비음수 행렬 분해와 K-means를 이용한 주제기반의 다중문서요약

(Topic-based Multi-document Summarization Using Non-negative Matrix Factorization and K-means)

박 선 † 이 주 홍 ††
(Sun Park) (Ju-Hong Lee)

요 약 본 논문은 K-means과 비음수 행렬 분해(NMF)를 이용하여 주제기반의 다중문서를 요약하는 새로운 방법을 제안하였다. 제안방법은 비음수 행렬 분해를 이용하여 가중치가 부여된 용어-문장 행렬을 희소(sparse)한 비음수 의미특징 행렬과 비음수 변수 행렬로 분해함으로써 직관적으로 이해할 수 있는 형태의 의미적 특징을 추출할 수 있고, 주제와 의미특징간의 유사도에 가중치를 부여하여 유사도는 높으나 실제 의미 없는 문장이 추출되는 것을 막는다. 또한 K-means 군집을 이용하여 문장에 포함된 노이즈를 제거함으로써 문서의 의미가 요약에 편향되게 반영하는 것을 피할 수 있고, 추출된 문장에 부여된 순위순서대로 정렬하여 보여 줌으로써 응집성을 높인다. 실험 결과 제안방법이 다른 방법에 비하여 좋은 성능을 보인다.

키워드 : 다중문서 요약, 비음수 행렬 분해, 군집, 주제기반 요약, 가중치 유사도

Abstract This paper proposes a novel method using K-means and Non-negative matrix factorization (NMF) for topic-based multi-document summarization. NMF decomposes weighted term by sentence matrix into two sparse non-negative matrices: semantic feature matrix and semantic variable matrix. Obtained semantic features are comprehensible intuitively. Weighted similarity between topic and semantic features can prevent meaningless sentences that are similar to a topic from being selected. K-means clustering removes noises from sentences so that biased semantics of documents are not reflected to summaries. Besides, coherence of document summaries can be enhanced by arranging selected sentences in the order of their ranks. The experimental results show that the proposed method achieves better performance than other methods.

Key words : multi-document summarization, non-negative matrix factorization, clustering, topic-based summarization, weighted similarity

1. 서 론

문서 요약은 문서의 기본적인 내용을 유지하면서 문서의 양을 줄이는 작업이다. 즉 문서에서 가장 중요한 내용을 추출하는 것이다. 문서 요약 방법은 문서 내용전체를 요약하는 일반 요약과 사용자나 사용자 그룹의 특별한 요구에 따라 요약하는 사용자 기반(혹은 주제나 질의 기반) 요약으로 나눌 수 있다. 또한 적용 문서의 개수에 따라서 나눌 수 있는데, 단일문서요약은 하나의 문서를 요약하는 것이며, 다중문서요약은 관련된 여러 개의 문서를 요약하는 것이다[1].

Radev와 Hovy는 다중문서요약을 위해서 고려해야 하는 세가지 문제점을 언급하였다. 첫째는 중복성을 찾아서 제거하는 것이며, 둘째는 문서들 간의 중요한 차이점을 식별하는 것이다. 셋째는 요약의 일관성을 보장하는 것이다[2]. 중복성은 문서에 포함된 용어들이나 개념들이 얼마나 반복되는가에 나타내며, 다양성이나 차이점

† 정 회 원 : 호남대학교 컴퓨터공학과 교수
sunpark@honam.ac.kr
†† 종신회원 : 인하대학교 컴퓨터 정보공학과 교수
juhong@inha.ac.kr
논문접수 : 2007년 5월 31일
심사완료 : 2008년 2월 12일

정보과학회논문지 : 소프트웨어 및 응용 제35권 제4호(2008.4)

은 문서에 포함된 용어들이나 개념들이 다른 정도를 나타낸다[3]. 또한 일관성은 요약을 사용자가 쉽게 읽을 수 있거나 주어진 주제와의 적합성의 정도를 나타낸다[4,5].

본 논문은 Nonnegative Matrix Factorization(NMF)과 군집방법을 이용하여 주제기반의 다중문서를 요약하는 새로운 방법을 제안하였다. NMF는 Lee와 Seung이 제안한 방법으로서, 인간이 객체의 부분 정보의 조합으로 객체를 인식할 때 비 음수 자료들의 덧셈만을 사용하는 것에 착안하여, 비음수로 표현된 대량의 객체자료들로부터 부분 객체정보들을 추출하여, 개개의 객체들을 추출된 부분 객체들의 비음수 선형조합으로 표현할 수 있게 하는 방법이다. 이 방법은 원 비음수 행렬이 두 개의 비음수 희소행렬로 분해되므로 대량의 정보를 효율적으로 처리할 수 있다[6,7].

제안된 방법은 다음과 같다. 다중문서로부터 문장을 분해하고, 분해된 문장들은 Vector Model에 따라서 vector로서 표현된다[8]. 이들 문장은 *Kmeans*을 이용하여 유사한 문장의 집합으로 군집된다. 중복 문장이 많을수록 문장간의 유사도가 높기 때문에 군집 내의 문장들의 중복성이 나타나고, 군집들 간의 유사도의 차이에 의해서 다양성이 표현된다. 문장수가 적은 군집은 노이즈로 간주되어 제거된다. 각각의 군집을 나타내는 행렬은 NMF를 이용하여 비음수 의미 특징 행렬(NSFM, non-negative semantic feature matrix)과 비음수 의미변수 행렬(NSVM, non-negative semantic variable matrix)로 분해된다. 군집 내의 문장 벡터들은 의미특징 벡터에 가중치인 의미변수를 곱한 값의 선형합으로 표시된다. 의미특징 벡터는 문장의 내부특징을 나타내며, 의미변수는 문장 내에서 의미특징의 중요도를 나타낸다. 주제와의 유사도값이 제일 큰 의미특징을 선택하고, 이 의미특징에 대한 가중치 값이 제일 큰 문장을 추출한다. 추출된 문장들을 사용자가 읽기에 적합한 순서대로 정렬한다.

제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 의미특징벡터와 의미변수벡터는 음수 값이 없으므로 문장들이 직관적으로 이해하기 쉬운 의미특징들의 조합으로 분석될 수 있다. 따라서 의미적으로 주제와 유사도가 높은 문장을 선택할 가능성을 높여준다는 장점이 있다. 둘째, 의미특징벡터와 의미변수벡터는 희소하기 때문에 문장이 좁은 범위의 의미를 나타내는 적은 개수의 의미특징들로 분해되므로 주제와 가까운 문장을 쉽게 찾을 수 있다는 장점이 있다. 셋째, 주제와 의미 특징간의 유사도에 가중치를 부여하여 주제와 실질적으로 유사한 의미 특징을 선택하므로 주제와의 유사도는 높으나 실제로는 문서에서 별로 중요하지 않는 문장이 추출되는 것을 피할 수 있다. 넷째, 문장을 군집함으로써 중복 정보를 적절히 처리할 수 있고 노이즈를 제거할 수 있기 때문에 요약 문장의 추출시 편향된 문서의 고유구조의 반영을 피할 수 있고 문서요약의 정확도를 높일 수 있다. 다섯째, 추출문장을 사용자의 필요에 맞게 정렬하여 일관성을 높일 수 있다.

본 논문의 구성은 다음과 같다. 제2장에서는 다중문서 요약에 관한 관련연구를, 제3장에서는 제안한 다중문서 요약방법에 대하여 기술한다. 제4장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제5장에서 결론은 맺는다.

2. 관련연구

문서요약은 접근방법에 따라서 언어학적 접근방법과 통계적 접근방법으로 나눌 수 있다. 언어학적 접근방법은 언어적 지식에 기반한 문서 내부의 관계를 이용하는 방법이다. 언어학적 접근 방법은 문서를 요약하기 까지 비교적 복잡한 언어처리 과정과 워드넷, 시소러스 등의 외부의 추가적인 언어 지식 자원을 필요로 한다. 통계적 방법은 용어 출현 빈도, 제목, 문장 길이, 문장의 위치, 암시 단어나 구 등의 특징을 사용하여 각 문장이나 문단의 중요도 값을 구하여 중요도 값이 높은 문장이나 문단을 추출하는 방법이다. 문장에 의미구분이 필요한 경우에 구분하지 못하는 점과 실제로는 중요한 문장이나 통계값이 낮아서 구분하지 못하는 문제를 가지고 있다. 이러한 문제를 해결하기 위해서 MMR, LSA, 군집 방법들이 이용되고 있다[1].

다음은 언어학적 접근방법에 속하는 문서요약 방법들이다. Harabagiu와 Lacatusu는 주제 문장에 기반한 문장추출 방법과 문장 정렬방법에 의한 다중문서요약 방법을 제안하였다[9]. 이 방법은 자연어처리를 이용하여 주제를 추출하고 문서를 요약한다. 그러나 많은 주제 검색 과정을 요구하기 때문에 계산 비용이 많이 든다. Sakurai와 Utsumi는 시소러스를 사용한 질의 기반 방법을 제안하였다. 이들이 제안한 방법은 먼저 시소러스를 사용하여 질의와 가장 관련이 있는 핵심 문장을 추출하고, 나머지 문서들로부터 요약을 보충할 부분을 추출하여 문서를 요약하였다. 이들의 방법은 긴 문서를 요약 할 때에는 효과적이나 요약 문장이 짧을 때에는 좋은 성능을 보장하지 못한다[10].

다음은 통계학적 접근방법에 속하는 문서요약 방법들이다. Goldstein외 저자들은 MMR(maximal marginal relevance)를 적용한 방법을 제안하였다. 이 방법은 문서와 질의의 유사도, 선택 단락과 이전 선택 단락의 유사도 등을 계산하여 문서를 요약한다. 이 방법은 단순히 통계적 처리만을 사용하기 때문에 복잡한 자연어처리방법이나 정보추출 기법에 비하여 잘못 요약하는 경우도

발생한다[11]. Nomoto와 Matsumoto는 문서의 다양성 개념을 이용한 문서요약 방법을 제안하였다. 이들의 방법은 변형된 K-means 군집방법을 이용하여 문서에서 다양한 주제 영역들을 찾는다. 그리고 각각의 주제 영역에서 문장들의 중복성을 제거하여 가장 중용한 문장을 선택한다[12]. LSA는 고차원의 자료 공간의 축을 변경하여 원본 자료를 가장 잘 대표 할 수 있는 새로운 축을 찾아 원본 자료의 잠재구조를 이용하는 방법이다[13,14]. Gong과 Liu는 연관 척도(Relevance Measure)와 LSA를 이용한 두 가지 문서요약 방법을 제안하였다. 첫 번째 방법은 연관척도를 이용하는 방법으로, 문서를 문장 후보 집합으로 분해하고, 후보문장 집합에 연관척도를 이용하여 상위 점수의 문장을 추출한다. 두 번째 방법은 LSA를 이용한 방법으로, 문장집합에 SVD (singular value decomposition)을 하여 고유값 행렬로 분해한다. 분해된 좌 고유값 행 벡터에서 가장 큰 요소값과 일치하는 문장을 추출하여 문서를 요약한다[2]. Hachey외 저자들은 질의 지향의 다중문서요약을 위하여 MMR과 LSA(Latent Semantic Analysis)를 이용한 방법을 제안하였다[14]. 이 방법은 표현 오차를 최소화하도록 의미특징이 추출되므로 의미특징 벡터의 용어 가중치가 음수 값을 가질 수 있다. 따라서 의미특징의 직관적인 의미가 불분명하여, 직관적으로 의미를 파악하기 어렵도록 요약 문장이 추출될 수 있다는 문제가 있다[15]. Park외 저자들은 NMF를 이용한 군집기반의 다중문서요약 방법을 제안하였다. 이 제안 방법은 문장을 군집하고, 주제와 의미특징 간의 유사도를 이용하여 다중문서를 요약하였다. 이 방법은 문장을 군집하여 노이즈를 제거함으로써 주제가 문장구조에 편향되게 반영하는 것을 피하여 요약의 질을 높였으나, 주제와 의미특징 간의 유사도는 높으나 실제 문장에서는 중요하지 않는 문장을 추출 할 수 있으며, 요약문장에 대한 일관성을 고려하지 않았다. 또한 이러한 문제를 해결하기 위하여 가중유사도를 적용한 방법도 제안하였다[4,5].

3. 주제기반의 다중문서요약

본 장에서는 NMF를 기반으로 문장을 추출하여 다중문서요약을 할 수 있는 방법을 제안한다. 제안 방법은 전처리 단계와 문서군집단계, 문장 추출에 의한 문서요약 단계로 이루어진다. 다음 장에서 세 단계에 대하여 자세히 기술한다. 다음 그림 1은 NMF에 의한 다중문서요약 방법의 개요이다.

본 논문에서 행렬 X의 j번째 열벡터는 X_{*j}로, i번째 행벡터는 X_{i*}로, i번째 행과 j번째 열의 원소는 X_{ij} 표시한다. 표 1은 본 논문에서 사용된 기호와 기호의 의미이다.

3.1 전처리

전처리 단계는 주어진 문서를 각각의 문장으로 분리 후, 불용어(stopword) 제거, 어근추출(stemming), 가중치 계산으로 이루어진다. 이후 용어 빈도(term frequency) 벡터를 생성하고 식 (1)을 이용하여 가중치를 계산한다[8,13,16].

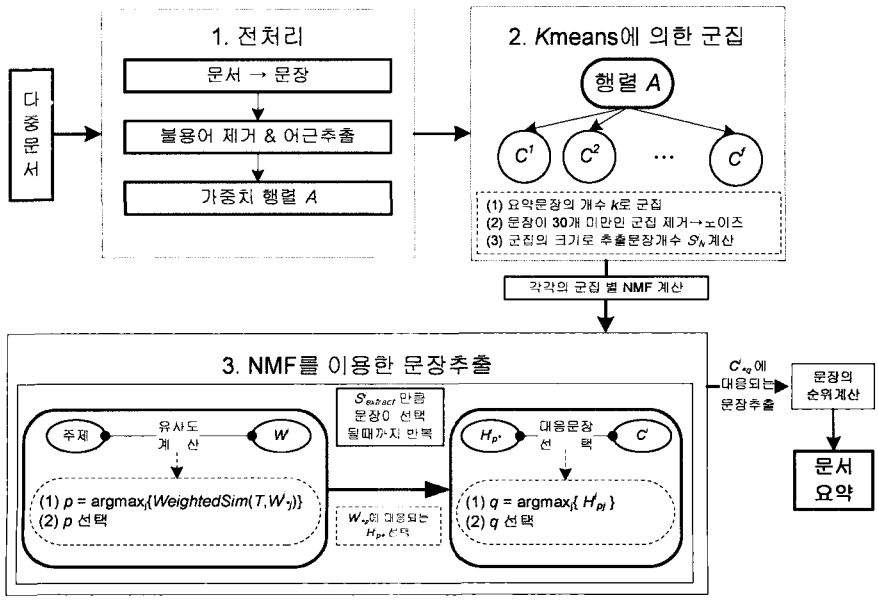

그림 1 NMF를 이용한 주제기반의 다중문서요약 방법의 개요

표 1 논문에 사용된 기호와 기호의 의미

기호	의미
A	용어 m과 문장 n으로 이루어진 $m \times n$ 가중치 행렬
m	용어의 개수
n	문장의 개수
r	의미특징의 개수
W	$m \times r$ 의미특징 행렬
H	$r \times n$ 의미변수 행렬
A_{ji}	i번째 문장에서 j번째 용어가 출현한 빈도의 가중치 원소
L_{ji}	i번째 문장에서 j번째 용어의 출현 빈도
$G(j)$	j번째 용어를 위한 전역 가중치
$n(j)$	j번째 용어를 포함한 문장의 개수
C^i	i번째 군집의 가중치 행렬
s_i	군집 C^i의 열의 개수
f	군집의 개수
f_{noise}	노이즈 군집의 개수
$f' = f - f_{noise}$	노이즈 군집을 제거한 군집의 개수
$d(\ )$	거리척도 함수
$sim(\ )$	코사인 유사도 함수
e_i	군집 C^i에서 추출할 문장의 개수
$TCsim(\ )$	주제와 군집간의 유사도 함수
r_i	군집 C^i에서 포함된 의미특징의 개수
k	사용자가 원하는 요약문의 개수
p	주제와 의미특징 열 벡터간의 유사도가 가장 큰 열의 위치
q	p번째 행의 가장 큰 의미 변수 원소의 위치
T	주제벡터로 $T = (T_1, \cdots, T_m)$
$WeightedSim(\ )$	주제와 의미벡터간의 가중치 유사도 함수

$$A_{ji} = L_{ji} \cdot G(j) \tag{1}$$

$$G(j) = \log(n/n(j)) \tag{2}$$

3.2 문장의 군집

본 논문에서는 Kmeans를 이용하여 문장을 군집한다. Kmeans은 n개의 객체를 주어진 K개의 군집으로 분할하는 알고리즘이다[17]. 문장을 군집하기 위해서 가중치 행렬 A에서 코사인유사도[18]에 기반한 식 (3)의 거리척도를 사용하여 Kmeans 알고리즘을 수행한다.

$$d(A_{*a}, A_{*b}) = 1 - sim(A_{*a}, A_{*b}) \tag{3}$$

$$sim(A_{*a}, A_{*b}) = \frac{A_{*a} \cdot A_{*b}}{|A_{*a}| \times |A_{*b}|} = \frac{\sum_{j=1}^{m} A_{ja} \times A_{jb}}{\sqrt{\sum_{j=1}^{m} A_{ja}^2} \times \sqrt{\sum_{j=1}^{m} A_{jb}^2}} \tag{4}$$

여기서, A_{*a}와 A_{*b}는 행렬 A의 a번째와 b번째 열벡터로서 a번째와 b번째 문장의 가중치 벡터이다. 또한 이것들은 비음수 값을 가지므로 $0 \le sim(\ ) \le 1$이고 따라서 $0 \le d(\ ) \le 1$이다.

문장 군집의 행렬 C^i는 A 행렬의 열 벡터들의 부분집합이고, 서로 disjoint하며 다음과 같은 성질을 만족한다.

$$\{A_{*j} \mid j = 1,\ldots,n\} = \bigcup_{i=1}^{f} \{C_{*l}^i \mid l = 1,\ldots,s_i\},$$

$$C^i \cap C^j \neq \phi, \quad i \neq j \tag{5}$$

3.3 의미특징과 의미변수를 이용한 문장 추출과 정렬

NMF의 정의는 다음과 같다. 주어진 비음수 행렬 A를 식 (6)과 같이 비음수 의미 특징 행렬(NSFM) W와 비음수 의미 변수 행렬(NSVM) H의 곱으로 분해하는 것이다.

$$A \approx WH \tag{6}$$

여기서 r은 일반적으로 m이나 n보다 작게 선택하여 행렬 W나 행렬 H의 크기가 행렬 A의 크기 보다 작게 한다.

본 논문에서는 $\tilde{A} = WH$가 근사값을 얻을 수 있도록, 각 A의 열 벡터의 Euclidean 거리가 최소가 되도록 하는 Lee와 Seung이 제안한 목표함수를 사용한다[6,7]. 식 (7)과 같은 Frobenius norm을 사용한 목표함수를 이용한다.

$$\Theta_E(W,H) \equiv \|A - WH\|_F^2 \equiv \sum_{j=1}^{m}\sum_{i=1}^{n}(X_{ji} - \sum_{l=1}^{r} W_{jl}H_{li})^2 \tag{7}$$

W와 H의 원소 값을 갱신하기 위하여 $\Theta_E(W,H)$ 값이 수렴 허용오차 보다 작게 되거나 지정한 반복횟수를 초과할 때까지 식 (8)과 (9)를 반복한다[6,7].

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T A)_{a\mu}}{(W^T W H)_{a\mu}} \tag{8}$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(AH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \tag{9}$$

예1) 다음은 식 (8)과 식 (9)를 이용하여 A 행렬을 W와 H 행렬로 분해한 예이다. r = 2, 수렴할 반복 수는 50이고, 수렴 허용오차가 0.001이다. W와 H 행렬의 초기값은 각각 0.5이다.

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

문장 추출 단계는 다음과 같다. 노이즈 군집을 제거한 f'개 군집의 문장행렬에 NMF 알고리즘을 적용하여 행렬 W^i와 행렬 H^i를 식 (10)과 같이 계산한다.

$$C^i \approx W^i H^i, \ i = 1, 2, \cdots, f', f' = f - f_{noise} \tag{10}$$

행렬 C^i의 j번째 열벡터 C_{*j}^i는 j번째 문장의 가중치 벡터를 나타내며, 행렬 W^i의 l번째 의미특징 열벡터 W_{*l}^i와 의미변수 H_{lj}^i의 선형조합으로 식 (11)과 같이 표현된다. 즉, l번째 의미 특징벡터 W_{*l}^i의 C_{*j}^i 문장벡터 내에

서의 가중치가 의미변수 H^i_{ij} 이다.

$$C^i_{*j} \approx \sum_{l=1}^{r} H^i_{il} W^i_{*l} \tag{11}$$

예2) 다음은 예제(1)의 열벡터 A_{*3}를 식 (11)과 같이 의미특징벡터와 의미변수의 선형조합으로 나타낸 예이다.

$$\begin{bmatrix}3\\6\\9\\12\end{bmatrix} \approx 7.2\times\begin{bmatrix}0.15\\0.66\\1.15\\1.61\end{bmatrix} + 1.2\times\begin{bmatrix}1.60\\0.97\\0.57\\0.41\end{bmatrix} = \begin{bmatrix}3.03\\5.93\\8.95\\12.08\end{bmatrix}$$

A_{*3} H_{13} W_{*1} H_{23} W_{*2} $\tilde{A}_{*3}$

식 (12)는 주제와 의미특징 벡터간의 실제 유사도를 나타내는 식이다. 주제와 의미특징벡터간의 단순 유사도 ($sim(\ )$)는 그 의미특징이 문장들에서 실제로 표현되고 있는 중요도를 반영하고 있지 않다. 따라서 의미특징이 문장들에서 실제로 표현되고 있는 중요도를 반영하기 위해서 전체 문장들에서 사용되는 중요도의 합의 비율을 곱한다. 식 (12)를 사용함으로써 주제와의 유사도는 높으나 전체 문서들에서 별로 중요하게 나타나지 않는 문장들이 선택되는 오류를 피할 수 있다.

$$WeightedSim(T, W^i_{*p}) = sim(T, W^i_{*p}) \times \frac{\sum_{v=1}^{s_i} H^i_{pv}}{\sum_{u=1}^{r_i}\sum_{v=1}^{s_i} H^i_{uv}} \tag{12}$$

식 (13)은 주제와 군집간의 유사도를 나타내는 식이다.

$$TCsim(T, C^i) = \sum_{p=1}^{r_i} WeightedSim(T, W^i_{*p}) \tag{13}$$

군집 C^i에서 추출하는 문장의 개수 e_i는 다음 식 (14)과 같이 정의한다. 군집과 주제의 유사도가 커질수록 많은 문장이 추출되고, 또한 군집에 포함된 문장의 개수가 많을수록 많은 문장이 추출된다.

$$e_i = \left\lfloor k \times \frac{s_i \times TCsim(T, C^i)}{\sum_{j=1}^{f'} \left(s_j \times TCsim(T, C^j)\right)} \right\rfloor \tag{14}$$

본 논문에서 제안한 NMF와 군집을 이용한 다중문서 요약 알고리즘은 다음과 같다.

1. 문서를 개개의 문장으로 분해한다. 요약할 문장의 개수를 k라 한다.
2. 문장집합을 전처리하여 가중치 행렬 A를 구성한다.
3. 행렬 A에서 f개의 군집들, C^i i=1, 2, …, f,을 구한다. 군집들에서 노이즈 군집을 제거한다. 식 (14)을 적용하여 각각의 군집에 대한 추출문장 개수 e_i를 계산한다. 여기서 i=1, 2, …, f'이고, $f' = f - f_{noise}$ 이다.
4. 행렬 C^i에서 식 (10)의 비음수 행렬 W^i, H^i를 구한다.
5. 각 군집 C^i에 대해서 아래의 단계를 수행한다.
   a. $p = \arg\max_{1\le j\le r_i} \{WeightedSim(T, W^i_{*j})\}$ 를 선택한다.
   b. $q = \arg\max_{1\le j\le s_i} \{H^i_{pj}\}$ 를 선택한다.
   c. C^i_{*q}에 대응되는 문장을 요약문장집합에 넣는다.
   d. e_i 개수만큼의 문장이 선택할 때까지 a에서 c단계를 반복한다. 여기서 a, b를 반복 수행할 때에는 먼저 선택된 것을 제외하고 제일 큰 값을 선택한다.
6. 식 (12)의 유사도가 큰 순으로 추출문장을 최종요약문으로 정렬한다.

위 5.a단계에서 식 (12)를 이용하여 주제 T와 가장 유사한 의미특징벡터 W^i_{*p}를 선택한다. 5.b단계에서 p번째 의미특징벡터를 가장 많이 반영하고 있는 문장을 선택하기 위해서 의미변수벡터 H^i_{p*}의 p번째 행에서 가장 큰 값 H^i_{pq}을 가진 q열을 선택한다. 만일 두 개 이상의 의미특징벡터를 반영하여 q열을 선택하려면 5.a단계와 5.b단계를 다음과 같이 수정하면 된다. 주제와 두 번째로 유사한 p'번째의 의미특징벡터를 함께 고려하여 좀더 정확한 요약 문서를 선택하려면, 5.a에서 $p = \arg\max_{1\le j\le r_i} \{WeightedSim(T, W^i_{*j})\}$ 와 $p' = \arg\max_{1\le j\le r_i, j\ne p} \{WeightedSim(T, W^i_{*j})\}$ 를 선택하고, 5.b에서 $q = \arg\max_{1\le j\le s_i} \{H^i_{pj} \times WeightedSim(T, W^i_{*p}) + H^i_{p'j} \times WeightedSim(T, W^i_{*p'})\}$ 를 선택한다.

3.4 LSA와 NMF에 의한 요약 방법의 비교설명

LSA에 의한 문서요약 방법은 식 (15)의 고유값 분해(SVD, singular value decomposition)를 적용하여 문서를 요약한다. 이 방법은 행렬 A를 U, D, V^T등의 3개의 행렬로 분해한다.

$$A = UDV^T \tag{15}$$

여기서 U는 AA^T의 고유벡터(좌 고유벡터, left singular vector)들의 $m\times m$ 직교정규(orthonormal) 행렬이고, V는 A^TA의 고유벡터(우 고유벡터, right singular vector)들의 $n\times n$ 직교정규(orthonormal) 행렬이다. $D = diag(\sigma_1, \sigma_2, \ldots, \sigma_n)$는 대각원소(diagonal elements)가 비음수의 고유값이고 내림차순으로 정렬되어 있는 $n\times n$ 대각행렬이다.

LSA에서, 행렬 A의 i번째 열벡터 A_{*i}는 i번째 문장의 가중치 벡터를 나타내며, 의미 특징벡터인 좌 고유벡터 U_{*j}들의 선형조합으로 식 (16)과 같이 표현된다. 즉, j번째 의미특징벡터 U_{*j}의 A_{*i} 문장벡터 내에서의 가중치가 $\sigma_j V_{ij}$ 이다.

$$A_{*i} \approx \sum_{j=1}^{r} \sigma_j V_{ij} U_{*j} \tag{16}$$

V^T의 첫 번째 행에서부터, 가장 큰 값인 열에 대응되는 문장을 요약문장으로 선택한다[6].

다음 예는 LSA와 NMF를 문서에 적용한 예이다. 표 2는 "Tourism in Great Britain"란 주제에 관련된 20개의 문서를 문장으로 추출한 일부분을 나타낸 것이다. 표 2의 문장집합을 전처리하여 용어문장 행렬 A를 생성한다. A는 396개의 용어와 57개의 문장으로 구성된다.

표 3은 A행렬을 SVD로 분해하여 얻은 10개의 의미특징벡터 U_{*j}들과 문장 S20에 대한 가중치($\sigma_j V_{20j}$)를 보여 주고, 선택된 문장 S20을 원본 문장의 용어빈도로서 보여 주고, 각 의미특징 벡터와 가중치($\sigma_j V_{20j}$)를 곱한 값의 합으로서 나타낸 값을 보여 주고 있다. LSA방법에서 V의 첫 번째 열의 값들이 가장 큰 고유값과 곱해지므로 첫 번째 의미특징 벡터(U_{*1})가 가장 가중치가 높다는 것을 의미하고 있다. 따라서 여러 문장 중에서 U_{*1}의 가중치를 가장 크게 가지고 있는 문장을 요약 문장으로 선택하는 것이다.

표 4는 A를 NMF로 분해하여 얻은 10개의 의미특징 벡터 W_{*j}와 문장 S20의 가중치(H_{j20})를 보여 주고, 선택된 문장 S20을 원본 문장의 용어빈도로서 일부분을 보여 주고, 각 의미특징 벡터와 가중치(H_{j20})를 곱한 값의 합으로서 나타낸 값을 보여 주고 있다.

표 3과 표 4를 비교하여 보면 표 3에서는 의미특징 벡터의 값에 음수가 많이 나타나 있으며 0이 아닌 값이 그리 많지 않은데 비해 표 4에서는 의미특징 벡터의 값에 음수 값이 없고 0의 값이 많다는 것을 알 수 있다. 즉 LSA 방법에 비해서 NMF방법으로 구한 의미특징 벡터는 좀더 희소하게 하여 좁은 범의의 의미를 나타내고 있고 음수 값이 없으므로 직관적으로 이해하기 쉬운 형태로 표현 되어 있음을 알 수 있다. 또한 문장에 대한 각 의미특징 벡터의 가중치 값도 LSA방법에 비해서 NMF 방법으로 구한 값이 더 희소하다는 사실도 알 수 있다. 이것은 LSA방법이 문장을 비직관적인 많은 의미특징벡터들의 선형 조합으로 표시하는데 비해서 NMF방법은

표 2 "Tourism in Great Britain"란 주제와 관련된 문장집합의 일부분

문장번호	내용
S1	TOURIST arrivals to the UK in 1991 are forecast to recover sharply after the steep decline earlier this year caused by the Gulf war. The British Tourist Authority said incoming tourist numbers had already increased significantly after falling 18 per cent in the first two months of this year from the levels of the corresponding period of 1990.
...	...
S20	The increases were achieved in spite of a fall in the number of North American visitors Visits by North Americans fell 6 per cent to 600,000 in the first quarter However, the number of visitors from western Europe rose 12 per cent to 23m - higher than in any previous first quarter. A RECORD 185m tourists visited Britain in the 12 months to March, 8 percent more than the previous year - and the British Tourist Authority said yesterday that it was expecting even higher numbers this year.
...	...
S57	She said wealthier Americans appeared to have returned this summer, along with an increasing number of continental Europeans and visitors from as far away as Venezuela. increasing number of continental Europeans and visitors from as far away as Venezuela.

표 3 LSA를 이용한 문장의 표현

Term		의미특징										문장 S20	
		U_{*1}	U_{*2}	U_{*3}	U_{*4}	U_{*5}	U_{*6}	U_{*7}	U_{*8}	U_{*9}	U_{*10}	원본	$\sum_{j=1}^{10}\sigma_j V_{20j} U_{*j}$
...	...	...	...	...	...	...	...	...	...	...	...	...	...
13	war	-0.0438	0.029	-0.004	0.011	0.091	-0.094	0.031	0.053	-0.014	0.120	0	0.000
14	british	-0.113	-0.016	-0.050	0.087	-0.057	0.031	-0.133	0.107	0.179	-0.014	1	1.000
15	author	-0.168	-0.004	-0.169	0.274	-0.071	0.055	0.051	-0.014	0.025	-0.229	1	1.000
16	income	-0.023	-0.005	0.015	0.007	-0.021	-0.053	-0.056	-0.074	-0.016	0.016	0	0.000
17	increase	-0.112	-0.014	0.025	-0.045	0.115	0.058	0.136	0.066	-0.115	0.183	1	1.000
18	significantly	-0.014	0.021	-0.011	0.012	-0.007	-0.001	-0.002	0.014	0.017	0.017	0	0.000
19	fall	-0.062	-0.051	-0.042	-0.036	0.067	0.127	-0.050	0.063	-0.004	-0.043	1	1.000
20	cent	-0.434	-0.144	-0.157	-0.369	-0.126	0.184	-0.270	-0.057	-0.075	-0.033	3	3.000
21	month	-0.105	0.053	0.002	-0.033	-0.016	0.156	-0.059	0.057	0.116	0.042	1	1.000
22	level	-0.046	0.053	-0.037	0.041	-0.018	-0.013	-0.004	0.025	-0.019	0.027	0	0.000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
396	Return	-0.007	-0.004	0.007	0.007	0.036	-0.031	0.005	0.016	0.007	0.032	0	0.000
가중치σjV20j		-2.963	1.766	-0.040	0.222	0.018	-0.838	1.070	-0.562	1.864	1.675		

표 4 NMF를 이용한 문장의 표현

Term		의미특징										문장 S20	
		W_{*1}	W_{*2}	W_{*3}	W_{*4}	W_{*5}	W_{*6}	W_{*7}	W_{*8}	W_{*9}	W_{*10}	원본	$\sum_{i=1}^{10} H_{i20} W_{*i}$
...	...	...	...	...	...	...	...	...	...	...	...	...	...
13	war	0.000	0.000	0.000	0.000	1.116	0.455	0.419	0.168	0.000	0.039	0	0.112
14	british	0.000	0.679	0.439	2.113	0.296	0.000	0.000	0.000	0.406	0.128	1	0.895
15	author	0.000	0.600	0.000	2.173	0.000	1.099	0.000	0.206	2.769	0.000	1	1.051
16	income	0.000	0.000	0.351	0.274	0.307	0.000	0.000	0.000	0.000	0.031	0	0.111
17	increase	0.068	0.000	0.000	0.264	0.675	0.770	1.260	1.384	0.000	0.759	1	1.011
18	significantly	0.000	0.000	0.000	0.315	0.330	0.000	0.000	0.000	0.000	0.000	0	0.126
19	fall	0.000	0.6221	0.000	0.557	0.000	0.000	0.000	1.073	0.000	0.198	1	0.964
20	cent	1.610	4.514	3.339	1.766	0.028	1.569	0.000	3.002	0.000	0.096	3	2.984
21	month	0.463	0.534	0.000	2.103	0.314	0.000	0.000	0.107	0.000	0.210	1	0.950
22	level	0.000	0.280	0.000	0.000	0.670	0.524	0.206	0.000	0.000	0.000	0	0.022
...	...	...	...	...	...	...	...	...	...	...	...	...	...
396	Return	0.000	0.000	0.000	0.000	0.000	0.000	0.089	0.104	0.000	0.081	0	0.068
가중치 H_{j20}		0.000	0.068	0.002	0.401	0.000	0.005	0.001	0.651	0.000	0.000		

문장을 더 적은 수의 보다 직관적인 의미특징 벡터들의 선형 조합으로 표현한다는 것을 의미한다.

LSA방법은 가장 가중치가 높은 의미특징 벡터에 대응되는 V의 첫 번째 행에서 요약 문을 선택하고, 두 번째로 중요한 의미특징 벡터에 대응되는 V의 두 번째 행에서 요약문을 선택하는 방법을 선택하는데 반해서, NMF방법은 의미특징 벡터와 주어진 주제와의 가중치가 가장 높은 것을 먼저 선택하여 이에 대응되는 H의 행에서 해당 의미특징 벡터에 대한 가중치 값을 가장 크게 가진 문장을 선택한다. 따라서 LSA 방법에 비해서 NMF방법이 좀더 주제와도 가깝고, 의미적으로도 이해되기에 쉬운 문장들을 선택할 수 있는 가능성이 더 높다고 말할 수 있다.

4. 성능평가

제안방법에 대한 실험은 DUC[1]에서 평가방법으로 사용되고 있는 ROUGE[2](Recall-Oriented Understudy for Gisting Evaluation)를 이용하였다[18]. DUC은 전문가들이 작성한 이상적인 요약문과 제안된 시스템이 만든 요약문을 비교하여 각 시스템의 성능을 평가하는 국제회의다. 본 논문에서 제안한 방법의 성능을 평가하기 위하여 DUC2005의 평가자료로 실험하였다. DUC 2005의 평가자료는 50개의 주제와 주제에 관련된 25-50개의 관련된 문서로 구성되어 있다[19]. 다음 표 5는 DUC2005의 평가자료에 대한 특성을 나타낸 것이다.

본 논문에서는 DUC2005의 주어진 50개의 주제를 질의로 하는 주제기반의 다중문서요약에 대하여 실험을 하였다. ROUGE는 ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, ROUGE-SU 등의 자동평가 방법을 포함하고 있다[18]. 각 방법은 전문가가 만든 참조 요약문과 제안 시스템이 만든 후보 요약문 사이의 재현율, 정확율, f-measure를 측정한다. ROUGE-N은 식 (17)의 n-gram 재현율을 이용하여 재현율, 정확율, f-measure를 측정하는 방법이다.

표 5 DUC2005 평가자료에 대한 특성

문서집합의 속성	DUC2005
주제/군집의 수	50
총 문서의 수	1591
주제별 평균문서의 수	32
문서당 평균 문장	23
문서당 최소문장의 수	3
문서당 최대문장의 수	96

$$ROUGE-N = \frac{\sum_{S \in \{Reference Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference Summaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (17)$$

여기서, $gram_n$은 n-gram의 길이이고, $Count_{match}(gram_n)$은 참조요약문과 자동요약문이 동시에 발생한 최대 n-gram의 수이다. ROUGE-L은 두 문장에서 공통적으로 나타나는 가장 긴 단어를 순서에 관계없이 사용하여 재현율, 정확율, f-measure를 측정하며 식 (18)과 같다.

$$R_{lcs} = \frac{LCS(X,Y)}{m}, P_{lcs} = \frac{LCS(X,Y)}{n}, F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \quad (18)$$

여기서, m은 참조 요약문 X의 길이, n은 자동 요약문 Y의 길이 이고, $LCS(X,Y)$는 X와 Y간의 공통적으로나 나타나는 가장 긴 단어의 길이이다. R은 재현율이며 P는 정확율이고, β는 P_{lcs}/R_{lcs}이다. ROUGE-W는 두 문장에서 공통적으로 나타나는 가장 긴 단어의 개수

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

표 6 제안방법과 비교실험 결과

ROUGE	Thesaurus			LSA			*K*means			Clustering+NMF			제안방법		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
1-gram	0.256	0.287	0.313	0.409	0.218	0.283	0.380	0.199	0.273	0.415	0.295	0.334	0.445	0.348	0.348
L	0.244	0.262	0.298	0.388	0.207	0.268	0.368	0.197	0.258	0.394	0.280	0.317	0.421	0.340	0.340
W	0.071	0.105	0.103	0.109	0.108	0.108	0.102	0.098	0.099	0.112	0.156	0.124	0.120	0.157	0.137
SU	0.084	0.114	0.091	0.127	0.067	0.087	0.117	0.062	0.087	0.139	0.107	0.111	0.148	0.122	0.122

에 대해 연속적인 일치 여부를 고려하여 측정하도록 가중치를 부여하여 재현율, 정확율, *f*-measure를 측정하며 식 (19)과 같다.

$$R_{skip2}=\frac{SKIP2(X,Y)}{C(m,2)}, P_{skip2}=\frac{SKIP2(X,Y)}{C(n,2)},$$

$$F_{skip2}=\frac{(1+\beta^2)R_{skip2}P_{skip2}}{R_{skip2}+\beta^2P_{skip2}} \quad (19)$$

여기서 *SKIP2(X,Y)*는 *X*와 *Y*사이에 일치하는 *skip-bigram*의 수이고, β는 P_{skip2}와 R_{skip2}의 중요 계수이고 *C*는 조합함수이다. ROUGE-S는 단어쌍 사이에 임의의 공간을 허용하여 재현율, 정확율, *f*-measure를 측정한다. ROUGE-SU는 ROUGE-S에 *uni-gram*을 추가하여 재현율, 정확율, *f*-measure를 측정한다.

제안방법을 두 가지 방법으로 비교하여 성능을 평가하였다. 첫 번째 실험은 Thesaurus, LSA, *K*means, Clustering+NMF,등에 의한 요약방법들과 비교실험 하였다. 두 번째 실험은 DUC2005에 참가한 32개 요약시스템과 비교하는 실험을 하였다. 두 실험에서 DUC2005의 시험데이타와 ROUGE 평가방법이 사용되었다. 실험에 사용된 각각의 비교 방법들은 실험환경에 맞도록 재구현 하였다.

실험1) 첫 번째 실험은 Thesaurus, LSA[13], *K*means, Clustering+NMF[4], 제안방법을 비교평가 하였다. Thesaurus[10]는 Sakurai와 Utsumi가 제안한 방법으로, 실험환경에 맞도록 Moby thesaurus II[3]를 이용하여 재구현하였다. *K*means 방법은 K-means를 이용하여 문장을 군집하고, 각각의 군집들로부터 주제와의 유사도가 가장 높은 문장을 추출하는 방법이다. 표 6은 비교실험 결과의 재현율, 정확율, *f*-measure 값이다. 그림 2, 3, 4는 재현율, 정확율, *f*-measure에 대한 실험결과를 비교하여 그래프로 나타낸 것이다.

그림 2는 비교방법들의 재현율 평가 결과를 나타낸다. 실험결과 제안방법이 가장 좋은 성능을 보인다. 그 다음으로 Clustering+NMF, LSA, *K*means 순으로 성능이 우수하다. Thesaurus의 성능이 최저이다.

그림 3과 그림 4는 정확율과 *f*-measure 평가결과를 나타낸다. 실험결과 제안방법이 가장 좋은 성능을 보인다. 그 다음으로 Clustering+NMF, Thesaurus, LSA 순으로 성능이 우수하다. *K*means의 성능이 최저이다.

실험결과를 분석해 보면, 제안방법이 가장 좋은 성능을 보이며, *K*means의 성능이 최저이다. *K*means를 이

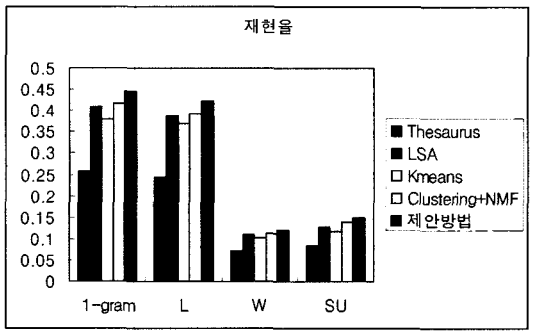

그림 2 제안방법과 비교방법 간의 재현율 실험결과

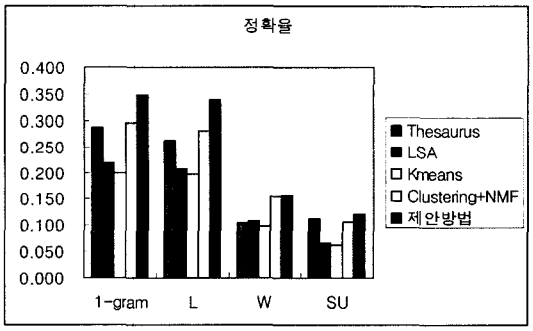

그림 3 제안방법과 비교방법 간의 정확율 실험결과

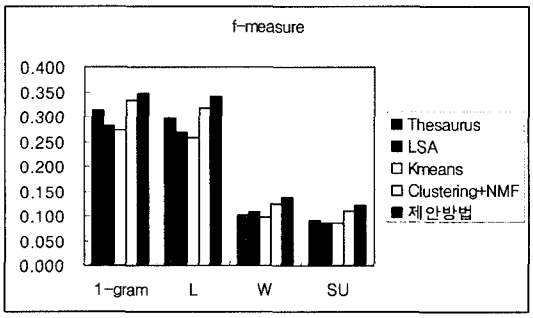

그림 4 제안방법과 비교방법 간의 F-measure 실험결과

3) http://en.wikipedia.org/wiki/Moby_Thesaurus#Thesaurus

표 7 DUC 2005 참가 요약시스템과 비교실험

ROUGE	제안방법	최고시스템	최저시스템	최고전문가	최저전문가
1-gram	0.3376	0.3797	0.2415	0.4585	0.4150
L	0.3196	0.3471	0.2048	0.4178	0.3934
W	0.1268	0.1297	0.0739	0.1589	0.1433
SU	0.1116	0.1343	0.0754	0.1756	0.1451

용하는 방법보다 LSA를 이용한 방법이 좀더 좋은 성능을 보인다. *K*means를 이용하여 노이즈를 제거하여 문서를 요약하는 것보다는 LSA를 이용하여 문서의 잠재 구조를 이용하는 것이 좀더 좋은 요약결과를 생성하는 것으로 보인다. 특히 용어의 thesaurus를 이용하여 문서를 요약하는 방법이 *K*means나 LSA에 비하여 더 좋은 결과를 나타냈다. 이는 용어들 간의 의미적 관계가 요약결과에 많은 영향을 미치는 것으로 보이다. Thesaurus 방법 보다 Clustering+NMF방법에서와 같이 *K*means를 이용하여 노이즈와 중복문장을 제거 한 후에 NMF를 이용하여 문서의 고유 구조를 반영하여 문서를 요약하는 것이 더 의미 있는 요약결과를 생성한다. 제안된 방법은 Clustering+NMF방법의 장점을 가지면서, 유사도에 가중치를 부여하여 유사도는 높으나 실제로는 중요하지 않는 문장이 추출되는 것을 방지하며, 추출문장을 정렬하여 최종문서로 요약함으로써 일관성을 높여서 요약의 질을 향상시켰기 때문에 가장 좋은 성능을 보인다.

실험2) 두 번째 실험은 DUC2005에 참여한 32개의 요약시스템들과 성능을 비교하였다. 표 7은 참가한 요약시스템들과 비교한 결과의 *f*-measure값이다. 여기서, 최고시스템은 참가한 요약시스템 중 ROUGE의 *f*-measure 값이 최고인 요약시스템이며, 최저 시스템은 ROUGE의 *f*-measure 값이 최저인 요약시스템이다. 최고전문가는 수작업으로 문서를 요약한 결과에 대한 ROUGE의 *f*-measure 값이 최고인 값을, 최저전문가는 ROUGE의 *f*-measure 값이 최소인 값을 나타낸다.

표 7에서 보는 것과 같이 사람이 직접 문서를 요약하는 것은 최저전문가라도 최고시스템을 사용하는 것보다 더 좋은 성능을 나타낸다. 전문가에 의한 요약문은 문서를 작성한 저자들의 다양한 표현방식, 문법 등을 고려하여 요약문을 생성한다. 최고시스템은 자연어처리를 통하여 어느 정도 저자의 표현방식을 고려하나 완벽하지는 않다. 제안방법은 ROUGE-W에서 최고시스템 대비 98%의 *f*-measure 값을 보인다. 제안방법은 문서내의 고유구조 상에서 일부 중요한 문장을 추출함으로써 문서를 작성한 저자의 의도를 충분히 반영하지 못하기 때문에 이러한 결과를 갖는 것으로 보인다.

5. 결 론

본 논문은 군집방법과 가중치가 부여된 의미 특징을 이용하는 새로운 주제기반의 다중 문서 요약 방법을 제안하였다. 주제기반의 다중문서요약은 모든 문서에 공통적인 정보를 포함하면서 주제에 관련된 정보를 포함하도록 중복성 제거와 차이점 식별과, 일관성을 보장하는 것이 중요하다. 제안된 방법은 문서의 고유특징과 주제간의 유사도를 이용하여 중요 주제가 요약결과에 직접 반영되고, 주제와 의미특징간의 유사도에는 가중치를 주어 유사도는 높으나 실제로는 의미 없는 문장이 추출되는 것을 피하여 차이점을 식별하였다. 또한 군집된 문장들로부터 중복정보와 노이즈를 제거함으로써 문서요약의 질이 높였다. 추출문장을 정렬하여 일관성을 높였다. 실험결과 대표적인 언어학적 접근방법과 통계학적 접근방법에 비교하여 제안방법의 성능이 더 좋은 것으로 나타났으며, DUC 2005에 참가한 시스템과 비교 결과 상위 성능의 시스템들과 비슷한 성능을 보였다.

앞으로의 연구는 다음과 같다. 제안된 방법에서는 문서를 작성한 저자의 표현방식이나 문법을 고려하지 않았다. 이 문제를 해결하기 위하여 주제나 용어에 가중치를 재부여 함으로써 문법이나 표현방식 등을 문서의 고유특징에 반영하면 더 정확한 결과가 나올 것으로 예상한다.

참 고 문 헌

[1] Mani, I., "Automatic Summarization," John Benjamins Publishing Company, 2001.

[2] Radev, D. R., Hovy, E. and Mckeown, K., "Introduction to the Special Issue on Summarization," Computational Linguistics, volume 28, 399-408, 2002.

[3] Goldstein. J., Mittal. V., Carbonell. J., Kantrowitz. M., "Multi-Document Summarization By Sentence Extraction," The Proceeding of the ANLP/NAACL Workshop, 2000.

[4] Park, S., Lee, J. H., Kim, D. W., Ahn, C. M., "Multi-document Summarization Based on Cluster Using Non-negative Matrix Factorization," In proceeding of SOFSEM, 2007.

[5] Park, S., Lee, J. H., Kim, D. W., Ahn, C. M., "Multi-document Summarization Using Weighted Similarity Between Topic and Clustering-Based Non-negative Semantic Feature," In proceeding of APWeb, 2007.

[6] Lee, D. D., Seung, H. S., "Learning the parts of

objects by non-negative matrix factorization," Nature 401:788-791, 1999.
[7] Lee, D. D., Seung, H. S., "Algorithms for non-negative matrix factorization," In Advances in Neural Information Processing Systems, volume 13:556-562, 2000.
[8] Ricardo, B. Y., Berthier, R. N., "Moden Information Retrieval," ACM Press, 1999.
[9] Harabagiu, S. Finley L., "Topic Themes for Multi-Document Summarization," In proceeding of ACM SIGIR, 202-209, 2005.
[10] Sakurai, T., Utsumi, A., "Query-based Multidocument Summarization for Information Retrieval," The Proceeding of NTCIR, 2004.
[11] Goldstein. J., Mittal. V., Carbonell. J., Callan. J., "Creating and Evaluating Multi-Document Sentence Extract Summaries," The Proceeding of CIKM, 165-172, 2000.
[12] Nomoto, T., Matsumoto, Y., "A New Approach to Unsupervised Text Summarization," In proceeding of ACM SIGIR, 26-34, 2001.
[13] Gong, Y., Liu, X., "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," In proceeding of ACM SIGIR, 19-25, 2001.
[14] Hachey. B., Murray. G., Reitter. D., "The Embra System at DUC 2005: Query-oriented Multi-document Summarization with a Very Large Latent Semantic Space," In Proceedings of the DUC, 2005.
[15] Xu, W., Liu X., Gong, Y., "Document Clustering Based On Non-negative Matrix Factorization," In proceeding of ACM SIGIR, 267-273, 2003.
[16] Chuang, W. T., Yang, J., "Extracting Sentence Segments for Text Summarization: A Machine Learning Approach," In Proceeding of ACM SIGIR, 152-159, 2000.
[17] Han. J., Kamber., M., "Data Mining Concepts and Techniques," Morgan Kaufmann, 2001.
[18] Chin-Yew, L., "ROUGE: A Package for Automatic Evaluation of Summaries," In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL, 2004.
[19] Hoa., H., D., "Overview of DUC 2005," In Proceedings of the DUC, 2005.

박 선
1996년 전주대학교 전자계산학과 졸업(학사). 2001년 한남대학교 정보산업대학원 정보통신학과 졸업(석사). 2007년 인하대학교 컴퓨터정보공학과 졸업(박사) 2008년~현재 호남대학교 컴퓨터공학과 전임강사. 관심분야는 정보검색. 데이타마이닝, 데이타베이스

이 주 홍
1983년 서울대학교 컴퓨터공학과 졸업(학사). 1985년 서울대학교 대학원 컴퓨터공학과 졸업(석사). 2001년 한국과학기술원 컴퓨터공학과 졸업(박사). 2001년~현재 인하대학교 컴퓨터공학부 부교수. 관심분야는 데이타마이닝, 데이타베이스, 정보검색, 신경망, 기계학습