Query Based Summarization Using Non-negative Matrix Factorization

Sun Park[1], Ju-Hong Lee[1,*], Chan-Min Ahn[1], Jun Sik Hong[2], and Seok-Ju Chun[3]

[1] Department of Computer Science and Information Engineering, Inha University, Korea
{sunpark, ahnch1}@datamining.inha.ac.kr,
juhong@inha.ac.kr
[2] Department of Electronic Engineering, Youngdong University, Korea
jnskhong@youngdong.ac.kr
[3] Department of Computer Education, Seoul National University of Education, Korea
chunsj@snue.ac.kr

Abstract. Query based document summaries are important in document retrieval system to show the concise relevance of documents retrieved to a query. This paper proposes a novel method using the Non-negative Matrix Factorization (NMF) to extract the query relevant sentences from documents for query based summaries. The proposed method doesn't need the training phase using training data comprising queries and query specific documents. And it exactly summarizes documents for the given query by using semantic features and semantic variables without complex processing like transformation of documents to graphs because the NMF have a great power to naturally extract semantic features representing the inherent structure of a document.

1 Introduction

Query based summarization of text document becomes important with increasing the amount of information available on the internet. Text document summaries can be either generic summaries or query based summaries. A generic summary presents an overall sense of the documents' contents. A query based summary presents the contents of the document that are related to the user's query. Query based summarization are tailored to the requirements of a particular user. The summary takes into account some representation of the user's interests, which can range from user model to profiles recording subject area terms or even a specific query containing terms that are deemed expressive of a user's information need [10, 11].

The recent studies for query based summarization are as follows: Berger and Mittal proposed a method that uses frequently-asked question (FAQ) for training a query-relevant summarization. Each frequently-asked question document is comprised of questions and answers about specific topic [2].

Bosma proposed a method using the Rhetorical Structure of the document. It transforms a document into a weighted graph, in which each vertex represents a sentence and the weight of an edge represents the distance between the two sentences. It is difficult to apply the Rhetorical Structure Theory to multimodal documents without extensive

* Corresponding author.

B. Gabrys, R.J. Howlett, and L.C. Jain (Eds.): KES 2006, Part III, LNAI 4253, pp. 84 – 89, 2006.

modifications [3]. Varadarajan and Hristidis proposed a method to create query specific summaries by adding structure graph to documents by extracting associations between their fragments [15]. Mani and Bloedorn used graphs to formalize relations between sentences inside a document for multi-document summarization [12].

Sakurai and Utsumi proposed a method that generates the core part of the summary from the most relevant document to a query, and then the additional part of the summary, which elaborates on the topic, from the other documents. Their method has a beneficial effect on long summaries. But its performance is not satisfactory for the specific task [13].

Saggion used the content reduction which is a process of sentence elimination. The content reduction method removes sentences from a pool of candidate sentences until the desired compression is achieved [14].

The NMF can find a parts representaion of the data because non-negative constraints are compatible with the intuitive notions of combining parts to form a whole, which is how the NMF learns a parts-based representation [8, 9].

In this paper, we propose a novel method that makes query-based summaries by extracting sentences using the similarity between query and Non-negative Semantic Feature vectors obtained from the NMF.

The proposed method in this paper has the following advantages: First, it is an unsupervised text summarization method that doesn't require the training data comprising queries and query specific documents. Second, the NMF have a great power to naturally extract semantic features representing the inherent structure of a document. By virtue of the power of the NMF, it also can select sentences that are highly relevant to a given query because it can chooses the sentences related to the query relevant semantic features that well represent the structure of a document. Third, it can be applied to the query based summarization for multi-documents.

The rest of the paper is organized as follows: Section 2 describes the proposed method and section 3 shows the experimental results. We conclude the paper in section 4 with future researches.

2 Query Based Summarization

In this section, we propose a method that creates query-based summaries by selecting sentences using the NMF. The proposed method consists of the preprocessing step and the summarization step.

2.1 Preprocessing

In the preprocessing step, after a given document is decomposed into individual sentences, we remove all stopwords and perform words stemming. Then we construct the weighted term-frequency vector for each sentence in document by Equation (1) [1, 4, 5].

Let $T_i = [\ t_{1i}, t_{2i}, \ldots, t_{ni}\ ]^T$ be the term-frequency vector of sentence i, where elements t_{ji} denotes the frequency in which term j occurs in sentence i. Let $\boldsymbol{A}$ be $m \times n$ weighted terms by sentences matrix, where m is the number of terms and n is the number of sentences in a document. Let element A_{ji} be the weighted term-frequency of term j in sentence i.

$$A_{ji} = L(j, i) \cdot G(j, i) \tag{1}$$

Where $L(j, i)$ is the local weighting for term j in sentence i, and $G(j, i)$ is the global weighting for term j in the whole documents. That is,

$$L(j, i) = t_{ji} \tag{2}$$

$$G(j, i) = \log(N/n(j)) \tag{3}$$

Where N is the total number of sentences in the document, and $n(j)$ is the number of sentences that contain term j.

2.2 Query Based Summarization by NMF

In the summarization step, sentences are selected by using the NMF.

We perform the NMF on **A** to obtain the Non-negative Semantic Feature Matrix $\boldsymbol{W}$ and Non-negative Semantic Variable matrix $\boldsymbol{H}$ such that:

$$A \approx WH \tag{4}$$

Here $\boldsymbol{W}$ is an $m \times r$ matrix and $\boldsymbol{H}$ is an $r \times n$ matrix. Usually r is chosen to be smaller than n or m, so that the total sizes of $\boldsymbol{W}$ and $\boldsymbol{H}$ are smaller than that of the original matrix $\boldsymbol{A}$. This results in a compressed version of the original data matrix. We keep updating $\boldsymbol{W}$ and $\boldsymbol{H}$ until $\left\|A - WH\right\|^2$ converges under the predefined threshold. The update rules are as follows [8, 9]:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T A)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}} \tag{5}$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(AH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \tag{6}$$

A column vector corresponding to j'th sentence $A_{\cdot j}$ can be represented as a linear combination of semantic feature vectors $W_{\cdot l}$ and semantic variable H_{lj} as follows:

$$A_{\cdot j} = \sum_{l=1}^{r} H_{lj} W_{\cdot l} \tag{7}$$

Where $W_{\cdot l}$ is the l'th column vector of $\boldsymbol{W}$.

The powers of the two non-negative matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ are described as follows: All semantic variables (H_{lj}) are used to represent each sentence. $\boldsymbol{W}$ and $\boldsymbol{H}$ are represented sparsely. Intuitively, it make more sense for each sentence to be associated with some small subset of a large array of topics ($W_{\cdot l}$), rather than just one topic or all the topics. In each semantic feature ($W_{\cdot l}$), the NMF has grouped together semantically related terms. In addition to grouping semantically related terms together into

semantic features, the NMF uses context to differentiate between multiple meanings of the same term [8].

To evaluate the degree of similarity of the semantic feature vector $W_{\cdot l}$ with regard to the query $\vec{q}_j$ as the correlation between the vector $W_{\cdot l}$ and $\vec{q}_j$. This correlation can be quantified, for instance, by the cosine of the angle between these two vectors [1, 4, 5]. That is,

$$sim(W_{\cdot l}, \vec{q}_j) = \frac{W_{\cdot l} \cdot \vec{q}_j}{|W_{\cdot l}| \times |\vec{q}_j|} = \frac{\sum_{i=1}^{m} w_{i,l} \times q_{i,j}}{\sqrt{\sum_{i=1}^{m} w_{i,l}^2} \times \sqrt{\sum_{i=1}^{m} q_{i,j}^2}} \tag{8}$$

Where $|W_{\cdot l}|$ and $|\vec{q}_j|$ are the norms of the semantic feature vector and query vectors.

We propose the following query based summarization method:

1. Decompose the document ***D*** into individual sentences, and let *k* be the number of sentences for summarization.
2. Perform the stopwords removal and words stemming operations.
3. Construct the weighted terms by sentences matrix ***A*** using Equation (1).
4. Perform the NMF on the matrix ***A*** to obtain the matrix ***W*** and the matrix ***H*** using Equation (5) and (6).
5. Select a column vector $W_{\cdot p}$ of matrix ***W*** whose similarity to the query is the largest using Equation (8).
6. Select the sentence corresponding to the largest index value of the row vector $H_{p\cdot}$, and include it in the summary.
7. If the number of selected sentences reaches the predefined number *k*, then stop the algorithm. Otherwise go to step 5 to find the next most similar column vector excluding $W_{\cdot p}$.

In step 5, the fact that the similarity between $W_{\cdot p}$ and the query is largest means the *p*'th semantic feature vector $W_{\cdot \mathrm{p}}$ is the most relevant feature to the query. In step 6, it select the sentences that has the largest weight for the most relevant semantic feature.

3 Experimental Results

As an experimental data, we used Yahoo Korea News [6]. We gave 5 queries to retrieve news documents from Yahoo Korea News. The retrieved news documents are preprocessed using HAM (Hangul Analysis Module) which is a Korean language analysis tool based on Morpheme analyzer [7]. The evaluator was employed to manually create the query based summaries for the retrieved Yahoo Korea news documents. Table 1 provides the particulars of the evaluation data corpus.

Table 1. Particulars of the Evaluation Data Corpus (Yahoo Korea News)

Document attributes	values
Number of docs	50
Number of docs with more than 10 sentences	42
Avg sentences / document	16
Min sentences / document	2
Max sentences / document	29

We used the precision (P) to compare the performances of the two summarization methods, Saggion's method[14] and our method. We modified a Saggion's method for the experimental environment. Let S_{man}, S_{sum} be the set of sentences selected by the human evaluators, and the summarizer, respectively. The standard definition of the precision is defined as follows [1, 4, 5]:

$$P = \frac{|S_{man} \cap S_{sum}|}{|S_{sum}|} \tag{9}$$

The evaluation results are shown in figure 1.

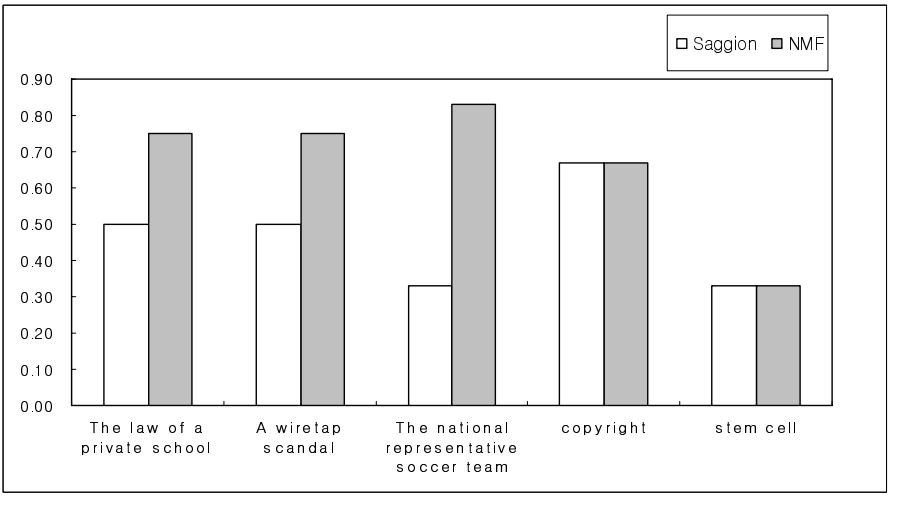

Fig. 1. Evaluation Results

Experimental results show that the proposed method surpassed the Saggion's method. This is because the NMF have the great power to grasp the innate structure of a document like human's cognition process [8].

4 Conclusions

In this paper, we proposed a novel method that makes query based summaries by extracting sentences using the NMF. The proposed method has the following

advantages: it doesn't require the training data comprising queries and query specific documents. By virtue of the power of the NMF that have a great power to naturally grasp the innate structure of a document. It can select sentences that are highly relevant to a given query. It also can be used to make the summaries for multi-documents.

In future work, we have a plan to evaluate our method on various term weighting schemes. And we will study the relation between Non-negative Semantic Feature Matrix $\boldsymbol{W}$ and Non-negative Semantic Variable Matrix $\boldsymbol{H}$ for performance elevation of the summarization.

Acknowledgment. This work was supported by the Brain Korea 21 Project in 2006.

References

1. Baeza-Yaters, R., Ribeiro-Neto, B.: Modern Information Retrieval, ACM Press (1999)
2. Berger, A., Mittal, V. O.: Query-Relevant Summarization using FAQs. In Proceeding of the 38th Annual Meeting on Association for Computational Linguistics (ACL'00), (2000)
3. Bosma, W.: Query-based Summarization using Rhetorical Structure Theory. In Proceeding of the 15th Meeting computational Linguistics in the Netherlands (CLIN'04), (2004)
4. Chakrabarti, S.: mining the web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann. 67-74 (2003)
5. Frankes, W. B., Baeza-Yates, R.: Information Retrieval : Data Structure & Algorithms, Prentice-Hall (1992)
6. Http://kr.news.yahoo.com (2005)
7. Kang, S. S.: Information Retrieval and Morpheme Analysis. HongReung Science Publishing Co. (2002)
8. Lee, D. D. and Seung, H. S.: Learning the parts of objects by non-negative matrix factorization, Nature, 401:788-791 (1999)
9. Lee, D. D. and Seung, H. S.: Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems, volume 13, pages 556-562 (2001)
10. Mani, I.: Automatic Summarization. John Benjamins Publishing Company (2001)
11. Mani, I., Maybury, M. T.: Advances in automatic text summarization. The MIT Press (1999)
12. Mani, I., Bloedorn, E.: Multidocument summarization by graph search and matching. In Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI'97), (1997)
13. Sakurai, T., Utsumi, A.: Query-based Multidocument Summarization for Information Retrieval. In Proceeding of the Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Summarization Workshop (NTCIR'04), (2004)
14. Sassion. H.: Topic-based Summarization at DUC 2005. In Proceedings of the Document Understanding Conference 2005 (DUC'05), (2005)
15. Varadarajan, R., Hristidis, V.: Structure-Based Query-Specific Document Summarization. In Proceeding of the ACM Fourteenth Conference on Information and Knowledge Management (CIKM'05), (2005)