다중요인모델에 기반한 텍스트 문서에서의 토픽 추출 및 의미 커널 구축

(Multiple Cause Model-based Topic Extraction and Semantic Kernel Construction from Text Documents)

장 정 호 † 장 병 탁 ††
(Jeong-Ho Chang) (Byoung-Tak Zhang)

요 약 문서 집합 내의 개념 또는 의미 관계의 자동 분석은 보다 효율적인 정보 획득과 단어 이상의 개념 수준에서의 문서간 비교를 가능케 한다. 본 논문에서는 다중요인모델에 기반하여 텍스트 문서로부터 토픽들을 추출하고 이로부터 의미 커널(semantic kernel)을 구축하여 문서간 유사도를 측정하는 방안을 제시한다. 텍스트 문서는 내재된 토픽들의 다양한 결합에 의해 생성된다고 가정하며 하나의 토픽은 공통 주제에 관련되거나 적어도 자주 같이 나타나는 단어들의 집합으로 정의한다. 다중요인모델은 은닉층을 갖는 하나의 네트워크 형태로 표현되며, 토픽을 표현하는 단어 집합은 은닉노드로부터의 가중치가 높은 단어들로 구성된다. 일반적으로 이러한 다중요인 네트워크에서의 학습과 추론과정을 용이하게 하기 위해서는 근사적 확률 추정 기법이 요구되는데, 본 논문에서는 헬름홀츠 머신에 의한 방법을 활용한다. TDT-2 문서 집합에 대한 실험에서 토픽별로 관련 있는 단어 집합들을 추출할 수 있었으며, 4개의 텍스트 집합에 대한 문서 검색 실험에서는 다중요인모델의 분석결과에 기반한 의미 커널을 사용함으로써 기본 벡터공간 모델에 비해 평균정확도 면에서 통계적으로 유의한 수준의 성능 향상을 얻을 수 있었다.

키워드 : 다중요인모델, 헬름홀츠 머신, 은닉의미자질, 의미 커널

Abstract Automatic analysis of concepts or semantic relations from text documents enables not only an efficient acquisition of relevant information, but also a comparison of documents in the concept level. We present a multiple cause model-based approach to text analysis, where latent topics are automatically extracted from document sets and similarity between documents is measured by semantic kernels constructed from the extracted topics. In our approach, a document is assumed to be generated by various combinations of underlying topics. A topic is defined by a set of words that are related to the same topic or cooccur frequently within a document. In a network representing a multiple-cause model, each topic is identified by a group of words having high connection weights from a latent node. In order to facilitate learning and inferences in multiple-cause models, some approximation methods are required and we utilize an approximation by Helmholtz machines. In an experiment on TDT-2 data set, we extract sets of meaningful words where each set contains some theme-specific terms. Using semantic kernels constructed from latent topics extracted by multiple cause models, we also achieve significant improvements over the basic vector space model in terms of retrieval effectiveness.

Key words : multiple-cause model, Helmholtz machine, latent semantic feature, semantic kernel

1. 서 론

인터넷의 발달과 이에 따른 정보량의 폭발적 증가로 온라인 텍스트나 전자화된 문서의 양이 크게 증대되고 있다. 하지만 이러한 데이타의 방대함이 곧바로 사용자들의 정보 획득 용이성으로 이어지는 것은 아니며, 오히려 정보 과부하(information overload)를 발생시켜 다양한 주제에 관련된 문서에 대한 검색과 조직화를 어렵게 하는 측면이 있다. 이에 따라 자동화된 텍스트 분석에 대한 요구가 증대되고 있으며, 이를 위해 문서 검색을

· 이 연구는 과학기술부 뇌신경정보학연구사업(BrainTech)과 국가지정연구실사업(NRL)에 의해 지원되었음
† 비 회 원 : 서울대학교 컴퓨터공학부
jhchang@bi.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
btzhang@cse.snu.ac.kr
논문접수 : 2003년 10월 22일
심사완료 : 2004년 3월 5일

비롯한 문서 분류, 문서 군집화 등에 기계학습이나 통계적 알고리즘 기반의 방법론들을 활용한 연구가 활발히 진행되고 있다.

이러한 연구의 일환으로, 최근 문서 집합으로부터의 개념 추출과 의미 관계 분석을 자동화하기 위한 연구들이 많은 관심을 받고 있는데, 대표적으로 LSA(latent semantic analysis)[1], NMF(non-negative matrix factorization)[2], spherical K-means[3] 등의 방법론들이 있다. 이러한 방법론들은 기본적으로 은닉변수 모델(latent variable models)의 관점에서 다시 생각해 볼 수 있는데, 은닉변수 모델에서는 관찰되는 많은 데이타들이 어떠한 (일반적으로 관찰변수의 수보다 상대적으로 적은) 숨겨진 요인들에 의해 생성된다고 가정한다. 은닉변수 모델은 입력 데이타에 내재된 특징 패턴을 이러한 은닉요인들에 의해 효과적으로 표현하는 것을 목적으로 하며, 텍스트 문서에 대해 적용할 때 텍스트 내에 내재된 의미 구조들을 발견하고 이를 요약적으로 제시하기 위한 도구로서 유용하게 적용될 수 있음이 실험적으로 확인되었다[1-4].

또한 은닉변수모델에 기반한 위의 방법론들은 정보검색 및 문서 분류 성능의 향상에 도움이 될 수 있다. 정보의 탐색 및 획득을 위한 핵심 내용 중의 하나는 사용자의 정보 요구에 대한 정보의 관련성(relevance)에 관한 것인데[5], 텍스트 문서의 경우 이는 문서의 표현 방식과 직결된 문제라고 할 수 있다. 기존의 기본적인 'bag-of-words' 방식은 간단하면서도 어느 정도 좋은 성능을 내지만, 단순히 단어 매칭에만 의존할 뿐 단어들 간의 의미관계를 고려하지 않는다는 점에서 문제가 있으며[6,7], 이는 유의어와 다의어에 관련된 문제로 볼 수 있다. 이러한 문제점에 대처하기 위해, 문서의 벡터공간 표현 시 단어간의 의미적 유사성에 관한 정보를 포함시킬 수 있으며, 이를 위한 방법으로는 Generalized VSM, LSI 등을 들 수 있다. Generalized VSM 기법은 문서 집합 내에서의 단어의 공기 정보를 이용하여 단어간 관련성 파악을 시도한다. LSI(latent semantic indexing) 기법은 이와 달리 차원 축소에 기반한 문서 인덱싱 방법으로서, 단어-문서 행렬의 선형적 차원 감소를 통해 단어나 문서들을 의미 공간(semantic space)으로 사상(mapping)하고 이로부터 단어나 문서들간의 의미 관계 파악을 시도하는 것으로서, 이를 통한 의미 관계 분석이 문서 검색 등에서 유용하게 적용될 수 있음이 실험적으로 확인되었다[8-10]. 그러나, LSI에서의 특이값 분해(singular value decomposition)에 의한 단어-문서 행렬의 순차적 재복구(reconstruction) 기반의 의미공간 구축 과정은 일반적으로 전역적(global) 자질을 추출하는 경향이 있으며, 각 차원의 토픽 관점에서의 해석 및 특정 토픽에 관련된 지역적(local) 자질의 추출이 곤란하다는 단점이 있다[3,4].

본 논문에서는 은닉변수모델 일종인 다중요인 모델(multiple-cause model)을 이용한, 텍스트 문서로부터의 토픽(topic) 추출 및 이를 활용한 의미 커널 구축을 통해 문서간 유사도를 측정하기 위한 방법을 제시한다. 토픽은 의미적으로 유사하거나 적어도 동일 주제에 관련된 단어들의 집합으로 정의되며, 이는 다중요인모델을 구성하는 은닉노드들에 의해 파악된다. 추출된 토픽을 통해 단어간의 의미적 상관성을 고려할 수 있으며 GVSM과 달리 LSI에서와 같이 축소된 공간상에서의 의미 표현을 시도한다. 또한 이러한 다중요인 모델에서는 하나 이상의 이산적(discrete) 토픽들의 조합으로서의 단어 및 문서 생성을 가정하고 또한 확률적 학습에 의해 모델을 추정함으로써, LSI에서와 같은 전역적 자질이 아닌 특정 개념 또는 토픽에 관련된 지역적 자질을 추출할 수 있을 뿐 아니라 k-means 군집화와 같은 단일요인 모델에 비해 각 토픽에 대해 보다 평활화(smooth)된 단어 분포를 얻을 수 있다.

2절에서는 다중요인 모델과 모델의 학습을 위한 헬름홀츠 머신(Helmholtz machine)에 대해 설명하며, 3절에서는 다중요인모델에 의한 텍스트 분석을 설명하고 학습된 모델로부터의 의미 커널 구축에 기반한 문서간 유사도 측정에 대해 서술한다. 4절에서는 실세계 문서 데이타에 대한 토픽 단어 분석 실험과, 다중요인모델로부터 구축된 의미 커널 기반의 질의-문서 유사도 측정을 통한 문서 검색 실험을 보인다. 마지막으로 5절에서는 결론 및 향후 연구방향을 제시한다.

2. 다중요인모델(Multiple-Cause Models)

다중요인모델은 은닉변수 모델의 일종으로서, 하나 이상의 내재된 은닉요인들의 조합에 의해 데이타가 생성된다고 가정한다[11,12]. 이 모델은 다중요인 네트워크(multiple-cause network)로 표현할 수 있는데, 여기서는 이를 텍스트 문서에 대한 분석 관점에서 설명한다.

2.1 다중요인 네트워크

그림 1은 다중요인 네트워크(multiple-cause network)의 전형적인 예이다. 네트워크 상의 각 은닉노드 z_k는 하나의 토픽을 의미하고 문서 내 각 단어는 해당 입력 노드에 대응된다. 문서집합 $D=\{d_1, d_2, \cdots, d_N\}$에서 각 문서 d_n이 서로 독립이고 동일한 분포를 따른다고 가정하면, 모델에서의 D에 대한 로그 확률 값은 식 (1)과 같다.

$$\log P(D|G) = \sum_{n=1}^{N} \log P(d_n|G) = \sum_{n=1}^{N} \log \left(\sum_{z} P(z|G) P(d_n|z, G) \right) \quad (1)$$

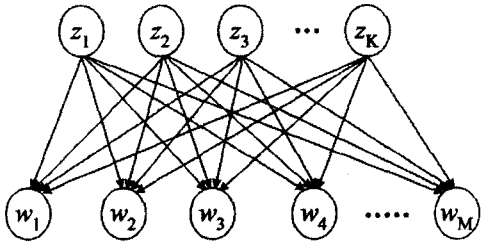


그림 1 다중요인 네트워크

G는 모델의 매개변수 집합이며, z는 토픽들의 가능한 조합 중의 하나를 의미하는 것으로서 $z=(1,0,1,\cdots,0)$과 같이 표현된다. 그림 1의 네트워크 구조와 의존성 분리(dependency separation)[12,13] 특성으로부터, 하나의 토픽 구성이 결정되면 문서 내 각 단어들은 서로 조건부 독립이라고 가정한다. 즉,

$$P(d_n|z, G) = \prod_{m=1}^{M} P(w_m|z, G) \quad (2)$$

만약 문서 내 단어의 출현 여부만 고려하여 문서를 이진벡터로 표현한다면, 각 변수 w_m은 이진 변수이다. $P(w_m|z)$은 일반적으로 시그모이드(sigmoid) 함수에 의해 계산된다[11].

$g_{km}(1\le m\le M,\ 1\le k\le K)$는 은닉노드 z_k로부터 입력노드 w_m으로의 연결선의 가중치이며, g_{0m}은 입력노드에 대한 바이어스(bias)이다.

$$P(w_m|z, G) = \frac{1}{1+\exp\left(-\sum_{k=0}^{K} g_{km} z_k\right)} \quad (3)$$

위와 같은 설정에서, 다중요인 모델은 학습과 추론시 은닉노드의 수가 증가함에 따라 각 데이타(문서)에 대해 가능한 토픽 조합(이진 벡터 z의 가능한 형태) 수가 지수적으로 증가하게 되며, 이에 따라 EM 알고리즘과 같은 기본적인 최대 우도(maximum likelihood) 접근법 적용이 어렵게 된다는 점이 문제가 된다. 예를 들어 토픽의 수가 20개(K=20)만 되어도, 각 데이타에 대해 고려해야 할 토픽 조합 수는 2^{20}(~1,000,000)개가 된다. 따라서 이러한 계산 복잡도 문제를 해결하기 위해 근사적 방법(approximation)이 일반적으로 많이 사용되는데, 본 논문에서는 헬름홀츠 머신(Helmholtz machine)에 의한 근사화 방식과 wake-sleep 알고리즘에 의한 모델 학습 방식을 활용한다.

2.2 헬름홀츠 머신 (Helmholtz Machine)

헬름홀츠 머신[11]은 다중요인모델이나 계층적 생성모델에서의 학습과 추론 과정을 용이하게 하기 위한 근사화 방법 중의 하나로서, 뉴런 형태의 확률적 처리 단위로 구성된 다층 연결망으로 표현된다. 일반적으로 다중요인 네트워크 상에서 데이타 d_n에 대한 사후 확률(posterior probability) 분포 $P(z|d, G)$에 대한 직접적 추정이 곤란할 경우, 보다 간단한 형태의 확률 분포 함수 $Q(z|d_n)$를 도입함으로써 이에 대한 근사적 추정을 시도할 수 있다. 헬름홀츠 머신에서는 다중요인 네트워크에서의 생성 네트워크에 인식 네트워크(recognition network)를 추가함으로써, 이러한 Q 함수를 제공한다. 그림 2는 그림 1의 다중요인 모델 학습을 위한 헬름홀츠 머신의 예를 보인다.

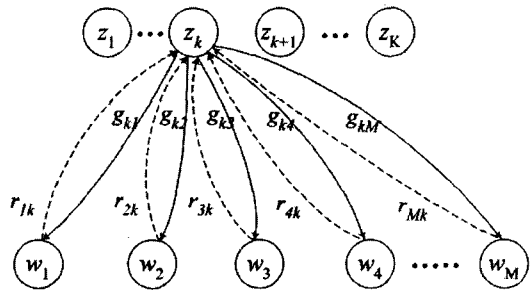


그림 2 은닉층이 하나인 헬름홀츠 머신. 실선은 은닉층으로부터의 생성모델(generative model)을 위한 연결선을 나타내고, 점선은 입력층으로부터의 인식모델(recognition model)을 위한 연결선이다.

인식네트워크는 입력층으로부터 은닉층으로의 상위 연결들에 의해 구현되며, $Q(z|d_n)$은 연결선의 가중치 집합 R을 매개변수로 하는 함수 $Q(z|d_n, R)$에 의해 추정된다. 그림 2와 같이 은닉노드들간의 측위(lateral) 연결이 존재하지 않으면, 인식 네트워크 상에서 은닉 노드 $z_k(1\le k\le K)$들은 입력 노드의 값이 주어질 때 서로 조건부 독립이고 이는 다음과 같이 표현된다.

$$Q(z|d_n, R) = \prod_{k=1}^{K} Q(z_k|d_n, R) \quad (4)$$

$Q(z_k|d_n, R)$는 생성모델에서와 같이 시그모이드 함수에 의해 계산되며,

$$Q(z_k|d_n, R) = \frac{1}{1+\exp\left(-\sum_{m=0}^{M} r_{mk} w_m\right)} \quad (5)$$

$r_{mk}(1\le m\le M,\ 1\le k\le K)$는 인식네트워크 상에서 입력노드 w_m으로부터 은닉노드 z_k로의 연결선의 가중치이며, r_{0k}는 은닉노드 z_k에 대한 바이어스(bias)이다. 이러한 인식 네트워크에 의한 근사적 방법을 통해, 헬름홀츠 머신은 사후(posterior) 확률 분포를 추정하는 절차를 간소화할 수 있는 방안을 제공한다.

3. 다중요인 모델에 기반한 텍스트 문서 분석

3.1 텍스트 분석을 위한 모델 학습 및 토픽단어 추출

2장에서 언급한 바와 같이 본 논문에서는 다중요인 네트워크에서의 학습을 위해 헬름홀츠 머신을 이용한다. 각 입력노드는 하나의 단어에 대응되고, 각 은닉노드는 문서에 내재된 은닉의미자질(latent semantic feature)을 표현하며 이는 해당 노드로부터의 연결 가중치가 높은 입력노드, 즉 단어들의 집합으로 정의한다.

그림 2와 같은 헬름홀츠 머신에서의 생성모델과 인식모델의 매개변수 G와 R의 추정을 위해서는 온라인 자기감독(self-supervised) 학습 알고리즘인 wake-sleep 알고리즘[14]을 이용한다. wake-sleep 알고리즘은 주어진 데이타 집합의 각 데이타에 대해 wake-단계와 sleep-단계의 두 단계를 번갈아가며 반복적으로 실행함으로써 모델 학습을 진행한다. Wake-단계에서는 인식모델을 이용하여 각 은닉노드들이 활성화될 확률을 구하고 이 확률에 따른 샘플링을 통하여 은닉노드들의 실제 활성화 여부를 결정한다. 그리고 각 잠재의미에 대한 샘플링결과와 실제 입력 데이타를 이용하여 생성모델의 매개변수 값을 갱신한다. Sleep-단계에서는 wake-단계와는 반대로, 앞서 수정된 생성모델의 매개변수를 이용하여 하나의 가상 데이타를 생성해 낸다. 이렇게 결정된 은닉노드와 입력노드의 값들을 이용하여 인식모델의 매개변수 값을 갱신한다. 각 단계에서의 생성, 인식모델의 매개변수들은 간단한 지역 델타 규칙(local delta rule)에 의하여 갱신된다[11,12]. 이러한 과정은 Generalized EM 알고리즘[15]과 연관지어 생각해 볼 수 있는데, wake-단계는 실제 생성모델의 매개변수를 학습하므로 Maximization 단계에, sleep-단계는 인식모델을 학습을 통해 주어진 데이타의 은닉층에서의 표현을 결정하므로, Expectation 단계에 근사하다고 할 수 있다[12]. 표 1은 텍스트 문서에 대한 모델 학습을 위한, wake-sleep 알고리즘 기반의 전반적인 학습 과정이다.

헬름홀츠 머신은 기본적으로 이진노드로 구성된 네트워크인데 문서 데이타 분석시 문서 내 단어의 출현 빈도를 고려함으로써 보다 효과적인 토픽 추정에 도움이 될 수 있다[16]. 본 논문에서는 인식모델에 의한 문서 d에 대한 자질 z_k의 샘플링 확률 $Q(z_k|d)$ 추정시 네트워크 상의 각 입력노드의 값을 해당 단어의 빈도수로 하며, 이는 표 1의 wake-단계에서만 적용된다.

3.2 문서간 유사도 측정을 위한 의미 커널

정보검색을 위한 벡터공간모델(vector space model)[17]에서 두 문서 d_1와 d_2의 유사도는 기본적으로 다음과 같이 표현할 수 있다[6].

$$sim(d_1, d_2) = (S^T d_1) \cdot (S^T d_2) = d_1^T S S^T d_2, \quad (6)$$

$$d_1 = (t_1(w_1), t_1(w_2), \ldots, t_1(w_M))^T$$

$$d_2 = (t_2(w_1), t_2(w_2), \ldots, t_2(w_M))^T$$

$t_i(w_m)$은 단어 $w_m (1 \le m \le M)$의 문서 d_i 내에서의 정보를 나타내는 것으로 보통 d_i 내에서의 w_m의 존재 유무나 빈도수를 나타낸다. 행렬 S는 기본 벡터 공간에서의 문서를 다른 자질공간으로 사상하기 위한 변환행렬이다. Cristianini[7]와 Siolas[18]는 이를 다시 커널(kernel) 관점에서 정의하였는데, d_i와 d_j가 주어질 때 특정 자질공간으로의 사상(mapping) ϕ에 대해 커널 함수는 $k(d_i, d_j) = \langle \phi(d_i), \phi(d_j) \rangle$로 표현된다. 식 (6)에서 변환행렬 S는 결국 ϕ를 정의하는 것으로 볼 수 있으며, 이 때 $\phi(d) = S^T d$로 주어진다.

표 1 wake-sleep 알고리즘에 기반한 텍스트 문서에 대한 학습

입력: (D, K)
- D: 단어-문서 행렬 (단어 수: N, 문서 수: M).
- K: 은닉의미자질(은닉노드)의 수.

출력: (G, R), G: 생성네트워크의 매개변수, R: 인식네트워크의 매개변수.

wake-sleep 알고리즘:

1. G, R 초기화
2. wake 단계:
   - 2-1. 문서 $d = \{w_1, w_2, \ldots, w_M\}$에 대해 인식네트워크 상에서 각 은닉의미자질(은닉노드)에 대해 확률 p_k를 계산하고(식 (5)), 샘플링을 통해 활성화 여부 결정.
2-2. 2-1의 결과에 기반하여 G 갱신
3. sleep 단계:
   - 3-1. 각 은닉의미자질(은닉노드)의 활성화 여부를 샘플링에 의해 결정한 후, 이로부터 생성네트워크 상의 각 단어(입력노드)의 활성화 확률 p_m을 계산하고(식 (3)), 샘플링을 통해 가상 문서 생성.
3-2. 3-1의 결과에 기반하여 R 갱신
4. 모든 문서에 대해 2, 3 과정을 반복 수행

변환행렬이 $S= I_M$(I_M은 $M\times M$ 단위행렬)일 때는 단어벡터 공간에서의 두 문서간의 내적에 의한 유사도이며, 역문서빈도를 해당 단어에 대한 가중치로 부여할 경우 변환행렬 S는 대각 행렬 $S_{IDF}=diag(idf(w_1), idf(w_2), idf(w_3),\cdots, idf(w_M))$로 표현된다. 이와 같이 S가 대각행렬인 경우($S= I_M$인 경우도 포함하여)에는 각 단어들에 의해 표현되는 축이 벡터공간상에서 직교(orthogonal)를 이루므로 단어들의 상관관계를 고려할 수 없다는 점에서 그 문제점이 지적되곤 한다[1,6,7,18].

기본 벡터공간모델(basic VSM)의 이러한 문제점을 어느 정도 해소하고자 하는 노력의 일환으로, 단어간의 상관관계 또는 의미 관계를 고려할 수 있는 방향으로 S를 정의할 수 있으며 이의 도입에 의한 문서간 유사도 또는 커널 함수 $k(d_i, d_j)$를 "의미 커널(semantic kernel)"이라 한다[7,18]. Wong[19]은 Generalized VSM (GVSM)을 제안하였는데, 이는 문서집합 내에 나타나는 단어들의 공기 정보를 이용하는 방식이다. N개의 문서로 구성된 문서 집합을 $M\times N$ 단어-문서 행렬 D로 표현할 때, GVSM에서의 변환행렬은 $S_{GVSM}= D$로 생각할 수 있다. LSI(latent semantic indexing)[1,10] 기법에서는 D에 대한 SVD(singular value decomposition)를 수행하고($D= U S V^T$) 상위 $k(\le rank(D))$개의 left singular vector들을 이용하여 문서간 유사도를 정의한다.

$$k_{LSI}(d_1, d_2)=(I_k U^T d_1)\cdot(I_k U^T d_2)= d_1^T U_k U_k^T d_2 \quad (7)$$

식 (6)과 식 (7)을 비교해 볼 때, LSI 적용에서의 변환행렬은 $S= U_k$로 정의된다. 말뭉치 기반이 아니고 외부 지식을 활용하는 경우도 있는데, Siolas[18]는 온라인 어휘 의미망인 워드넷(WordNet)[20] 상에서 단어들 간의 의미적 유사성을 계산하고 이로부터 단어-단어 행렬 W를 구성하여 변환행렬 $S_{WN}= W$로 사용하였다.

본 논문에서는 텍스트 문서에 대한 의미 커널을 구축하기 위하여 헬름홀츠 머신에 의한 학습 결과를 이용한다. 헬름홀츠 머신의 생성네트워크 상에서의 은닉노드에서 입력노드에 이르는 가중치의 집합 G를 활용하며, G는 $M\times K$(K는 은닉변수 개수) 행렬로 주어진다. 즉 $S_{MCM}= G$로 정의된다.

$$k_{MCM}(d_1, d_2)=(G^T d_1)\cdot(G^T d_2)= d_1^T GG^T d_2 \quad (8)$$

보통 $K\ll M$이며, 행렬 G의 각 열에서 가중치가 높은 단어들은 서로 그 연관도가 높거나 적어도 해당 토픽에 관련된 단어들로 간주된다.

4. 실 험

4.1 문서 집합으로부터 토픽 단어 추출

TDT-2 문서 집합에서 비교적 관련 문서가 많은 주제에 관한 10,685 문서를 선택하여 토픽 단어 추출에 관한 실험을 하였다. 기본적인 불용어(stop words)는 제거하고, 10개 이상의 문서에서 나타난 단어 중 문서와의 상호정보량이 큰 6,000개의 단어들을 선택하였으며, 스테밍 과정은 거치지 않았다. 전체 문서에 대한 단어 w의 상호 정보량 $I(w)$는 식 (9)와 같다[21].

$$I(w)= p(w)\sum_d p(d|w)\log\frac{p(d|w)}{p(d)} \quad (9)$$

위 식에서 $p(d|w)$와 $p(w)$는 다음과 같이 계산되며,

$$p(d|w)=\frac{n(d, w)}{\sum_d n(d, w)}, \qquad p(w)=\frac{\sum_d n(d, w)}{\sum_w\sum_d n(d, w)}$$

$n(d, w)$는 문서 d에서의 단어 w의 빈도수이다.

실험에서 헬름홀츠 머신의 각 입력노드는 개별 단어에 대응되어 총 6,000개의 입력노드를 가지며, 표 2는 은닉변수의 수를 64개로 설정한 학습 결과로 추출된 단

표 2 은닉변수가 64인 헬름홀츠 머신 학습 결과 추출된 토픽 단어들의 예

rupiah, bailout, traders, inflation, imf, stabilize, currencies, investor, restructuring, recession, banking, indonesian, devaluation, baht, slump
unscom, palaces, nerve, unrestricted, biological, tariq, compounds, invasion, scud, verify, aziz, iraqi, inspector, severest, anthrax
congressional, capitol, gingritch, newt, mccain, hearings, senators, trent, bipartisan, democrat, conservatives, legislation, republicans, rep, tennessee
intern, lewinsky, whitewater, impeachment, monica, paula, perjury, tripp, subpoenaed, starr, lindsey, clintons, hillary, ginsburg, recordings
palestinian, militants, terrorism, jerusalem, israelis, egypt, gaza, ramadan, holy, refugees, civilians, islamic, jewish, netanyahu
nagano, tokyo, ioc, winter, olympics, athletes, organizers, skater, hockey, snowboarding, sweden, norway, medalist, downhill, medals
pope, cuba, visit, fidel, havana, john, pontiff, communist, cubans, church, human, catholic, vatican, rome, roman
cigarrettes, farmers, producers, marketing, manufacturers, smokers, products, advertising, smoking, nicotine, industries, philip, customers, illnesses, taxes
attorneys, courts, testified, lawsuit, jury, prosecution, testify, lawyers, client, jurors, misconduct, attorney, courtroom, prosecutors, courthouse
viagra, impotence, pill, patients, pfizer, drugs, doctors, aids, medical, researchers, medicine, male, physical, treatment, dr

표 3 TDT-2 문서 집합 중 "미국에서의 담배 관련 법안 및 소송"에 관한 기사

> The next few months may well decide the financial and *legal* future of the *tobacco* industry. (중략)... In proposing the settlement, the tobacco companies are hoping to buy legal peace from a growing number of *lawsuits*. (중략).. the companies would receive protection against most *cigarette*-related *lawsuits* and would be shielded against punitive damages. (중략) it will not back *legislation* to restrict government authority over nicotine. *Tobacco farmers*, who were not included in the June proposal, are also lobbying for compensation under any *legislative* plan...

어 집합 예를 보인다. 각 행은 하나의 은닉변수 z에 대응되며, 선택된 단어들의 집합은 생성네트워크에서 가중치, 즉 g_{zw}의 값이 큰 상위 단어들이다. 각 단어 집합들을 검토해 보면 이러한 단어들이 각기 특정 토픽에 관련된 것을 알 수 있다. 예를 들어, 첫 번째 단어 집합은 주로 "아시아 경제 위기"에 관련된 것이며, 두 번째 단어 집합은 "이라크 무기 사찰"에 관한 내용이며, 여섯 번째는 "동계올림픽"에 관련된 내용을 표현하는 단어 집합으로 판단된다.

표 3은 TDT-2 문서 집합 중 "미국에서의 담배 관련 법안 및 소송"에 관련된 내용이다. 은닉 토픽 중 해당 문서에 대한 확률값이 높은 토픽들을 추출하였을 때, 표 2에서 3번째(미국 의회 관련), 8번째(담배 산업 관련), 9번째(사법 관련) 등이 포함되었다. 물론 해당 문서를 하나의 토픽으로 간주할 수도 있겠지만, 다중요인모델에서는 이와 같이 세부 주제의 조합으로 해당 문서를 표현할 수 있다는 것을 예시한다.

4.2 다중요인모델 기반 의미 커널을 이용한 문서 매칭

3.2 절에서 서술된 의미 커널을 활용하여 4개의 문서 집합(MED, CACM, CISI, CRAN)에 대한 문서 검색 실험을 하였으며 각 문서 집합의 구성은 표 4와 같다.

기본 VSM 이외에 GVSM, k-means 클러스터링, LSI, 다중요인모델을 이용한 데이타 기반의 의미 커널에 대해서 실험을 하고 그 성능을 비교하였다.[1] 기본 VSM 이외의 나머지 네 방법론에서의 질의문 q와 문서 d간의 유사도 $sim_S(q, d)$는 기본 VSM에서의 유사도 $k_{BVSM}(q, d)$와 각 방법론 기반의 의미 커널 $k_S(q, d)$에 의한 유사도의 평균을 취하였다.

$$sim_S(q, d) = \frac{1}{2} k_{BVSM}(q, d) + \frac{1}{2} k_S(q, d) \qquad (10)$$

각 질의어와 문서는 단어 빈도수(term frequency) 벡터로 표현되며, 5가지의 모든 방법론에 대해 단어의 역문서빈도를 고려한다. 표 5는 이러한 방법론들을 요약 제시한다.

다중요인모델, LSI의 경우에는 은닉 차원의 수를 32, 48, 64, 80, 96, 128로 변화시켜 나가면서 실험하였으며, k-means 클러스터링의 경우에도 유사한 방식으로 클러스터의 수가 32, 48, 64, 80, 96, 128인 경우에 대해 실험을 하였다. 그림 3은 각 문서 집합에 대한, 기본 VSM 기법과 의미 커널 정의를 위한 4가지 방법을 이용했을 때의 정확율-재현율(precision-recall) 그래프이며, 표 6은 각 기법의 평균정확도를 보인다. 다중요인모델, k-means 클러스터링, LSI의 경우에는 은닉변수 또는 클러스터 개수 설정에서 가장 높은 성능을 보이는 경우를 제시하였으며, 또한 k-means 클러스터링과 다중요인모델은 10번 수행한 후에 그 평균값과 표준편차를 명시하였다.

그림 3과 표 6에서 전체적으로 볼 때는 다중요인모델과 LSI 기반의 의미 커널에 의한 성능이 다른 방법론에 비해 조금 더 우수한 성능을 보였다. GVSM과 k-means 클러스터링의 경우 CACM과 CISI 데이타에 대해서는 기본 VSM보다 오히려 그 성능이 저하되었는데 이러한 결과로 볼 때, 단순한 단어 공기 정보 이용이 문서간 유사도 측정에서 반드시 도움이 되지는 않음을 알 수 있다.

각 문서 집합 별로 의미 커널에 의한 성능 향상 정도에 약간의 편차를 보이는데, 이러한 차이의 분석을 위해 각 문서 집합의 질의에 대한 관련 문서들의 특성을 살펴보았다. 각 질의문 q에 대한 최적의 토픽은 주어진 관련문서들의 집합 T로써 정의된다고 가정할 때, 만약 T가 q에 대한 비관련문서들과 적절히 구분될 수 있다면 다중요인모델 등에 의한 문서 분석 기법은 T에 한정된 속성(단어 분포 등)들이 잘 반영된 의미 자질들을 추출할 수 있을 것이다. 그러나 그런 경우가 아니라면, 추출된 자질은 T에 의해 정의되는 토픽에 대한 구체적 자질을 잘 반영할 수 없을 것이다. 이러한 가정 하에, 질의문들에 대한 관련 문서들의 비관련문서와의 이산도(dispersion) 측정을 통해 각 문서에 대한 정량적 검토를 해보았다. 질의 q에 대한 관련문서 집합 T의 중심 벡터를 c라 하고 T에 속한 문서들의 c에 대한 유사도 측정치의 평균을 s라 할 때, 해당 질의에 대한 이산도는 s보다 상위 값을 가지는 비관련문서들의 전체 비

표 4 MED, CACM, CISI, CRAN 문서 집합의 문서 및 질의문 수

	MED	CACM	CISI	CRAN
문서(초록) 수	1,033	3,204	1,460	1,398
질의 수	30	51	76	225

1) 검색 질의어(query)는 짧은 질의의 하나의 문서로 취급한다.

표 5 각 방법론에 따른 변환 행렬 및 유사도 측정 방식. GVSM, LSI, 다중요인모델의 변환행렬의 경우 3.2절에 설명된 행렬을 사용한다. k-means 클러스터링의 경우 행렬 C의 각 열은 spherical k-means 알고리즘[3] 수행 후의 각 문서 클러스터의 중심벡터를 포함한다.

구 분	커 널	변환행렬 S	유사도
기본 VSM(BVSM)	k_{BVSM}	S_{IDF}	$k_{BVSM}(q, d)$
Generalized VSM(GVSM)	$k_S = k_{GVSM}$	$S_{IDF} \times D$	$\frac{1}{2}(k_{BVSM}(q, d) + k_{GVSM}(q, d))$
k-means (KM)	$k_S = k_{KM}$	$S_{IDF} \times C$	$\frac{1}{2}(k_{BVSM}(q, d) + k_{KM}(q, d))$
LSI	$k_S = k_{LSI}$	$S_{IDF} \times U$	$\frac{1}{2}(k_{BVSM}(q, d) + k_{LSI}(q, d))$
다중요인모델(MCM)	$k_S = k_{MCM}$	$S_{IDF} \times G$	$\frac{1}{2}(k_{BVSM}(q, d) + k_{MCM}(q, d))$

표 6 MED, CACM, CISI, CRAN 문서 집합에 대한 평균 정확도

	BVSM	GVSM	LSI	k-means	MCM
MED	0.525	0.589	0.621	0.609±0.013	0.648±0.009
CACM	0.288	0.250	0.258	0.233±0.003	0.351±0.007
CISI	0.221	0.193	0.224	0.205±0.003	0.235±0.002
CRAN	0.384	0.410	0.411	0.411±0.001	0.435±0.001

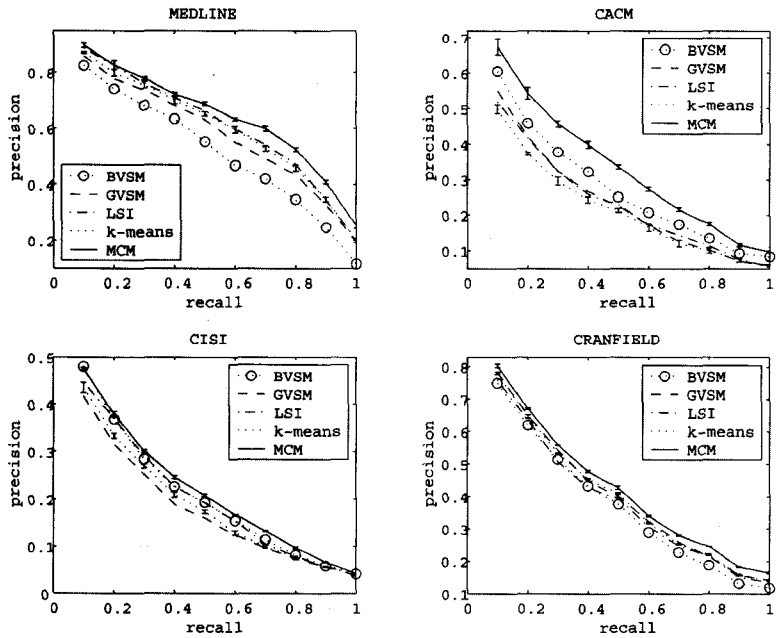


그림 3 MED, CACM, CISI, CRAN 문서 집합에 대한 재현율-정확율 그래프

관련 문서에 대한 비율로 정의하였다. 실험에 사용된 4개 문서 집합의 모든 질의에 대한 평균 이산도는 각각 MED=0.22%, CRAN=0.12%, CACM=2.1%, CISI=5.7%였다. CISI의 경우 가장 이산도가 높은데, 실험에서도 CISI에 대해서는 의미 커널에 의한 방법의 성능 향상의 정도가 미미하였을 뿐 아니라, 기본 성능 역시 다른 문서 집합에 비해 상당히 낮음을 알 수 있다. 이와는 달리, MED 데이타와 CRAN 데이타의 경우 상대적으로 아주 낮은 값을 갖는데, 표 6의 결과에서 볼 수 있듯이 기본 BVSM에 대해 나머지 모든 의미 커널 방법론들이 일정한 수준의 성능 향상을 이루었다. 따라서, 의미 커널에 의한 방법의 유효성과 각 문서 집합에 대한 성능

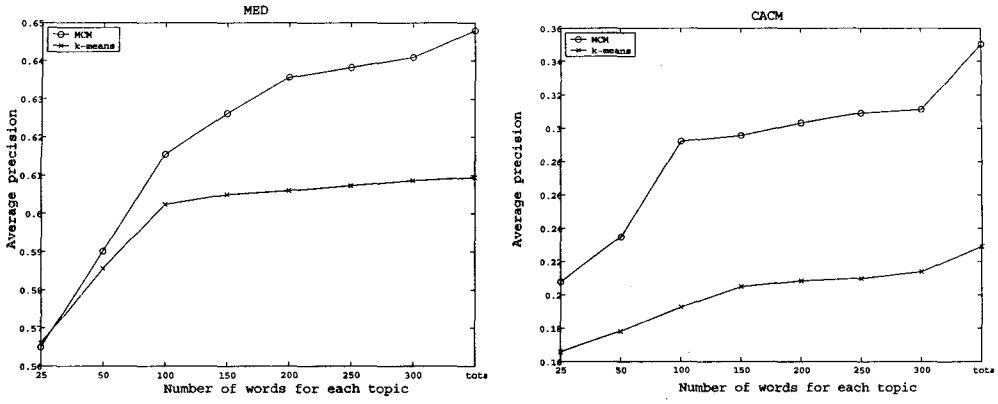


그림 4 토픽당 단어수에 따른 다중요인모델과 k-means의 성능 변화

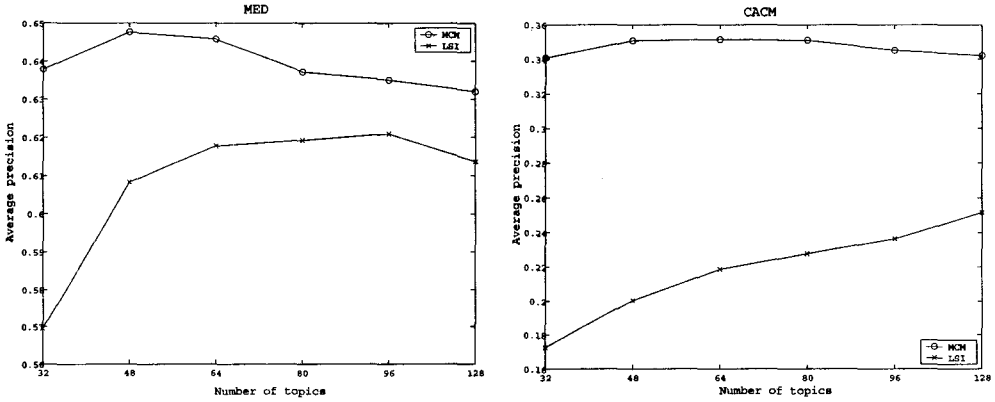


그림 5 추출된 토픽 수에 따른 다중요인모델과 LSI의 성능 변화

의 편차는, 질의문에서 요구하는 관련문서들의 집합과 이를 직접 반영하지 않는 데이타 기반의 무감독 학습이 추출하는 토픽들의 괴리 정도에 의해 설명할 수 있을 것이다.

그림 4는 다중요인모델과 k-means 군집화에 대해 토픽을 정의하는 단어들의 개수에 따른 성능을 비교한 것으로서, 추출된 각 토픽에 대해 가중치가 큰 상위 N (=25, 50, 100, 150, 200, 250, 300) 개의 단어들만으로 해당 토픽을 정의하고 각 경우에 대해 그 성능의 변화를 살펴보았다. 다중요인모델의 경우 토픽을 정의하는 관련 단어의 수가 증가할수록 k-means 군집화에 의한 방법에 비해 성능 면에서 보다 큰 향상을 이루는 경향을 볼 수 있다. 이는 다중요인모델의 경우 문서에 대해 하나 이상의 토픽 활성화가 가능하고 또한 확률적 학습 과정을 활용함으로써 특정 토픽 관련의 지역적 자질을 추출할 수 있음과 더불어 각 토픽의 관련 단어들에 대한 보다 평활(smooth)한 분포를 획득할 수 있기 때문인 것으로 판단된다. 그림 5는 MED 데이타와 CACM 데이타에 대해 각 차원 수에 따른 MCM에 의한 방법과 LSI에 의한 방법의 성능을 비교한 것이다. 두 데이타에 대해 다중요인모델의 경우 상대적으로 적은 차원에서의 성능 향상이 두드러진데 이를 통해, MCM의 경우 LSI에 비해 보다 지역적 자질의 추출이 가능함으로써 효과적인 토픽 분석이 가능함을 알 수 있다.

여러 가지 방법에 의한 성능 차이의 통계적 유의성을 측정하기 위해 다중요인모델에 의한 성능을 기준으로 다른 방법론들에 의한 성능과의 paired t-test와 paired Wilcoxon signed-rank test[22]를 수행하였다(표 7). 서로 다른 두 기법 A와 B의 비교시, 두 검증 방법 모두 각 질의(query)에 대한 A와 B의 점수 차이를 평가척도로 하는데, paired t-test는 A와 B에 의한 성능 차이의 크기를 각 질의문에 대한 차이의 분산과 비교하는 평가법이며, paired Wilcoxon signed-rank test는 각 질의문에 대한 점수차를 그 절대값의 전체 질의문에서의 순위값으로 대체하여 A와 B의 성능을 비교하는 평가법으로 paired t-test에 대응되는 비모수적(non-

표 7 4개의 문서 집합에 대한 다중요인모델 기반 의미커널에 의한 성능의 통계적 유의성 테스트 결과. 각 항에 명시된 기호의 의미는 p-value에 따른 유의도(significance)를 나타낸다. 즉, '≫': p-value < 0.01, '>': 0.01 ≤ p-value < 0.05, '~': p-value > 0.05.

	paired t-test				paired Wilcoxon signed-rank test			
	BVSM	GVSM	k-means	LSI	BVSM	GVSM	k-means	LSI
MED	≫	≫	≫	>	≫	≫	≫	≫
CACM	≫	≫	≫	≫	≫	≫	≫	≫
CISI	>	≫	≫	~	≫	≫	≫	>
CRAN	≫	≫	≫	≫	≫	≫	≫	≫

parametric) 기법 중의 하나이다. 표 7에서 다중요인모델에 의한 성능의 다른 방법론들에 의한 성능과의 비교시, 대부분의 경우(MED 데이타에 대한 BVSM 기법 및 MED, CISI 데이타에 대한 LSI 기법과의 비교는 제외)에 p-value가 0.01 이하였으며 이는 표 6에 나타난 성능 차이가 통계적으로 유의미함을 의미한다.

5. 결 론

본 논문에서는 은닉변수모델의 하나인 다중요인모델 기반의 텍스트 문서 분석을 통해 토픽 단어를 추출하고 이를 활용한 의미 커널 구축 기법을 제시하였다. 다중요인모델에서 하나의 문서는 하나 이상의 토픽들의 결합으로 표현되며, 하나의 토픽 또는 주제는 각 은닉노드에 의해 파악된다. 이러한 모델에서의 학습을 용이하게 하기 위한 근사적 방법으로서, 헬름홀츠 머신과 wake-sleep 알고리즘에 의한 학습 기법을 활용하였다. 다중요인 네트워크 상에서 은닉노드들에 의해 정의되는 토픽들은 단어들에 대한 가중치 집합으로 표현하였으며, 이를 통해 단어들의 상관관계를 고려한 개념 수준에서의 문서간 비교를 위한, 말뭉치 기반의 의미 커널을 구축하였다.

토픽 단어 추출을 위한 TDT-2 문서 데이타에 대해 헬름홀츠 머신 학습에 의한 다중요인모델을 적용함으로써, 추출된 단어 집합이 특정 토픽을 어느 정도 잘 표현함을 제시된 예를 통해 확인할 수 있었으며, 하나 이상의 토픽의 조합으로서 문서가 표현될 수 있음을 보였다. 4개의 표준 문서 집합인 MED, CACM, CISI, CRAN 문서 데이타에 대한 실험에서는 기본 벡터공간모델을 비롯한 다양한 의미 커널의 성능을 비교하였다. 다중요인모델에 기반한 의미 커널은 단순 단어 기반의 기본 벡터공간 모델에 비해서는 네 개의 문서 집합에 대해 모두 향상된 성능을 보였으며, Generalized VSM 등의 다른 의미 커널 구축방법론들에 비해서도 대부분 통계적으로 유의미한 수준의 성능 향상을 보였다.

논문에서 제시된 다중요인모델에서는 은닉층이 하나이고 은닉노드간의 연결이 존재하지 않는다. 따라서 텍스트 분석시 토픽간의 상관관계는 고려하지 않는데 이의 분석이 가능하다면 텍스트 문서에 대한 보다 효과적인 표현 및 이해가 가능할 것이다. 은닉노드간 측위 연결(lateral connection)이 있거나 계층적 구조를 갖는 모델에 기반한 학습이 이를 위한 가능한 방법이 될 수 있을 것이다. 또한 다중요인모델에서 토픽 추출을 위한 최적 자질 노드 수 결정은 데이타의 효율적 모델링이라는 모델 자체적인 측면 뿐 아니라 추출된 의미 자질을 이용한 효과적 문서간 유사도 측정이라는 면에서 중요한 문제이다. 향후 이에 대한 심도있는 연구가 필요하며, 최적 모델 선정이라는 관점에서 통계적 선택 척도나 교차 검증(cross validation)등에 의한 방법을 고려해 볼 수 있을 것이다.

참 고 문 헌

[1] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. A., Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990.

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

[3] Dhillon, I. and Modha, D., Concept decomposition for large sparse text data using clustering. *Machine Learning*, vol. 42, pp. 143-175, 2001.

[4] Kolenda, T., Hansen, L. K. and Sigurdsson, S., Independent components in text. In *Proceedings of ICA'99*, 1999.

[5] van Rijsbergen, C. J., *Information Retrieval*, London: Butterworths, 2nd Edition, 1979.

[6] Jiang, F. and Littman, M. L., Approximate dimension equalization in vector-based information retrieval. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 423-430, 2000.

[7] Cristianini, N., Shawe-Taylor, J., and Lodhi, H., Latent semantic kernels. *Journal of Intelligent Information Systems*, vol. 18, no. 2/3, pp. 127-152, 2002.

[8] M. W. Berry, S. T. Dumais, and G. W. O'Brien.,

Using linear algebra for intelligent information retrieval. *SIAM Review*, vol. 37, no. 4, pp. 573-595, 1995.

[9] Dumais, S. T., Furnas, G. W., Landauer, T. K., and Deerwester, S., Using latent semantic analysis to improve information retrieval. In *Proceedings of CHI'88*, pp. 281-285, 1988.

[10] Dumais, S. T., Latent semantic indexing (LSI): TREC-3 report, In *Proceedings of the Text Retrieval Conference (TREC-3)*, pp. 219-230, 1995.

[11] Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S., The Helmholtz machine. *Neural Computation*, vol. 7, pp. 889-904, 1995.

[12] Frey, B. J., *Graphical Models for Machine Learning and Digital Communication*, The MIT Press, 1998.

[13] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[14] Hinton, G. E., Dayan, P., Frey, B. J., Neal, R. M., The wake-sleep algorithm for unsupervised neural networks. *Science* 268, pp. 1158-1161, 1995.

[15] Dempster, A. P, Laird, N. M., and Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, vol. 39, pp. 1-38, 1977.

[16] Chang, J.-H. and Zhang, B.-T., Using stochastic Helmholtz machine for text learning, In Proceedings of International Conference on Computer Processing of Oriental Languages, pp. 453-458, 2001.

[17] Salton, G. and McGill, M. J., *Introduction to Modern Information Retrieval, McGraw-Hill*, 1983.

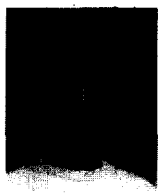
[18] Siolas, G. and d'Alche-Buc, F., Support vector machines based on a semantic kernel for text categorization, In *Proceedings of the International Joint Conference on Neural Networks*, vol. 5, pp. 205-209, 2000.

[19] Wong, S. K. M., Ziarko, W., and Wong, P. C. N., Generalized vector space model in information retrieval, In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 18-25, 1985.

[20] Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[21] Slonim, N. and Tishby, N., Document clustering using word clusters via the information bottleneck method. In *Proceedings of ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 208-215, 2000.

[22] Hull, D., Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 329-338, 1993.



장 정 호

1995년 서울대학교 컴퓨터공학과 학사
1997년 서울대학교 컴퓨터공학과 석사
1997년~현재 서울대학교 컴퓨터공학부 박사과정. 관심분야는 기계학습, 은닉변수모델, 텍스트 마이닝, 생물정보학

장 병 탁

정보과학회 논문지 : 소프트웨어 및 응용 제 31 권 제 3 호 참조