*2009 International Conference on Computer Engineering and Applications*
*IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore*

# User-focused Automatic Document Summarization using Non-negative Matrix Factorization and Pseudo Relevance Feedback

Sun Park[+]

Department of Computer Engineering, Honam University, Gwangju, Korea

**Abstract.** This paper proposes an automatic document summarization method using the pseudo relevance feedback (PRF) and the non-negative matrix factorization (NMF) to extract sentences relevant to a user's interesting for user-focused summary. The proposed method can improve the quality of document summaries because the inherent structure of the documents are well reflected by using the semantic features and the semantic variables calculated by NMF. Also it can provide an automatic relevance judgment on query expansion without the intervention of user. The experimental results demonstrate that the proposed method achieves better performance the other methods.

**Keywords:** document summarization, NMF, pseudo relevance feedback

## 1. Introduction

With the fast growth of the Internet access by personal users, it has increased the necessity of the personalized information retrieval and information summarization. Most of the information summarization is based on the user-oriented document summarization techniques. The document summarization is the process of reducing the size of documents while maintaining their basic outlines. The document summarization can be either generic summaries or user-focused summaries. A generic summary presents an overall sense of the documents' contents whereas a user-focused summary presents the contents of the document that are related to the user's query [9]. The relevance feedback is a query reformulation method which refines the current query using the documents that have been identified as relevant by the user. It constructs the new query that will be moved towards the relevant documents and away from the non-relevant ones [16]. However, it needs an intervention of user for a relevance judgment on documents. The *PRF* can provide an automatic relevance judgment on sentences without the user's intervention [4, 5]. The *NMF* represents individual object as the non-negative linear combination of part information extracted from a large volume of objects. This method can deal with a large volume of information efficiently since an original non-negative matrix is decomposed into sparsely distributed representation of two non-negative matrices [7, 8].

In this paper, we propose a new user-focused document summarization method using *pseudo relevance feedback* (*PRF*) and *non-negative matrix factorization* (*NMF*). The proposed method has the following advantages. First, it can provide an automatic relevance judgment on sentences without the intervention of user and it can minimize a semantic gap between user's concept for summarization and the vector representation for the query by using the *PRF*. Second, it can easily grasp the inherent structure of a document by using the semantic feature and the semantic variable and it can improve the quality of document summarization.

The rest of the paper is organized as follows: Section 2 describes the related works regarding document summarization, In Section 3, the proposed summarization method is introduced. Section 4 shows the evaluation and experimental results. Finally, we conclude in Section 5.

[+] Corresponding author. Tel.: + 82-62-940-5422; fax: +82-62-940-5422.
*E-mail address*: sunpark@honam.ac.kr.

## 2. Related Works

The previous studies for user-focused document summarization are as follows: Han et al. proposed a text summarization using relevance feedback with query splitting [5]. 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 may produce poor summaries of documents in the case that it has insufficient information for query splitting. Berger and Mittal proposed a document summarization method using frequently-asked question (FAQ) [1]. FAQ document is comprised of 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. Sakurai and Utsumi proposed a query based multi-document summarization method using a thesaurus. Their method generates the core part of the summary from the most relevant document to the query using a thesaurus, and then gets the additional part of the summary, which elaborates upon the query, from the other documents. Their method has beneficial effect on long summaries whereas its performance is not satisfactory for short summaries [17].

Park et al. proposed a query based summarization method using NMF [10]. This method extracts sentences using the cosine similarity between a query and semantic features. However, this method might produce poor document summarization in the case that initial user query does not reflect the user's requirement. Park et al. also proposed a multi-document summarization method based on clustering using NMF [11, 12, 13]. This method clusters the sentences and extracts sentences using the cosine similarity measure between a topic and semantic features. This method improves the quality of summaries and avoids the topic to be deflected in the sentence structure by clustering sentences and removing noise.

Park et al. proposed a query based document summarization method using NMF and relevance feedback [14]. This method 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 proposed a query based document summarization using NMF based pseudo relevance feedback [15]. This methods use pseudo relevance feedback to extract sentence without the intervention of user. However, this method might extract meaningless sentence by using simple cosine similarity function for query expansion.

## 3. Document Summarization by Using PRF and NMF

In this paper, we propose a user-focused automatic document summarization method using NMF and PRF. The proposed method consists of the preprocessing phase, pseudo relevance feedback phase, and sentence extraction phase. We will give a full explanation of three phases as Figure 1.

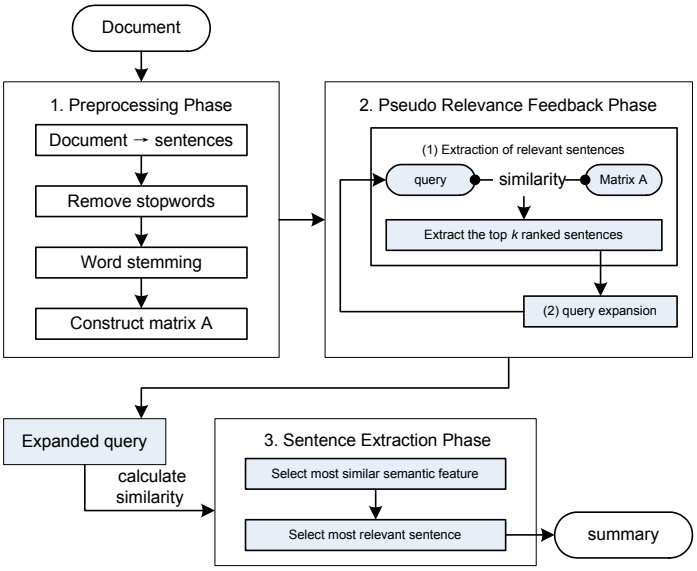


**Fig. 1:** User-focused document summarization method

## 3.1. Preprocessing phase

In this paper, we define the matrix notation as follows: Let $X_{*j}$ be $j'$th column vector of matrix $X$, $X_{i*}$ be $i'$th row vector, and $X_{ij}$ be the element of $i'$th row and $j'$th column. In the preprocessing phase, documents are decomposed into individual sentences, stop-words are removed, and word stemming is performed. Then the term-frequency vectors for all sentences in documents are constructed [3, 16]. Let $m \times n$ matrix $A$ be [ $A_{*1}$, $A_{*2}$, …, $A_{*n}$ ], where $m$ is the number of terms and $n$ is the number of sentences in documents. The column vector $A_{*i}$ is the term-frequency vector of $i$'th sentence.

## 3.2. Pseudo relevance feedback phase

The *pseudo relevance feedback* phase is described as follows. We calculate the cosine similarity between the initial query and a sentence vector by using equation (1) and then we select the top $k$ ranked sentences having the high similarity values. The cosine similarity function between the sentence vector $A_{*i}$ and query $Q$ is computed as follows [3, 16].

$$sim(A_{*i}, Q) = \frac{A_{*i} \cdot Q}{|A_{*i}| \times |Q|} = \frac{\sum_{j=1}^{m} A_{ji} \times q_j}{\sqrt{\sum_{j=1}^{m} A_{ji}^2} \times \sqrt{\sum_{j=1}^{m} q_j^2}} \tag{1}$$

Where query vector $Q = (q_1, q_2, \ldots, q_m)$, $q_j$ denotes the $j'$th term frequency of query, $m$ denotes the number of terms. We perform the query expansion by using the extracted top $k$ ranked sentences. The query expansion method is computed as follow:

$$Q^{new} = Q^{old} + \frac{\sum_{t=1}^{k} w_t \times A_{*i_t}}{\sum_{t=1}^{k} w_t}, \qquad w_t = sim\left(Q^{old}, A_{*i_t}\right) \tag{2}$$

Where $Q^{new}$ is a new expanded query vector of current query $Q^{old}$, $A_{*i_t}$ is a $t'$th sentence in the set of top $k$ ranked sentences, $w_t$ is the weight which is the cosine similarity value between current query $Q^{old}$ and $A_{*i_t}$.

## 3.3. Sentence extraction phase

The sentence extraction phase uses the *NMF* to extract the sentences for document summary. Please refer to the *NMF* method for document summarization [10]. *NMF* is to decompose a given matrix $A$ into a non-negative semantic feature matrix $W$ and a non-negative semantic variable matrix $H$ as shown in Equation (3) [7, 8].

$$A \approx WH \tag{3}$$

*NMF* algorithm keeps updating $W$ and $H$ until the object function $J$ converges under the predefined threshold or exceeds the number of repetition by using the Equation (4).

$$J = \left\| A - WH \right\|^2 \tag{4}$$

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T V)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}}, \qquad W_{i\alpha} \leftarrow W_{i\alpha} \frac{(VH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \tag{5}$$

Where $A$ is $m \times n$ non-negative matrix, $W$ is $m \times r$ non-negative matrix, and, $H$ is $r \times n$ non-negative matrix, $r$ is the number of semantic feature vectors. Usually $r$ is chosen to be smaller than $m$ or $n$, so that the total sizes of $W$ and $H$ are smaller than that of the original matrix $A$. A column vector corresponding to $j'$th sentence, $A_{*j}$, can be represented as a linear combination of semantic feature vectors $W_{*l}$ and semantic variable $H_{lj}$ as follows:

$$A_{*j} = \sum_{l=1}^{r} H_{kj} W_{*l} \tag{6}$$

The sentence extraction phase is described as follows. We decompose the matrix $A$ into semantic feature matrix and semantic variable matrix by using *NMF* as shown Equation (3). We calculate the similarity between query and semantic feature vectors and select the semantic feature vector having the highest

similarity value by using Equation (1). We select the semantic variable vector corresponding to the selected semantic feature vector. We extract the sentence corresponding to the largest value of semantic variable. We repeat these steps until the predefined number of sentences to be summarized is reached.

## 4. Experiment and Results

We use the *ROUGE-N* (recall-oriented understudy for gisting evaluation) to evaluate the proposed method. The ROUGE[1] has been applied by Document Understanding Conference (DUC[2]) for performance evaluation [2]. The DUC is the international conference for performance evaluation of the proposed system by comparing manual summaries with summaries of the proposed system [6]. As experimental data, we use the testing data of DUC 2005. We conducted the performance evaluation of the user-focused document summarization using 8 topics of DUC 2005. *ROUGE-N* estimates recall, precision, and *f*-measure between experts' reference summaries and candidate summaries of the proposed system using *N*-gram [2].

In this paper, we conducted two kinds of experiments to evaluate the performance of the proposed method. In the first experiment, we identify the optimal number of the relevant sentences for query expansion. In the second experiment, we compared the proposed method with other methods.

**(Experiment 1)** We evaluated the performance of the query expansion with respect to the number of relevant sentences *k*. We calculate the similarity values between initial query and sentences vectors to select top *k* ranked relevant sentences. We conducted the performance evaluation using *f*-measure with respect to the number of relevant sentences. Figure 2 shows the number of relevant sentences affecting the performance of document summarization. We changed the number of relevant sentences from 1 to 10 to verify the effectiveness of the query expansion. The evaluation results are shown in Figure 2. It shows the best performance when the number of relevant sentences is 3.

**(Experiment 2)** We evaluated three different summarization methods such as *QuerySplitting*, *NMF*, and *PRF+NMF*. In Figure 3, the *QuerySplitting* denotes Han' method [5], the *NMF* denotes the document summarization method only using *NMF* [10], and the *PRF+NMF* denotes the proposed method. Figure 3 shows the average *f*-measure of three document summarization method using *ROUGE-N* evaluation processes. The proposed method shows better performance than the *NMF* whereas the *QuerySplitting* shows the lowest performance. The *QuerySplitting* 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 can not influence the user' requirement properly to summarize the document, in the case that the initial user' query is biased. However, the proposed method 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.

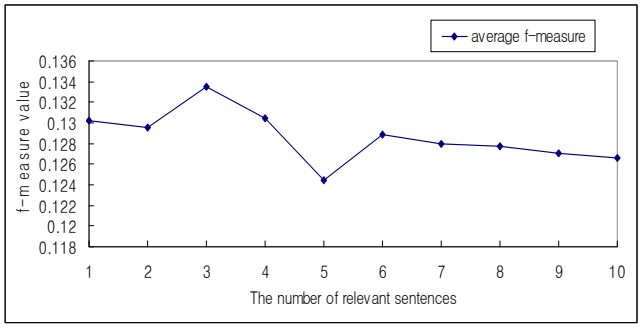


**Fig. 2:** Effectiveness of query expansion with respect to the number of relevant sentences

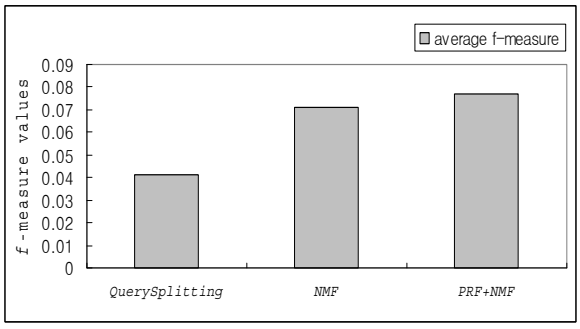


**Fig. 3:** Performance comparison with three methods

[1] http://www.isi.edu/~cyl/ROUGE/
[2] http://www-nlpir.nist.gov/projects/index.html

## 5. Conclusion

This paper proposed an automatic user-focused document summarization using the *PRF* and the *NMF*. The proposed method expands the initial user query without user's intervention to reduce the semantic gap between user's concept and query sentence vector representation. Besides, it can extract meaningful sentences since it reflects the inherent structure of a document with respect to semantic features and semantic variables by NMF. The performance of the proposed method is about 46.8% in average *f*-measure better than that of the *QuerySplitting* method, and about 8.2% in average *f*-measure better than that of the *NMF* method.

## 6. References


[1] A., Berger, V. O. Mittal. Query-Relevant Summarization using FAQs. *Proc. of Annual Meeting of the Association for Computational Linguistics*. 2000.

[2] C.Y. Lee. ROUGE: A Package for Automatic Evaluation of Summaries. *Proc. of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL*. 2004.

[3] W. B. Frankes, B. Y. Ricardo. *Information Retrieval: Data Structure & Algorithms*. Prentice-Hall, 1992.

[4] J. Goldstein, M. Kantrowitz, V. Mittal and J. Carbonell. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. *Proc. of ACM-SIGIR*. 1999, 121-128.

[5] K. S. Han, D. H. Bea, H. C. Rim. Automatic Text Summarization Based on Relevance Feedback with Query Splitting. *Proc. of International Workshop on Information Retrieval with Asia Languages*. 2000, 201-202.

[6] H. D. Hoa. Overview of DUC2005. *Proceedings of the Document Understanding Conference*. 2005.

[7] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 1999, 788-791.

[8] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *In Advances in Neural Information Processing Systems*, 13, 2001, 556-562.

[9] I. Mani. Automatic Summarization. *John Benjamins Publishing Company*, 2001.

[10] S. Park, J. H. Lee, C. M. Ahn, J. S. Hong. S. J. Chun. Query Based Summarization using Non-negative Matrix Factorization. *Proc. of Knowledge-Based Intelligent Information and Engineering Systems*. 2006, 84-89.

[11] S. Park, J. H. Lee, D. H. Kim, C. M. Ahn. Multi-document Summarization Based on Cluster Using Non-negative Matrix Factorization. *Proc. of Annual International Conference on Software Seminar*. 2007, 761-770.

[12] S. Park, J. H. Lee, D. H. Kim, C. M. Ahn. Multi-document Summarization Using Weighted Similarity between Topic and Clustering-based Non-negative Matrix Factorization. *Proc. of the Annual International Conference on Asia Pacific Web*. 2007, 761-770.

[13] S. Park, J. H. Lee, Topic-based Multi-document Summarization Using Non-negative Matrix Factorization and K-means. *Journal of KIISE: Software and Applications*. 35(4), 2008, 255-264.

[14] S. Park, J. H. Lee, D. H. Kim, C. M. Ahn. Document Summarization using Non-negative Matrix Factorization and Relevance Feedback. *Proc. of International Conference on Hybrid Information Technology*. 2008. 301-306.

[15] S. Park. Automatic Query-based Document Summarization using NMF based on Pseudo Relevance Feedback. *Proc. of Pacific Rim Knowledge Acquisition Workshop*. 2008. 101-106.

[16] B. Y. Ricardo, R. N. Berthier. Moden Information Retrieval. *ACM Press*, 1999.

[17] T. Sakurai, A. Utsumi. Query-based Multi-document Summarization for Information Retrieval. *Proc. of NTCIR*. 2004.