

# 의미특징의 포괄적 중요도를 이용한 포괄적 문서 요약

## Generic Summarization Using Generic Important of Semantic Features

박 선\*, 이종훈\*\*

Sun Park\*, Jong-Hoon Lee\*\*

### 요 약

인터넷의 급속한 확산과 대량 정보의 이동은 문서요약을 더욱 필요 하게 하고 있다. 본 논문은 비음수 행렬 인수분해로 얻어진 비음수 의미 가변 행렬과 의미특징의 포괄적 중요도를 이용하여 문장을 추출하여서 포괄적 문서요약을 하는 새로운 방법을 제안하였다. 제안된 방법은 인간의 인식 과정과 유사한 비음수 제약을 사용한다. 이 결과 주제의 군집방법이나 잠재의미분석을 사용한 비지도 학습방법에 비해 더욱 의미 있는 문장을 선택하여 문서를 요약할 수 있다. 실험결과 제안방법이 다른 방법들에 비하여 좋은 성능을 보인다.

### Abstract

With the increased use of the internet and the tremendous amount of data it transfers, it is more necessary to summarize documents. We propose a new method using the non-negative semantic variable matrix (NSVM) and the generic important of semantic features obtained by Non-negative Matrix Factorization (NMF) to extract the sentences for automatic generic summarization. The proposed method use non-negative constraints which is more similar to the human's cognition process. As a result, the proposed method selects more meaningful sentences for summarization than the unsupervised method used the Latent Semantic Analysis (LSA) or clustering methods. The experimental results show that the proposed method achieves better performance than other methods.

Key words : non-negative variable matrix, non-negative matrix factorization, generic document summarization, unsupervised method

### I. 서 론

인터넷의 발전은 끊임없이 변화하는 정보를 증가시키고 있다. 이러한 정보의 증가는 정보 집합 내에 어떠한 정보가 있는지 검색할 수 있는 능력을 간절히 요구한다. 이 때문에 효율적인 정보검색을 위하

여 정보요약의 필요성이 증가하고 하고 있다. 현재 인터넷 상에서의 정보검색은 대부분 문자를 기반으로 이루어져 있기 때문에 문서요약에 대한 필요성이 점차 증가하는 추세에 있다.

문서 요약은 문서의 기본적인 내용을 유지하면서 문서의 량을 줄이는 작업이다[1]. 문서의 요약은 문

\* 호남대학교 컴퓨터공학과(Department of computer Engineering, Honam University)

\*\* 호남대학교 다매체영상학과(Department of MultiMedia & Film, Honam University)

· 교신저자 (Corresponding Author) : 이종훈

· 투고일자 : 2008년 9월 16일

· 심사(수정)일자 : 2008년 9월 17일 (수정일자 : 2008년 10월 15일)

· 게재일자 : 2008년 10월 30일

서 내용전체를 요약하는 포괄적 요약(generic summary)와 사용자의 질의에 따라 질의에 관련 있는 내용만을 포함하는 질의 중심 요약(query-based summary)으로 나눌 수 있다[1]. 또한 문서요약에 적용하는 방법에 따라서 지도 학습(supervised approaches)[2, 3],과 비지도 학습(unsupervised approaches) 방법[4, 5, 6, 7, 8, 9, 10, 11]으로 나눌 수 있다.

비지도 학습 방법은 문서를 요약하는데 사람이 작성한 요약문장이 필요 없는데 비하여, 지도 학습 방법은 요약 방법에 사용되는 특징이나 사람이 만든 요약문을 필요로 한다.

지도학습방법에 의한 최근의 문서요약은 다음과 같다. Chuang과 Yang은 문장을 추출하여 자동으로 문서를 요약하는 방법을 제안하였다[3]. 이 방법은 특징 벡터를 기반으로 한 중요 문장을 추출을 위하여 요약기(summarizer)를 훈련하는 비지도 학습 알고리즘을 이용하였다[3]. 이들의 방법은 훈련문서집합을 요구하며 빠른 시간 내에 요약을 해야 하는 응용 분야에는 적합하지 않은 단점을 가진다. Amini와 Gallinari는 준지도 학습 방법(semi-supervised learning)의 알고리즘을 이용한 문서 요약 방법을 제안하였다[2]. 이 방법은 대량의 레이블이 없는 자료를 위해 소량의 레이블 자료를 이용한다[2]. 이들의 방법은 요약결과와 관련된 문장이 필요하며, 요약 시스템의 훈련을 위해서 레이블링 해야 하는 단점을 갖는다.

비지도 학습 방법의 문서요약은 다음과 같다. LSA(latent semantic analysis, 잠재의미분석)를 이용한 방법으로, Gong과 Liu는 문서를 요약하였고[4], Sum과 Shen은 웹 페이지를 요약하였다[7]. LSA를 이용한 문서요약방법은 문장을 추출하기 위하여 고유값으로 구성된 의미특징을 이용한다. 고유값이 양수와 음수 값을 갖기 때문에 의미특징을 직관적으로 이해할 수 없으며, 의미가 적은 문장을 추출할 수 있는 단점이 있다[9]. 문서의 주제(topic)을 이용한 방법으로, Nomoto와 Matsumoto는 변형된 k-means를 이용하여 문서에서 다양한 주제를 찾은 후, 각 주제에 일치하는 문장을 선택하여 문서를 요약 하였다[5]. Zha는 문장과 용어(terms)로부터 특징점수를 계산하고, 이

를 이용하여 문장들을 주제그룹들로 군집하여 문서를 요약하였다[9]. 박선 등은 비음수 의미 가변 행렬을 이용한 포괄적 문서요약을 제안하였다. 이 방법은 의미특징의 값은 높으나 전체 문서들에서 별로 중요하게 나타나지 않는 문장들이 선택될 수 있는 문제를 가지고 있다[6, 10, 11].

본 논문은 비음수 행렬 분해(NMF, non-negative matrix factorization)에 의해 분해된 의미변수를 이용하여 포괄적 문서요약을 하는 새로운 방법을 제안하였다. 비음수 행렬 인수분해는 Lee와 Seung이 제안한 방법으로 다변량 자료를 유용하게 분해하는 알고리즘이다[8, 12, 13].

제안된 방법은 다음과 같다. 요약할 문서를 문장으로 분해하고, 분해된 문장들은 벡터모델에 따라서 벡터로 표현한다. 비음수 행렬 분해를 이용하여 문장 벡터 행렬을 의미특징 행렬과 의미변수 행렬로 분해한다. 의미 변수의 포괄적 중요도를 계산하고, 포괄적 중요도가 높은 의미변수 행벡터를 선택한다. 여기서 문장벡터들은 의미특징 벡터의 가중치인 의미변수를 곱한 값의 선형 합으로 표시된다. 의미특징 벡터는 문장의 내부 특징을 나타내며, 의미변수는 문장 내에서 의미특징의 중요도를 나타낸다. 즉, 문장과 대응되는 의미변수의 값이 크다는 것은 그 문장이 중요하다는 것을 의미한다. 선택된 의미특징 벡터에서 의미변수가 가장 큰 값과 대응되는 문장을 선택하고, 문장의 수가 미리 설정된 요약문의 개수를 만족할 때까지 의미변수의 값이 다음으로 큰 값과 대응되는 문장들을 선택하여 문서를 요약한다.

제안된 방법은 다음과 같은 장점을 갖는다. 첫째, NMF에 의해 찾아 지는 의미 특징(semantic feature)들이 비 음수 값을 갖기 때문에 잠재의미분석에 비해 의미 있는 문서요약 결과를 갖는다. 둘째, 의미변수만을 이용하여 적은 계산비용으로 쉽게 문장을 추출할 수 있다. 셋째, 의미변수에 포괄적 중요도를 계산하여서 의미특징의 값은 높으나 전체 문서들에서 별로 중요하게 나타나지 않는 문장들이 선택되는 것을 최소화 할 수 있다. 마지막으로, 비지도 학습방법으로 학습문장이 필요 없는 장점을 갖는다.

본 논문의 구성은 다음과 같다. 제2장에서는 비음수 행렬 분해 방법을, 제3장은 제안한 요약방법을, 제4장에서는 실험 및 평가에 대해 기술한다. 마지막

으로 제5장에서 결론은 맺는다.

### II. 비음수 행렬 분해

비음수 행렬 분해(NMF, non-negative matrix factorization)는 주어진 비음수 행렬로부터 비음수의 인수를 찾는 행렬분해 알고리즘이다[12, 13]. 본 논문에서 행렬  $X$ 의  $j$ 번째 열벡터는  $X_{*j}$ 로,  $i$ 번째 행벡터는  $X_{i*}$ 로,  $i$ 번째 행과  $j$ 번째 열의 원소는  $X_{ij}$  표시한다. 비음수 행렬 분해 알고리즘은 식(1)의 목표 함수  $J$ 가 0에 가깝게 수렴 할 때까지 식(2)과 식(3)를 이용하여 행렬  $W$ 와  $H$ 의 값을 동시에 갱신한다.

$$J = \|A - WH\|^2 \quad (1)$$

식(1)의 목적은 행렬  $A$ 를 비음수  $m \times r$  행렬  $W$ 와 비음수  $r \times n$  행렬  $H$ 로 분해하는 것이다. 여기서,  $A$ 는  $m$ 개의 용어와  $n$ 개의 문장으로 이루어진  $m \times n$  행렬이고,  $r$ 은 의미특징의 개수이다.

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T A)_{\alpha\mu}}{(W^T W H)_{\alpha\mu}} \quad (2)$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(A H^T)_{i\alpha}}{(W H H^T)_{i\alpha}} \quad (3)$$

행렬  $A$ 의  $j$ 번째 열벡터  $A_{*j}$ 는 행렬  $W$ 의  $i$ 번째 열벡터  $W_{*i}$ 와 행렬  $H$ 의 요소  $h_{ij}$ 가 선형조합을 이루며 식(4)과 같다.

$$A_{*j} = \sum_{i=1}^r h_{ij} W_{*i} \quad (4)$$

예1) 다음은 식(1)을 이용하여  $A$ 행렬을  $W$ 와  $H$ 행렬로 분해 한 예이다.  $r = 2$ , 수렴할 반복 횟수는 50 이고, 수렴 허용오차가 0.001이다.  $W$ 와  $H$ 행렬의 초기 값은 각각 0.5이다.

$$\begin{matrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix} & \approx & \begin{bmatrix} 0.15 & 1.60 \\ 0.66 & 0.97 \\ 1.15 & 0.57 \\ 1.61 & 0.41 \end{bmatrix} & \times & \begin{bmatrix} 6.11 & 6.68 & 7.18 \\ 0.09 & 0.60 & 1.22 \end{bmatrix} & = & \begin{bmatrix} 1.05 & 1.94 & 3.03 \\ 4.12 & 4.98 & 5.93 \\ 7.07 & 8.01 & 8.95 \\ 9.90 & 11.01 & 12.08 \end{bmatrix} \\ A & & W & & H & & \tilde{A} \end{matrix}$$

### III. 질의기반의 문서요약

본 장에서는 NMF를 기반으로 문장을 추출하여 포괄적 문서요약을 할 수 있는 방법을 제안한다. 제안 방법은 전처리 단계와 문서요약 단계로 이루어진다. 다음 에서 두 단계에 대하여 자세히 기술한다. 다음 그림1은 비음수 의미 가변 행렬을 이용한 포괄적 문서요약의 개요이다.

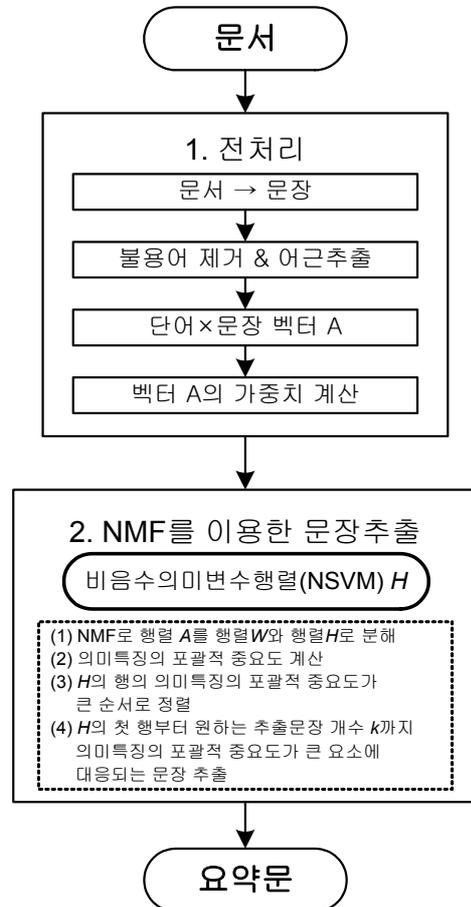


그림 1. 비음수 가변행렬의 중요도를 이용한 포괄적 문서요약

Fig. 1. Generic document summarization method using the important of semantic variable matrix

## 3-1 전처리 단계

전처리 단계는 주어진 문서를 각각의 문장으로 분해 후, 불용어 제거, 어근추출, 가중치 계산으로 이루어진다[14, 15, 16]. 이후 용어빈도 벡터를 생성하고 식(5)을 이용하여 가중치를 계산한다[14, 15, 16].

벡터  $T_i = [t_{1i}, t_{2i}, \dots, t_{ni}]^T$ 는  $i$ 번째 문장의 용어빈도이다. 여기서 요소  $t_{ji}$ 는  $i$ 번째 문장에서 출현한  $j$ 번째 용어의 빈도이다.  $i$ 번째 문장  $A_i$ 는 가중치가 부여된 용어빈도 벡터  $A_i = [a_{1i}, a_{2i}, \dots, a_{ni}]^T$ 로 표현되고, 벡터  $A_i$  요인  $a_{ji}$ 는 식(5)와 같이 정의된다.

$$a_{ji} = L(t_{ji}) \cdot G(t_{ji}) \quad (5)$$

여기서  $L(t_{ji})$ 는  $i$ 번째 문장에서  $j$ 번째 용어를 위한 지역 가중치(local weight)이고,  $G(t_{ji})$ 는 문서 전체에서  $j$ 번째 용어를 위한 전역 가중치(global weight)로 다음 식(6), (7)와 같이 정의된다.

$$L(i) = t(i) \quad (6)$$

$$G(i) = \log(N/n(i)) \quad (7)$$

여기서,  $t(i)$ 는 문장에서  $i$ 번째 용어가 출현한 빈도,  $N$ 은 문서에서 문장의 총 개수이다.  $n(i)$ 는  $i$ 번째 용어를 포함한 문장의 개수이다.

## 3-2 문서요약 단계

비음수 행렬 분해를 이용한 문서요약 단계는 다음과 같다. 문서를 전처리하여서 총  $m$ 개의 용어와  $n$ 개의 문장으로 이루어진  $m \times n$  행렬  $A$ 는  $A = [A_1, A_2, \dots, A_n]$ 로 나타내며, 각 행 벡터  $A_i$ 는  $i$ 번째 문장의 가중치 용어빈도 벡터이다.

주어진 행렬  $A$ 를 비음수 행렬 인수분해 하여 얻어지는 비음수 의미 특징 행렬(NSFM, non-negative semantic feature matrix)  $W$ 와 비음수 의미 변수 행렬(NSVM, non-negative variable matrix)  $H$ 는 다음 식(8)과 같다[8, 12, 13].

$$A \approx WH \quad (8)$$

여기서,  $A$ 는  $n \times m$  행렬이고,  $n \times r$  행렬  $W$ 와  $r \times m$  행렬  $H$ 는 행렬  $A$ 로부터 근사 값으로 분해된 행렬이며, 각각  $W = [w_1, w_2, \dots, w_r]$ ,  $H = [h_1, h_2, \dots, h_r]$ 로 나타낸다.

의미특징의 포괄적 중요도를 식 (9)와 같이 정의한다. 식(9)의 의미특징의 포괄적 중요도는 의미특징에 의해서 표현되는 주요주체가 문장에 얼마나 반영되는가를 의미한다. 즉, 식(9)를 사용함으로써 문서에 포함된 문장들 중 가장 중요한 주체를 포함한 문장을 선택 할 수 있다.

*Generic Important of Semantic Variable (GISV)*

$$= \left( \sum_{i=1}^n H_{pi} \right) \cdot \text{weight}(H) \quad (9)$$

$$\text{weight}(H_{j*}) = \frac{\sum_{q=1}^n H_{jq}}{\sum_{p=1}^m \sum_{q=1}^n H_{pq}} \quad (10)$$

여기서  $\text{weight}(H_{j*})$ 는 모든 의미특징들 중에서  $j$ 번째 의미특징 ( $W_{j*}$ )의 적합도와의 관계를 의미하고, 식(9)의 문장의 적합도는 의미특징에 의해서 표현되는 주요주체가 문장에 얼마나 반영되는가를 의미한다. 즉, 식(10)을 사용함으로써 의미특징의 값은 높으나 전체 문서들에서 별로 중요하게 나타나지 않는 문장들이 선택되는 것을 최소화 할 수 있다.

## 3-3 포괄적 문서요약 알고리즘

비음수 의미 변수를 이용한 포괄적 문서요약 알고리즘은 다음과 같다.

1. 전처리 단계를 수행한다.

(1) 문서를 개개의 문장으로 분해하고, 분해할 계수  $r$ 를 지정한다.

(2) 각각의 문장에 대한 불용어 제거 및 어근추출 후, 식(5)을 이용하여 용어빈도 벡터

의 가중치를 계산하여 용어문장 가중치 행렬  $A$ 를 구성한다.

2. 행렬  $A$ 에 식(2)과 식(3)을 적용하여 식(8)과 같은 비음수 행렬  $W, H$ 로 분해한다.
3. 의미변수 행렬  $H$ 에 식(9)과 식(10)을 이용하여 모든 의미변수 행벡터에 대하여 문장의 포괄적 중요도를 계산하여 정렬 한다. 포괄적 중요도가 큰 순서대로 요약문장개수인  $k$ 개만 큼 선택한다.
4. 선택된  $k$ 개의 행 벡터 각각에서, 행에서 가장 큰 요소 값을 가진  $q$ 열과 같은 열에 있는 행렬  $A$ 의 문장 벡터  $A_{*q}$ 에 대응되는 문장을 선택한다.
5. 선택한 문장을 포괄적 중요도가 큰 순서대로 정렬하여 요약문을 생성한다.

위의 3번째 단계에서 포괄적 중요도가 가장 높은  $H_{D*}$ 는 대응되는 의미특징 열 벡터  $W_{*p}$ 의 중요도를 나타낸다.

#### IV. 실험 및 평가

본 논문에서 제안한 방법을 실험하기 위하여 Reuters-21578 컬렉션과 야후코리아 뉴스 에서 200건의 기사를 무작위로 선택하여 실험 자료로 사용하였다. 제안방법을 비교하기 위하여 세 명의 평가자가 문서를 수동으로 요약하였다. 다음 표1은 평가 자료에 대한 특성을 나타낸다.

표 1. 평가자료에 대한 특성  
Table 1. Property of the test data set

문서의 속성	Reuters-21578	야후 코리아
문서의 수	100	100
10문장 이상인 문서의 수	35	32
문서당 평균 문장의 수	10.09	10.1
최소 문장의 수	3	3
최대 문장의 수	40	36

성능 평가는 문헌요약에서 주로 사용되는 정확률 (Precision), 재현율(Recall), F-measure를 이용하였

다[15]. 평가척도는 다음 식(11) 이다.

$$R = \frac{|S_{man} \cap S_{sum}|}{|S_{sum}|}, P = \frac{|S_{man} \cap S_{sum}|}{|S_{man}|}, F = \frac{2RP}{R+P} \quad (11)$$

여기서  $S_{man}, S_{sum}$ 은 각각 사람과 제안된 방법에 의하여 선택된 문장이다.

실험은 네 가지 요약방법들에 대하여서 성능을 비교 평가한 것이다. 그림 2, 3은 Reuters-21578 과 야후 코리아 뉴스에 대한 각각의 평균 재현율, 평균 정확률, 평균 F-measure 에 대한 평가 결과이다. 여기서 KMEANS(kmeans, k평균)방법은 K-means군집방법을 이용한 방법[5, 14], LSA는 Gong[4]의 방법으로 잠재 의미 분석을 이용한 문서 요약 방법, NMF은 이전에 저자들[6, 10, 11]이 제안한 방법으로 의미특징 변수를 이용한 문서요약 방법이다. GINMF(generic important NMF, 포괄적 중요 비음수 행렬)방법은 본 논문에서 제안한 방법이다.

그림 2에서 보는 것과 같이 Reuters-21578를 이용한 평가 결과에서는 제안 방법(GINMF)의 평균 재현율, 정확율, F-measure가 KMEANS에 비하여 19.64%, 24.32%, 33.6%가 높으며, LSA에 비해서는 12.50%, 8.12%, 29.04%가 높고, NMF에 비해서는 7.14%, 2.70%, 5.39%가 높다.

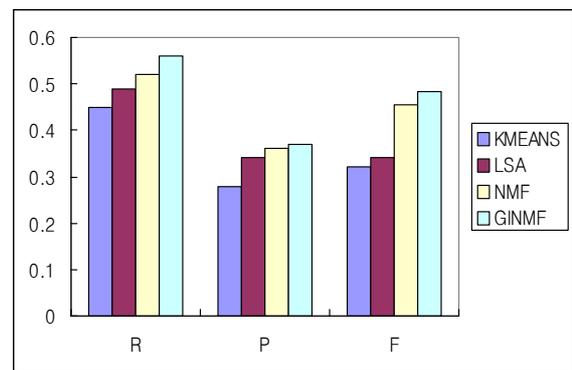


그림 2. Reuters-21578을 이용한 방법의 비교 결과  
Fig. 2. Result of comparison methods using Reuters-21578

그림 3에서 보는 것과 같이 야후 코리아 뉴스를 이용한 평가 결과에서는 제안 방법(GINMF)의 평균

재현율, 정확율, F-measure가 KMEANS에 비하여 21.05%, 29.78%, 35.84%가 높으며, LSA에 비해서는 11.65%, 10.11%, 26.55%가 높고, NMF에 비해서는 4.14%, 7.30%, 7.08%가 높다.

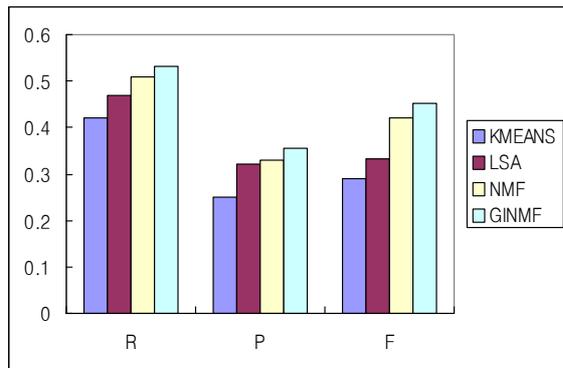


그림 3. 야후코리아 뉴스를 이용한 방법의 비교 결과  
Fig. 3. Result of comparison methods using Yahoo Korea News

성능평가결과 제안방법인 GINMF이 가장 좋은 결과를 나타내며, 그 다음으로 NMF가 좋은 결과를 나타내고, 다음으로 LSA 순으로 평가되었다. KMEANS가 가장 성능이 저조하다. 이는 KMEANS의 같은 주제별로 군집하여 요약하는 방법보다는 LSA의 문서의 잠재구조가 요약에 더 좋은 결과를 보이는 것을 알 수 있다. 또한 LSA보다는 NMF의 의미 특징이 좀 더 의미 있는 요약문을 생성하는 것을 알 수 있다. 제안방법은 의미 변수에 가중치를 주어서 의미특징의 값이 높으나 실제 중요하지 않는 문장이 추출되는 것을 최소화 시켜서 비교방법들에 비하여 더욱 의미 있는 문장을 추출함을 알 수 있다.

## V. 결 론

본 논문에서는 비음수 의미 변수에 포괄적 중요도를 적용하여서 포괄적 문서요약을 하는 방법을 제안하였다. 우리가 제안한 방법은 희박한 의미변수를 사용하기 때문에 의미적으로 중요한 문장을 추출하고, 의미변수를 비음수 값으로 표현하기 때문에 직관적으로 이해하기 쉬운 형태로 표현된다. 제안된 방법은 다음과 같은 장점을 갖는다. 제안방법은 희

박한 비음수 의미변수를 사용하기 때문에 잠재 의미 특징을 이용하는 방법에 비하여 더 중요한 문장을 추출한다. 의미변수의 값을 이용하여 문장을 추출하기 때문에 사용자가 원하는 하위주제에 적합한 문장을 추출한다. 또한, 의미변수에 포괄적 중요도를 계산하여서 의미특징의 값은 높으나 전체 문서들에서 별로 중요하게 나타나지 않는 문장들이 선택되는 것을 최소화 한다. 제안 방법은 비지도 학습방법으로 학습문장이 필요 없이 문서의 고유특징만을 이용하여 문장을 추출한다. 실험결과 군집을 이용한 방법, 잠재의미구조를 이용한 방법, 단순 의미 변수를 이용한 방법에 비하여 더 좋은 평가 결과를 보였다.

앞으로 제안 방법의 성능 향상을 위하여 다양한 종류의 가중치 및 전처리 방안에 대한 연구와, 문서의 크기에 따른 추출 문장 k의 개수를 자동으로 선택할 수 있는 방법에 대한 연구가 진행 되어야 할 것이다.

## 참 고 문 헌

- [1] I. Mani, M. T. Maybury, "Advances in Automatic Text," The MIT Press, 1999.
- [2] M. R. Amini, P. Gallinari, "The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization," *In Proceeding of ACM SIGIR'02*, pp.105-112, 2002.
- [3] W. T. Chuang, J. Yang, "Extracting Sentence Segments for Text Summarization: A Machine Learning Approach," *In Proceeding of ACM SIGIR'00*, pp.152-159, 2000.
- [4] Y. Gong, X. Liu, 'Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis', *In proceeding of ACM SIGIR'01*, pp.19-25, 2001.
- [5] T. Nomoto, 'Y. Matsumoto, "A New Approach to Unsupervised Text Summarization," *In proceeding of ACM SIGIR'01*, pp.26-34, 2001.
- [6] S. Park, "Generic Summarization Using Non-negative Semantic Variable," *Lecture Notes in Computer Science 5226*, Springer, pp.1052-1058, 2008.

- [7] J. T. Sum, D. Shen, H. J. Zeng, Q. Yang, Y. Lu, Z. Chen, "Web-Page Summarization Using Clickthrough Data," *In proceeding of ACM SIGIR'05*, pp.194-201, 2005.
- [8] W. Xu, X. Liu, Y. Gong, "Document clustering based on non-negative matrix factorization," *In ACM SIGIR*, 2003.
- [9] H. Zha, "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering," *In proceeding of ACM SIGIR'02*, pp.113-120, 2002.
- [10] 박선, 이주홍, 안찬민, 박태수, "비음수 의미 가변 행렬을 기반으로 한 자동 포괄적 문서 요약," 2006 *한국컴퓨터종합박술대회 논문집* vol. 33, No.1 (A), 2006.
- [11] 박선, 이주홍, 안찬민, 박태수, 김재우, 김덕환, "비음수 행렬 인수분해를 이용한 일반적 문서 요약," *제25회 한국정보처리학회 춘계학술발표대회 논문집*, 제13권, 제1호, 2006.
- [12] D. D. Lee, H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp.788-791, 1999.
- [13] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization," *In Advances in Neural Information Processing Systems*, vol. 13, pp.556-562, 2001.
- [14] S. Chakrabarti, "Mining the Web : Discovering Knowledge from Hypertext Data," Morgan Kaufmann, 2003.
- [15] W. B. Franke, R. Baeza-Yaes, "Information Retrieval : Data Structure & Algorithms," Prentice-Hall, 1992.
- [16] S. S. Kang, "Information Retrieval and Morpheme Analysis," HongReung Science Publishing Co., 2002.

### 박 선 (朴仙)



1996년 2월 : 전주대학교 전자계산학과(이학사)

2001년 8월 : 한남대학교 정보산업대학원 정보통신학과(공학석사)

2007년 8월 : 인하대학교 컴퓨터정보공학과 (공학박사)

2008~현재 : 호남대학교 컴퓨터공학과 전임강사

관심분야 : 정보검색, 데이터마이닝, 데이터베이스

### 이 종 훈 (이종훈)



2000년 8월 : 서울대학교 미술대학 서양화과(미술학사)

2007년 9월 ~ 현재 : 동국대학교 영상대학원 문화콘텐츠학과 석사과정

2001년 3월 ~ 2003년 3월 : 주식회사 포스트이즘 디자이너

2003년 3월 ~ 2004년 7월 : 씨지아이링크 주식회사 디자이너

2004년 7월 ~ 2005년 2월 : 레드어쓰 디자이너

2005년 3월 ~ 현재 : 호남대학교 다매체영상학과 전임강사

관심분야 : 영상, 콘텐츠, 디지털콘텐츠, 지적재산,