapest, Hungary, 2003 (May), pp. 391–392.

[7] C. C. Yang, F. L. Wang (2007), an information delivery system with automatic summarization for mobile commerce, Decision Support Systems Journal, Vol. 43, 2007, pp. 46– 61.

[8] T. D. Hao, Overview of DUC 2005, In proceeding of the document understanding conference (DUC'05), 2005.

[9] W. Frankes and R. Baeza-Yates, Information Retrieval: Data Structure & Algorithms, Prentice Hall, 1992.

[10] S. Liu, Enhancing e-business-intelligence-service: atopic-guided text summarization framework, Seventh IEEE International Conference on E-Commerce Technology (CEC), July 19–22, pp.493–496.

[11] Duc06: http://duc.nist.gov/

[12] C. Y. Lin, ROUGE: A package for automatic evaluation of summaries. In Proceedings of workshop on text summarization branches out, post-conference workshop of ACL, 2004.