Computing and Informatics, Vol. 35, 2016, 744–760

DOCUMENT SUMMARIZATION USING NMF AND PSEUDO RELEVANCE FEEDBACK BASED ON K-MEANS CLUSTERING

Sun Park, ByungRae Cha, JongWon Kim*

Gwangju Institute of Science and Technology
123 Cheomdangwagi-ro, Buk-gu
Gwangju 61005, Korea
e-mail: {sunpark, brcha, jongwon}@gist.ac.kr

Abstract. According to the increment of accessible text data source on the internet, it has increased the necessity of the automatic text document summarization. However, the performance of the automatic methods might be poor because the semantic gap between high level user's summary requirement and low level vector representation of machine exists. In this paper, to overcome that problem, we propose a new document summarization method using a pseudo relevance feedback based on clustering method and NMF (non-negative matrix factorization). Relevance feedback is effective technique to minimize the semantic gap of information processing, but the general relevance feedback needs an intervention of a user. Additionally, the refined query without user interference by pseudo relevance feedback may be biased. The proposed method provides an automatic relevance judgment to reformulate query using the clustering method for minimizing a bias of query expansion. The method also can improve the quality of document summarization since the summarized documents are influenced by the semantic features of documents and the expanded query. The experimental results demonstrate that the proposed method achieves better performance than the other document summarization methods.

Keywords: Document summarization, NMF, PRF, clustering, query expansion, semantic feature

* Corresponding author

1 INTRODUCTION

With the fast growth of the internet services access by users and device things, the amount of information with respect to accessible text data source has been explosively increasing and the vast information of various data source has been accumulating. It is difficult to read and understand the individual information from an enormous amount of information source on the internet for internet users and device things. Most of the internet users and device things need the method of distilling the core of information from the sea of information to produce an abridged version of the source data. Automatic document summarization method is the essential technology for information seeking and condensing goals on numerous sources of information. Besides, it becomes more and more important in many text based applications and electronic device things [1, 2, 3, 4, 5].

Traditional document summarization is the process of reducing the size of documents while maintaining their basic outlines. That is, the process should distill the most important information from the text document sources. The summary type can be either generic summary or query-based summary. A generic summary presents an overall sense of the documents' contents whereas a query based summary presents the contents of documents that are related to user's query. Document summarization method is divided into single-document summarization or multi-document summarization according to the scope of the summary target. The purpose of multi-document summarization is to produce a single summary from a set of related documents, whereas single-document summarization is intended to summarize only one document [2].

A manual document summarization shows better performance than the document summarization by means of a machine. A machine produces the summary by using statistical features of document whereas a human can generate more meaningful summary by grasping the concept of document. In other words, the qualities of document summarization methods may be poor since the semantic gap between the high level user's summary requirement and the low level vector representation of machine exists. Therefore, we need to reduce the semantic gap between the features captured by the machine and human's concepts to enhance the performance of the document summarization [1, 2, 3]. Recently, document summarization methods based on the user feedback (UF) [4, 5, 6, 7, 8, 9, 10] and NMF (non-negative matrix factorization) [11, 12, 13] try to narrow down the semantic gap between low level features and high level concepts. The UF is a query reformulation technique of an information retrieval field, which can be either a relevance feedback (RF) or a pseudo relevance feedback (PRF). The RF refines the current query using the documents that have been identified as the relevant ones by the user, however, it needs an intervention of a user for a relevance judgment on documents. The PRF can provide an automatic relevance judgment to expand query without user's intervention but it might get biased query during a query expansion process [5]. The NMF represents individual object as a non-negative linear combination of partial information extracted from a large volume of objects [7, 8, 9, 11, 12, 13, 14, 15, 16]. Advantages

of NMF are that the NMF has a great power to easily extract semantic features representing the inherent structure of data objects [14, 15]. However, there is still some disadvantage of NMF, it might limit successful decomposing semantic features from any data set as data objects viewed from extremely different viewpoints, or highly articulated objects [14, 16].

In order to resolve the above limitations of the document summarization methods based on NMF and PRF we propose a new document summarization method using NMF and pseudo relevance feedback based on K-means. The proposed method has the following advantages: First, it provides an automatic relevance judgment using query expansion method based on K-means clustering without the intervention of a user. Second, it can successfully decompose semantic feature from extremely different mixing topics of document since the expanded query by PRF helps us to minimize the biased semantic features by means of the clustered topics. Finally, the method can improve the quality of document summarization since it can minimize a semantic gap between the user's concept of summarization and the vector representation of machine, which uses the expanded query of the PRF and the selected sentences of NMF.

The rest of the paper is organized as follows: Section 2 mentions related work regarding the document summarization; Section 3 describes NMF; Section 4 describes the proposed summarization method; Section 5 describes the performance evaluation. Finally, in Section 6 we conclude this paper.

2 RELATED WORK

There are two approaches for document summarization method: supervised or unsupervised methods. The supervised methods [2, 4] typically make use of human-made summaries or extracts to find features or parameters of summarization algorithms, while unsupervised approaches [5, 6, 7, 8, 9, 11, 12, 13, 16, 17, 18, 19, 20, 21, 22] determine relevant parameters with no regard to human-made summaries. Recent studies for document summarization methods based on unsupervised techniques use graph based methods [22], NLP (natural language processing) [20, 21], MMR (maximal marginal relevance) [18], LSA (latent semantic analysis) [18], NMF (non-negative matrix factorization) [7, 8, 9, 11, 12, 13, 16], and UF (user feedback) [4, 5, 6, 7, 8, 9]. In this related work, the focus is placed on document summarization based on UF and NMF because our proposed method uses the feedback and semantic feature techniques based on NMF.

The recent studies for document summarization methods using UF and semantic feature are as follows: Han et al. [5] proposed a text summarization using relevance feedback with query splitting (QS). Their method can alleviate the problem that feedback gets biased query during a query expansion process by splitting the initial query into several pieces, while it might produce poor summaries of documents in the case that it has insufficient information for QS. Hu et al. [6] proposed the comments-oriented summarization method using a feature-biased and uniform

document approach, which generates comments-oriented summary in the form of extracted sentences from a given web document. Berger and Mittal [4] proposed a document summarization method using frequently-asked question (FAQ). Their method uses FAQ document, which comprises questions and answers for specific topic, as the training data. Their method needs to construct FAQ before summarizing documents and their result depends largely on the training data. Park et al. [7] proposed a query-based document summarization method using NMF and relevance feedback (RFNMF), which expands the query through relevance feedback to reflect user's requirement and extract meaningful sentences using the semantic features. However, this method needs to get feedback from user as to what sentences are relevant. Park et al. [8] proposed a user-focused automatic document summarization method using NMF and pseudo relevance feedback (PRFNMF), which can provide an automatic relevance judgment on sentences. However, this method may get the biased inherent semantics of the document to be reflected in summaries. Park et al. [9] proposed an automatic query-based personalized document summarization method using PRF with NMF, which can minimize reflecting biased inherent semantic features for document summarization. In the present study, our previous works (i.e. the conference papers) [7, 8, 9] were enhanced because they have advantages in clear extraction of the meaningful information when compared with the UF methods [4, 5, 6]. However, they are restricted within the structure of the original document set. Thus, the proposed method refines the structure of document set with reference to topics by means of clustering method.

3 NON-NEGATIVE MATRIX FACTORIZATION

In this paper, we define the matrix notation as follows: Let X_{*j} is the j^{th} column vector of matrix X, X_{i*} be the i^{th} row vector, and X_{ij} be the element of the i^{th} row and the j^{th} column.

NMF is to decompose a given $m \times n$ non-negative matrix A into multiplication of an $m \times r$ non-negative semantic feature matrix (NSFM), W, and an $r \times n$ non-negative semantic variable matrix (NSVM), H, as shown in Equation (1) [14, 15]:

$$A \simeq WH, \tag{1}$$

where r (i.e. the number of semantic features) is usually chosen to be smaller than m (i.e. the number of rows) or n (i.e. the number of columns) so that the total sizes of W and H are smaller than that of the matrix A. We use the Frobenius norm as Equation (2) to satisfy the approximation condition $\widetilde{A} = WH$ [14, 15].

$$\Theta_E(W,H) \equiv \|A - WH\|_F^2 \equiv \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - \sum_{l=1}^{r} W_{il}H_{lj}\right)^2 \tag{2}$$

This is lower bounded by zero, and clearly vanished if and only if $A = WH$. W and H are continuously updated until $\Theta_E(W,H)$ converges under the predefined

threshold or exceeds the number of repetitions. The update rules are as follows [14, 15]:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu}\frac{(AH^T)_{\alpha\mu}}{(W^THH)_{\alpha\mu}}, \quad W_{i\alpha} \leftarrow W_{i\alpha}\frac{(AH^T)_{i\alpha}}{(WHH^T)_{i\alpha}}. \tag{3}$$

Figure 1 shows the example using NMF algorithm: Let r be 3, the number of repetitions be 50, and the tolerance be 0.001. When the initial elements of W and H matrices are 0.5, it decomposes the matrix A into the W and H matrices, as shown in Figure 1.

$$\underset{A}{\begin{pmatrix} 0 & 3 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 4 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 \end{pmatrix}} \approx \underset{W}{\begin{pmatrix} 0 & 1.752 & 0.089 \\ 0 & 0.726 & 1.428 \\ 1.730 & 0 & 0.002 \\ 0.865 & 0 & 0.001 \end{pmatrix}} \times \underset{H}{\begin{pmatrix} 0 & 0 & 2.312 & 0 \\ 1.618 & 0.004 & 0 & 0.004 \\ 0.549 & 0.698 & 0.003 & 0.698 \end{pmatrix}}$$

Figure 1. Results of NMF algorithm from matrix A

4 DOCUMENT SUMMARIZATION USING PRF AND NMF

In this section, we propose a new document summarization method using NMF and pseudo relevance feedback based on K-means clustering. The proposed method consists of the preprocessing phase, the pseudo relevance feedback phase, and the document summarization phase.

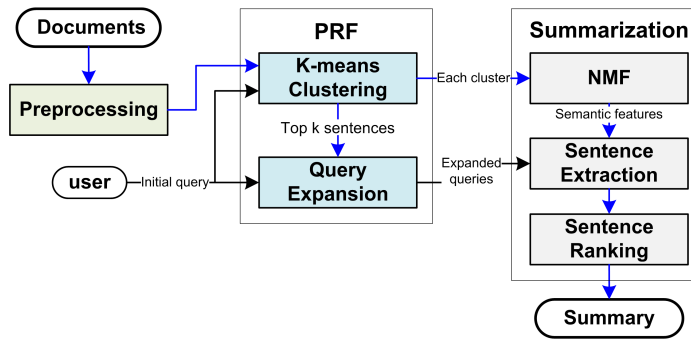

Figure 2. The proposed method using PRF based on clustering and NMF

4.1 Preprocessing Phase

In the preprocessing phase of generating document summaries, after a given English document is decomposed into individual sentences, we remove all stopwords by using Rijsbergen's stopwords list and perform word stemming by Porter's stemming algorithm [1, 3]. Then term sentence frequency matrix (i.e. term-frequency vector) is constructed for each sentence in the document. Let A be $m \times n$ terms sentences frequency matrix, where m is the number of terms and n is the number of sentences in the document set [3].

4.2 Pseudo Relevance Feedback Phase

The proposed pseudo relevance feedback phase consists of the clustering step and the query expansion step. The clustering step uses the K-means to cluster sentences. K-means clustering is a partition algorithm that splits given set of n object into K clusters. Given set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, K-means clustering aims to partition the n observations into k ($\leq n$) sets $S = (S_1, S_2, \ldots, S_k)$ so as to minimize the inter-cluster sum of squares (ICSS, sum of distance functions of each point in the cluster to the K center). In other words, its objective is to find [1, 2, 3]:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2. \tag{4}$$

K-means clustering is performed by using the cosine similarity measure [2, 3] with respect to the matrix A as shown in Equation (4). In this paper, the number of K is set by the number of extracted sentences for summarizing document.

$$d(A_{*a}, A_{*b}) = 1 - \mathrm{sim}(A_{*a}, A_{*b}), \tag{5}$$

$$\mathrm{sim}(A_{*a}, Q) = \frac{A_{*a} \cdot Q}{|A_{*a}| \times |Q|} = \frac{\sum_{i=1}^{m} A_{ij} q_i}{\sqrt{\sum_{i=1}^{m} A_{ij}^2} \times \sqrt{\sum_{i=1}^{m} q_i^2}}, \tag{6}$$

where a query vector $Q = (q_1, q_2, \ldots, q_m)$, q_i denotes the i^{th} term frequency of the query, m denotes the number of terms. We assume that K clusters are constructed and then the matrix A can be represented as Equation (6). The matrix of the p^{th} cluster of sentences, C_p, is a subset of the column vectors of matrix A. The C_p and C_q are disjointed and satisfy the following property:

$$\{A_{*j} | j = 1, \ldots, n\} = \bigcup_{p=1}^{K} \{C_{*l}^{p} | l = 1, \ldots, s_p\}, C^p \cap C^q \neq \phi, p \neq q, \tag{7}$$

where s_p is the number of column of C_p, n is the number of sentences, and K is the number of clusters.

In the query expansion step, the cosine similarity for the query expansion of PRF between the initial query and a sentence vector in clusters is calculated by using equation (5) , and then the top k ranked sentences having the high similarity values is selected. The query expansion is performed by using the extracted top k ranked sentences. The query expansion method using query point movement (QPM) [10] is computed as follows:

$$Q^{new} = Q + \frac{\sum_{t=1}^{k} w_t \times A_{*t}}{\sum_{t=1}^{k} w_t}, \quad w_t = \mathrm{sim}(Q, A_{*t}), \tag{8}$$

where Q^{new} is a new expanded query vector of current query Q, A_{*t} is a t^{th} sentence in the relevant sentences, w_t is a weight and it is a cosine similarity between current query Q and A_{*t} by using Equations (5) and (6).

4.3 Document Summarization Phase

The document summarization phases consist of sentence extracting step and sentence ranking step. The sentence extracting step is described as follows: Matrices W^p and H^p are constructed by applying the NMF algorithm to A^p as shown in Equation (9). The semantic feature W^p_{*l} having the largest similarity value is selected by using Equations (7)–(9) for the expanded query, and then the sentence having the largest weight with respect to this semantic feature is extracted. The extracted sentence is added to the candidate sentence set. Please refer to the NMF method for document summarization [12, 13].

$$A^p = W^i H^i, \quad i = 1, 2, \ldots, K, \tag{9}$$

where K is the number of clusters.

A column vector A^p_{*j} for j^{th} sentence of matrix A of p^{th} cluster is represented as a linear combination of semantic feature vectors (W^p_{*l}) and semantic variable (H^p_{lj}). That is, the weight of l^{th} semantic feature vector W^p_{*l} in sentence A^p_{*j} is H^p_{lj}.

$$A^p_{*j} = \sum_{l=1}^{r} H^p_{lj} W^p_{*l} \tag{10}$$

The powers of the two non-negative matrices W^p and H^p are described as follows: all semantic variables H^p_{lj} are used to describe how the j^{th} sentence of p^{th} cluster is structured using semantic features. The W^p and H^p are represented sparsely. Intuitively, it makes more sense for each sentence to be associated with some small subset of a large array of topics W^p_{*l}, rather than with just one topic or with all the topics. In each semantic feature W^p_{*l}, semantically related terms are grouped together by NMF. In addition to grouping semantically related terms together into semantic features, NMF uses context to differentiate between multiple meanings of the same term [15].

The sentence ranking step is described as follows: The ranking score of the candidate sentences is calculated by using Equation (12). As the number of duplicate sentences in the result set increases, the score of the sentence becomes higher.

$$Score_j = dn_j \times w_j, w_j = \mathrm{sim}(Q, A_{*j}) \tag{11}$$

where $Score_j$ is the ranking score of the j^{th} candidate sentence A_{*j} in the set of candidate sentences C_a, dn_j is the number of duplicate sentences among the set of candidate sentences C_a, and w_j is cosine similarity value between initial query Q and the candidate sentence A_{*j}.

Figure 3 shows the example of the sentence representation using Equation (10) and Figure 1. The column vector A_{*3} corresponding to third sentence is represented as a linear combination of semantic feature vectors W_{*l} and semantic variable column vector H_{*3}.

$$\begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix} \approx 2.312 \times \begin{pmatrix} 0 \\ 0 \\ 1.730 \\ 0.865 \end{pmatrix} + 0 \times \begin{pmatrix} 1.752 \\ 0.726 \\ 0 \\ 0 \end{pmatrix} + 0.003 \times \begin{pmatrix} 0.089 \\ 1.423 \\ 0.002 \\ 0.001 \end{pmatrix}$$

$$A_{*3} \qquad H_{13} \qquad W_{*1} \qquad H_{23} \qquad W_{*2} \qquad H_{33} \qquad W_{*3}$$

Figure 3. Example of sentence representation using semantic features and semantic variables from Figure 1

4.4 Document Summarization Algorithms

The proposed document summarization algorithm using PRF based on clustering and NMF is as follows:

Algorithm. *KPRFNMF*(D, Q)
Input: the document D, the query Q, the number of semantic feature vectors r,
the number of clusters and the number of extracted sentences K
Output: the term-frequency matrix A, the set of cluster C,
the expanded query Q^{new}, the set of sentences s,
the set of candidate sentences C_a, the set of ranking score RS,
the set of summarized sentences S
Method:
01: *Preprocessing*(D) {
02: $s \leftarrow$ *DecomposeDocument*(D);
03: $s \leftarrow$ *Stopword*(s);
04: $s \leftarrow$ *Stemming*(s);
05: $A \leftarrow$ *ContructMatrix*(s);

06: }
07: $KPRF(C, Q)$\{
08: $C \leftarrow kmeans(A, K)$;
09: $Q^{new} \leftarrow QueryExpansion(C, Q)$;
10: }
11: $Summarization(A, C, Q^{new})$\{
12: $[W^p, H^p] \leftarrow NMF(C)$;
13: for $j \leftarrow 1$ to n do {
14: select $e = \arg\max_{1 \le l \le r}[sim(Q^{new}, W^p_{*l})]$;
15: select $f = \arg\max_{1 \le j \le n}[H^p_{ej}]$;
16: $C_a \leftarrow C_{*f}$;
17: }
18: $RS \leftarrow Score(C_A, Q)$;
19: $S \leftarrow Extract(RS, C_a, K)$;
20: }

In lines 1 to 6, the preprocessing phase removes stop-words and performs word stemming, and then constructs the term frequency matrix. In lines 7 to 10, pseudo relevance phase uses K-means clustering and query expansion methods. In line 9, query expansion uses Equation (8). In lines 11 to 20, document summarization phase uses NMF and sentences ranking scores. In line 14, the semantic feature vector W^p_{*l} in p^{th} cluster most similar to query Q^{new} is selected. In line 15, the e^{th} column having the largest value H^p_{ej} in p^{th} cluster among the e^{th} row of H is selected in order to choose the sentence that has the largest weight with respect to the most relevant semantic feature vector (W^p_{*l}). In line 18, ranking scores of the set of candidate sentences are calculated. In line 19, the set of summarized sentences is extracted.

The number of sentences	Sentences
S1	A course on integral equations.
S2	Attractions for computer and evolution equations.
S3	Algorithms and computer implementations.
S4	Automatic differentiation of algorithms.
S5	Theory of delay differential equations.
Query	Computer algorithm

Table 1. The five sentences and the query

Example 1. We illustrate the example of sentence extraction with respect to the document summarization algorithm from line 12 to line 16. Table 1 shows five sentences and a query. A matrix A is generated by preprocessing a set of sentences in Table 1, and a matrix A is decomposed into a semantic feature matrix and a semantic variable matrix by using NMF. Figure 4 illustrates the sentence extraction process with respect to the five sentences and query in Table 1. In step 1 of Figure 4, we

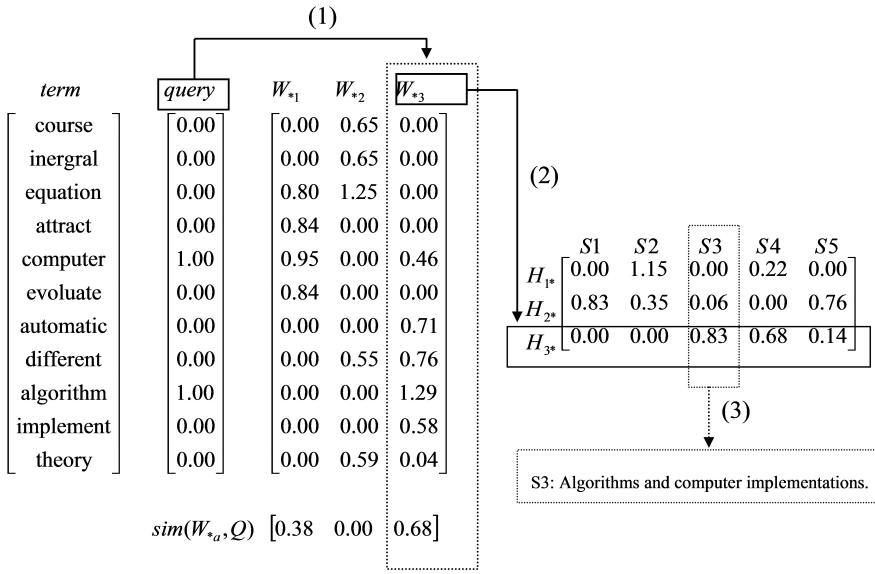

Figure 4. Example of the extracted sentence using similarity between query and semantic feature vectors

calculate the similarity between query and semantic feature vectors and select the semantic feature vector W_{*3} having the largest similarity value (0.68). In step 2 of Figure 4, the semantic variable vector H_{*3} that corresponds to the semantic feature vector W_{*3} is selected. In step 3 of Figure 4, the sentence S3 that corresponds to the semantic variable having the largest value (0.83) in semantic variable vector H_{*3} is extracted.

5 EXPERIMENTS

5.1 Data Set and Performance Evaluation Measure

The data set of Document Understanding Conference (DUC) 2007 is used for the performance evaluation of the proposed method. The DUC is the international conference for performance evaluation of the document summarization methods to compare manual summaries by experts with summaries of the proposed summarization system [23].

ROUGE (recall-oriented understudy for gisting evaluation) evaluation software package is used to evaluate the proposed method, where ROUGE has been applied by the DUC for the performance evaluation in the competition of document summarization methods. ROUGE includes five automatic evaluation methods, ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU. Each method estimates Re-

call, Precision, and f-measure between experts' reference summaries and candidate summaries of the proposed system. ROUGE-N uses n-gram recall between a candidate summary and a set of reference summaries [23, 24].

5.2 Evaluation Method

We implemented 7 kinds of summarization methods (i.e. KPRFNMF, PRFNMF, KWNMF, NMF, KMEAN, MMRLSA, QPRF). KPRFNMF denotes the proposed summarization method using NMF and PRF based on K-means. PRFNMF denotes the multi-document summarization method using non-negative matrix factorization and PRF [8, 9]. KWNMF denotes the multi-document summarization method using the weighted non-negative matrix factorization and K-means clustering [12, 13]. NMF denotes the document summarization method using non-negative matrix factorization [11]. KMEAN denotes the document summarization method using K-means clustering [2, 3], which clusters the sentences and from each cluster extracts the sentence where similarity value with respect to a given query is the largest. MMRLSA denotes Hachey's summarization method [18] using MMR and LSA. The QPRF denotes Han's summarization method that uses PRF based query splitting (QS) [5].

To compare the performances, we used the ROUGE evaluation software package which compares the various summary results from several summarization methods with the generated summaries of human beings. As a test data, we randomly selected 50 documents from DUC2007 data set. Each document has a summary done by human beings. Our methods (KPRFNMF, PRFNMF, KWNMF, NMF) and other 3 methods (KMEAN, MMRLSA, QPRF) produce summaries using test documents. Those summaries are input to ROUGE software to yield the ROUGE evaluation values.

5.3 Results and Discussions

We conducted the performance evaluation using ROUGE measure with respect to seven document summarization methods, which it compared the ROUGE results of the seven different summarization methods: KPRFNMF, PRFNMF, KWNMF, NMF, KMEAN, MMRLSA, and QPRF. Figure 5 shows the ROUGE results for the Average Recall. The ROUGE results for the average precision and the f-measure are shown in Figure 6 and Figure 7. In Figure 5, the average recall of KPRNMF is approximately 5.333 % higher than that of KMEAN, 4.603 % higher than that of MMRLSA, 3.528 % higher than that of QPRF, 3.013 % higher than that of NMF, 1.203 % higher than that of KWNMF, 0.778 % higher than that of PRFNMF. In Figure 6, the average precision of KPRFNMF is approximately 9.053 % higher than that of KMEAN, 3.548 % higher than that of MMRLSA, 3.298 % higher than that of QPRF, 2.025 % higher than that of NMF, 0.938 % higher than that of KWNMF, 0.523 % higher than that of PRFNMF. In Figure 7, the average f-measure of KPRFNMF is approximately 6.190 % higher than that of KMEAN, 4.315 % higher than that of

MMRLSA, 4.340 % higher than that of QPRF, 2.453 % higher than that of NMF, 0.975 % higher than that of KWNMF, 0.490 % higher than that of PRFNMF.

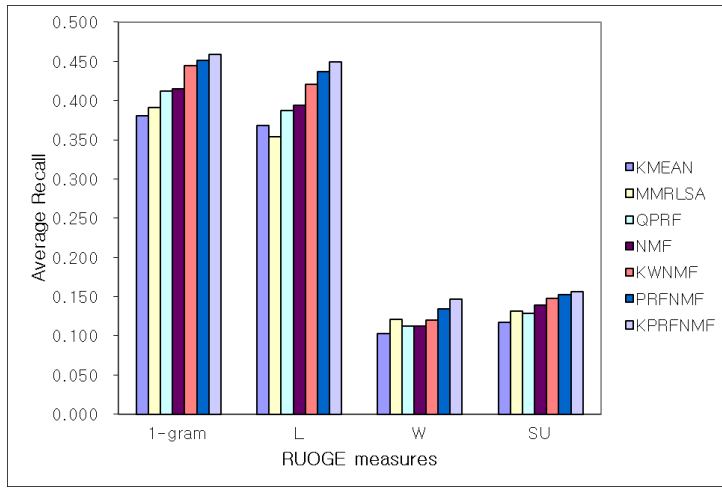

Figure 5. Performance comparison using average recall with respect to ROUGE measures

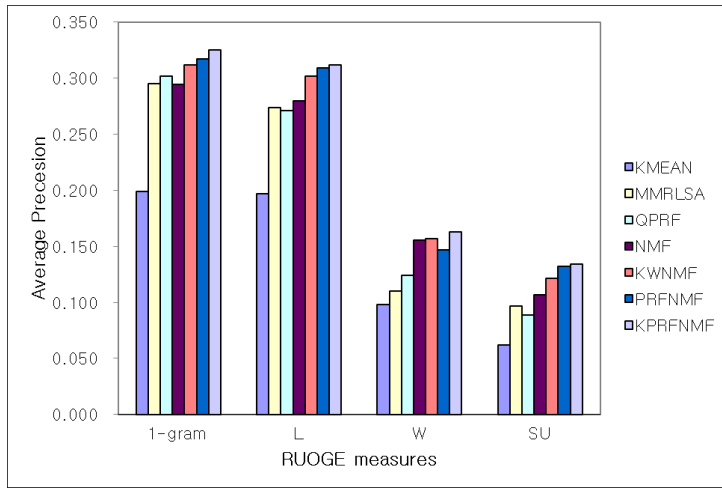

Figure 6. Performance comparison using average precision with respect to ROUGE measures

The result shows that ROUGE measures of MMRLSA has better performance than the KMEAN since the MMRLSA generates more meaningful summary by reflecting the latent semantics of documents. The QPRF shows the better performance than the MMRLSA since the QPRF extracts more important sentences using query

expansion. The NMF shows the better performance than the QPRF. The QPRF minimizes biased query expansion by splitting the initial query into several pieces whereas it may produce meaningless summary in the case that it has insufficient information for relevant sentences. The NMF uses the similarities between the initial query and the semantic features in documents. It cannot reflect the user's requirement properly to summarize the summarization, in the case that the initial user's query is biased. The KWNMF shows better performance than the NMF since it does not select less meaningful sentences by using the weighted similarity measure. The PRFNMF shows the better performance than the KWNMF. The PRFNMF uses the query expansion and the semantic features representing the inherent structure of a document so that it can improve the quality of document summaries. The KPRFNMF shows the best performance since it can minimize a semantic gap between manual summary and automatic summary by means of the semantic features by NMF and the PRF based on K-means clustering method.

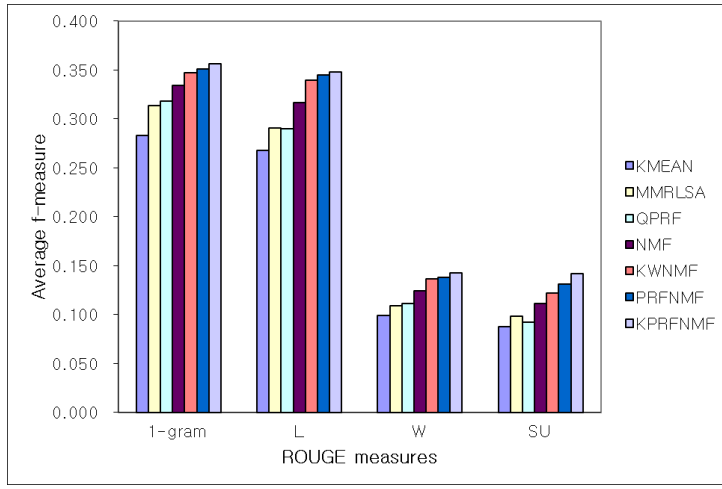

Figure 7. Performance comparison using average f-measure with respect to ROUGE measures

6 CONCLUSIONS AND FUTURE RESEARCH

In this paper, we propose a new document summarization method using pseudo relevance feedback based on document clustering and semantic features. The proposed method can reduce the semantic gap between high level user's requirement and low level vector representation of machine. Besides, it can minimize the bias of query expansion since it can well reflect the inherent structure of a document by using NMF, and it provides an automatic relevance judgment using clustering method on sentences without intervention of a user. Our experiment demonstrates that

the proposed technique (i.e. KPRFNMF) provides a significant improvement over KMEAN, MMRLSA, QPRF, NMF, KWNMF and PRFNMF in terms of ROUGE measures of recall, precision, and f-measure. As a future work, we plan to apply our method in the application of device things for an information summarization of internet users or device things. We anticipate that our application would help internet users or device things to improve more useful services.

Acknowledgement

The reserach was supported by the Software Convergence Technology Development Program through the Ministry of Science, ICT and Future Planning (S1070-15-1071), and the GRI(GIST Research Institute) Project through a grant provided by GIST in 2016.

REFERENCES

[1] Frakes, W. B.—Baeza-Yates, R.: Information Retrieval: Data Structures and Algorithms. Prentice-Hall, 1992.

[2] Mani, I.: Automatic Summarization. John Benjamins Publishing Company, 2011.

[3] Baeza-Yates, R.—Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, 1999.

[4] Berger, A.—Mittal, V. O.: Query-Relevant Summarization Using FAQs. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL '00), Hong Kong, October 2000, pp. 204–301.

[5] Han, K. S.—Bea, D. H.—Rim, H. C.: Automatic Text Summarization Based on Relevance Feedback with Query Splitting. Proceedings of the 5th International Workshop on Information Retrieval with Asian Language, Hong Kong, September 2000, pp. 201–202.

[6] Hu, M.—Sun, A.—Lim, E. P.: Comments-Oriented Document Summarization: Understanding Document with Readers' Feedback. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '08), Singapore, July 2008, pp. 291–298.

[7] Park, S.: Document Summarization Using Non-Negative Matrix Factorization and Relevance Feedback. Proceedings of International Conference on Hybrid Information Technology (ICHIT '08), Daejeon Korea, August 2008, pp. 301–306.

[8] Park, S.: User-Focused Automatic Document Summarization Using Non-Negative Matrix Factorization and Pseudo Relevance Feedback. Proceedings of International Conference on Computer Engineering and Applications (ICCEA '09), Manila, Philippines, June 2009, pp. 101–105.

[9] Park, S.—An, D. U.: Automatic Query-Based Personalized Summarization That Uses Pseudo Relevance Feedback with NMF. Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication (ACM ICUIMC '10), Suwon Korea, January 2010, Art. No. 61.

[10] Porkaev, K.—Chakrabarti, K.—Mehtotra, S.: Query Refinement for Multimedia Similarity Retrieval in MARS. Proceedings of 7th Annual ACM International Conference on Multimedia (ACM Multimedia '99), Los Angeles, October 1999, pp. 235–238.

[11] Park, S.—Lee, J. H.—Ahn, C. M.—Hong, J. S.—Chun, S. J.: Query Based Summarization Using Non-Negative Matrix Factorization. Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES '06), Bournemouth, UK, October 2006, pp. 84–87.

[12] Park, S.—Lee, J. H.—Kim, D. H.—Ahn, C. M.: Multi-Document Summarization Based on Cluster Using Non-Negative Matrix Factorization. Proceedings of the 33th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM '07), Harrachov, Czech Republic, January 2007. Lecture Notes in Computer Science, Vol. 4362, 2007, pp. 761–770.

[13] Park, S.—Lee, J. H.—Kim, D. H.—Ahn, C. M.: Multi-Document Summarization Using Weighted Similarity Between Topic and Clustering-Based Non-Negative Matrix Factorization. Proceedings of the 8th Annual International Conference on Asia Pacific Web (APWEB '07), Huang Shan, China, June 2007. Lecture Notes in Computer Science, Vol. 4505, 2007, pp. 108–115.

[14] Lee, D. D.—Seung, H. S.: Learning the Parts of Objects by Non-Negative Matrix Factorization. Nature, Vol. 401, 1999, pp. 788–791.

[15] Lee, D. D.—Seung, H. S.: Algorithms for Non-Negative Matrix Factorization. Advances in Neural Information Processing Systems, Vol. 13, 2001, pp. 556–562.

[16] Xu, W.—Liu X.—Gong, Y.: Document Clustering Based on Non-Negative Matrix Factorization. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '03), Toronto, Canada, July 2003, pp. 267–273.

[17] Goldstein, J.—Kantrowitz, M.—Mittal, V.—Carbonell, J.: Summarizing Text Documents: Sentence Selection and Evaluation Metrics. Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '99), CA USA, August 1999, pp. 121–128.

[18] Hachey, B.—Murray, G.—Reitter, D.: The Embra System at DUC 2005: Query-Oriented Multi-Document Summarization with a Very Large Latent Semantic Space. Proceedings of the 4th Conference of the Document Understanding Conferences (DUC '05), BC Canada, October 2005.

[19] Sanderson, M.: Accurate User Directed Summarization from Existing Tools. Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM '98), Bethesda, Maryland, November 1998, pp. 45–51.

[20] Sakurai, T.—Utsumi, A.: Query-Based Multidocument Summarization for Information Retrieval. Proceedings of NII Test Collection for IR Systems (NTCIR '04), Tokyo, Japan, April 2003.

[21] Sassion, H.: Topic-Based Summarization at DUC 2005. Proceedings of the 5th Conference of the Document Understanding Conferences (DUC '05), BC Canada, October 2005.

[22] Varadarajan, R.—Hristidis, V.: Structure-Based Query-Specific Document Summarization. Proceedings of the 14th International Conference on Information and Knowledge Management (CIKM '05), Bremen, Germany, November 2005, pp. 231–232.

[23] Hoa, H. D.: Overview of DUC 2006. Proceedings of the 5th Conference of the Document Understanding Conferences (DUC '05), BC Canada, October 2005.

[24] Lin, C. Y.: ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL (ACL '04), Barcelona Spain, July, 2004.

Sun Park received his B.Sc. degree in computer engineering from Jeonju University in 1996. He received his M.Sc. degree in information and communication engineering from Hannam University in 2001, and his Ph.D. degree in computer and information engineering from Inha University in 2007. From November 2013, he has been working at the School of Information and Communications, Gwangju Institute of Science and Technology (Gwangju, Korea), where he is Research Professor. Prior to becoming a researcher at GIST, he worked as Research Professor at Mokpo National University, as a postdoctoral scholar at Chonbuk National University, and as Professor in the Department of Computer Engineering, Honam University, Korea. His research interests include data mining, information retrieval, information summarization, convergent marine ICT, and IoT-cloud computing.

ByungRae Cha is Research Professor at the School of Information and Communications, and Super Computing and Collaboration Environment Technology (SCENT) Center, GIST, Korea. He received his Ph.D. degree in computer engineering from National Mokpo University in 2004 and his M.Sc. degree in computer engineering from Honam University in 1997. Prior to becoming Research Professor at GIST, he has worked as Research Professor in the Department of Information and Communication Engineering, Chosun University, and as Professor in the Department of Computer Engineering, Honam University, Korea. His research interests include computer security of IDS and P2P, neural networks learning, cloud computing, and Future Internet.

JongWon Kim received his B.Sc., M.Sc. and Ph.D. degrees from Seoul National University (Seoul, Korea), in 1987, 1989 and 1994, respectively, all in control and instrumentation engineering. In 1994–2001, he was a faculty member of KongJu National University (KongJu, Korea) and University of Southern California (Los Angeles, USA). From September 2001, he has joined Gwangju Institute of Science and Technology (Gwangju, Korea), where he is now Full Professor. From April 2008, he is serving as the Director of GIST SCENT (Super Computing CENTer). Also, he is leading Networked Computing Systems Lab. (renamed from Networked Media Lab.) that focuses on "Dynamic and resource-aware composition of media-centric services employing programmable/virtualized resources". His recent research interests cover topics such as software defined networking (SDN)/cloud computing (CC) for Future Internet testbed and smart media-centric services employing heterogeneous SmartX nodes.