헬름홀츠머신 학습 기반의 의미 커널을 이용한 문서 유사도 측정

장정호*° 김유섭** 장병탁*
*서울대학교 컴퓨터공학부 **한림대학교 정보통신공학부
jhchang@bi.snu.ac.kr° yskim01@hallym.ac.kr btzhang@bi.snu.ac.kr

Estimation of Document Similarity using Semantic Kernel Derived from Helmholtz Machines

Jeong-Ho Chang*° Yu-Seop Kim** Byoung-Tak Zhang*
*School of Computer Science and Engineering, Seoul National University
**Division of Information Engineering and Telecommunication, Hallym University

요 약

문서 집합 내의 개념 또는 의미 관계의 자동 분석은 보다 효율적인 정보 획득과 단어수준 이상의 개념 수준에서의 문서 비교를 가능하게 한다. 본 논문에서는 은닉변수모델을 이용하여 문서 집합으로부터 단어들 간의 의미관계를 자동적으로 추출하고 이를 통해 문서간 유사도 측정을 효과적으로 하기 위한 방안을 제시한다. 은닉변수 모델로는 다중요인모델의 학습이 용이한 헬름홀츠 머신을 활용하며 이의 학습 결과에 기반하여, 문서간 비교를 위한 의미 커널(semantic kernel)을 구축한다. 2개의 문서 집합 MEDLINE과 CACM 데이터에 대한 검색 실험에서, 제안된 기법을 적용함으로써 기본 VSM(Vector Space Model) 에 비해 20% 이상의 평균 정확도 향상을 이룰 수 있었다.

1. 서 론

인터넷의 발달과 이에 따른 정보량의 폭발적 증가로 온라인 텍스트나 전자화된 문서의 양이 크게 증대되고 있다. 하지만 이러한 데이터의 방대함이 곧바로 사용자들의 요구 정보 획득의 용이함을 의미하는 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, 문서 검색, 문서 분류, 문서 군집화 등의 작업을 자동화하기 위해 기계학습이나 통계적 알고리즘에 기반한 방법론이 많이 적용되고 있다.

정보탐색 및 획득을 위한 핵심 내용 중의 하나는 사용자의 정보 요구에 대한 정보의 관련성(relevance)에 관한 것인데[9], 이는 문서의 표현 방식과 직결된 문제라고 할 수 있다. 기본 벡터공간모델(VSM)에서의 'bag-of-words' 방식은 간단하면서도 어느 정도 좋은 성능을 내지만, 단어들 간의 의미관계를 고려하지 못한다는 점에서 문제가 있으며[1, 6], 이는 유의어와 다의어에 관련된 문제로 볼 수 있다. 이러한 문제를 극복하고자 하는 노력 중의 가장 대표적인 것이 Latent Semantic Analysis에 기반한 문서 인덱싱이며, 이 기법은 단어의 공기관계에 대한 분석을 통하여 문서들을 의미 공간(semantic space)으로 매핑하여 문서들 간의 의미관계를 파악하고자 시도한다[3, 7].

본 논문에서는 단어들간의 의미관계를 문서간 유사도 측정에 활용하기 위한 방안으로서, 은닉변수모델의 일종인 헬름홀츠 머신(Helmholtz machine)을 이용하여 텍스트 문서로부터 의미적으로 유사하거나 적어도 동일 주제에 관련된 단어들의 집합을 추출하고 이에 기반한 의미 커널(semantic kernel)을 구축하는 기법을 제시한다. 그리고 정보 검색에서의 문서간 유사도 측정[1] 실험을 통해 그 유용성을 확인한다.

2절에서는 의미 커널의 개념에 대해 설명하며 3절에서는 헬름홀츠머신과 이에 기반한 의미 커널 구축에 대해 설명한다. 4절에서는 2개의 문서 집합에 대해 의미 커널을 이용한 문서 검색 실험을 보이며, 5절에서는 결론 및 향후 연구방향을 제시한다.

1) 정보검색에서 사용자가 입력한 질의문을 하나의 짧은 문서로 간주할 경우 이는 결국 두 문서간의 유사도 계산 문제로 생각할 수 있다.

2. 의미 커널

정보검색을 위한 벡터공간 모델에서 두 문서 벡터 d_1와 d_2의 유사도는 일반적으로 다음과 같이 표현할 수 있다[1].

$$sim(d_1, d_2) = (P^T d_1)^T \cdot (P^T d_2) = d_1^T P P^T d_2 \qquad (1)$$

행렬 P는 기본 벡터 공간에서의 문서를 특정한 다른 자질공간으로 매핑시키기 위한 변환행렬로서 $P = I_M$ (M은 문서 집합 내 어휘 개수이고 I_M은 $M \times M$ 항등행렬)일 때는 원래 단어벡터 공간에서의 두 문서간 유사도를 의미한다. 하지만, 이 경우에는 각 단어들에 의해 표현되는 축이 벡터공간상에서 직교를 이루게 되므로(orthogonal) 단어들의 상관관계를 고려할 수 없다는 점에서 종종 그 문제점이 지적되곤 한다[1, 3, 5, 7].

이러한 문제점을 해결하기 위해서는 결국 단어간의 상관관계에 대한 정보를 고려할 필요가 있다. Wong[10]은 Generalized VSM(GVSM)을 제안하였는데 이는 문서집합 내에 나타나는 단어들 간의 공기정보를 이용하는 방식이다. N개의 문서로 구성된 문서 집합 D가 주어질 때, GVSM에서는 $P = D$이다. 외부 지식으로부터 이를 유도하는 경우도 있는데, 보통 시소러스 정보를 이용한다. Siolas와 d'Alche-Buc[8]은 온라인 어휘 의미망(semantic network)인 워드넷(WordNet)을 활용하여 단어들 간의 의미적 유사성을 계산하고 이를 P행렬로 사용하였다. LSI(latent semantic indexing)[3] 기법에서는 단어-문서 행렬 D에 대한 SVD(singular value decomposition)를 수행하고 ($D = USV^T$) 상위 k개의 left singular vector들을 이용하여 문서간 유사도는 다음과 같이 측정한다.

$$sim_{LSI}(d_1, d_2) = (I_k U^T d_1)^T \cdot (I_k U^T d_2) = d_1^T U_k U_k^T d_2 \quad (2)$$

Cristianini[1]는 SVM과 같은 커널(kernel) 방법론에서의 커널 관점에서 이러한 문서 유사도 측정을 설명하고 있다. 임의의 두 문서와 자질공간으로의 사상 ϕ에 대해 커널 함수는 $k(d_i, d_j) = < \phi(d_i), \phi(d_j) >$로 정의되며 이는 커널을 선택한다는 것은 묵시적으로 데이터(문서)가 표현된 자질 공간을 선택하는 것을 의미한다. 문서 유사도 측정에서의 변형행렬 P는 결국 ϕ를 정의하는 것으로 볼 수 있으며, 이 때 $\phi(d) = P^T d$로 주어진다. $k(d_1, d_2)$에서 행렬 P를 통해 단어들간의 의미 관계를 고려하여 문서간 유사도를 측정할 수 있을 때 이를 "의미 커널(semantic kernel)" 이라 한다 [1].

3. 헬름홀츠 머신에 의한 의미 커널 구축

여기에서는 헬름홀츠머신에 의한 문서 데이터 분석과 이에 가반한 의미 커널에서의 변환 행렬 P의 정의와 문서간 유사도 계산에 대해 서술한다.

헬름홀츠 머신(Helmholtz machine)[2]은 다중요인모델 계층적 생성 모델에서의 학습과 추론 과정을 용이하게 하기 위한 근사화 방법 중의 하나로서, 뉴런 형태의 확률적 처리 단위로 구성된 다층 연결망으로 표현된다. 그림 1은 은닉층이 하나인 간단한 헬름홀츠 머신을 보인다. 실선은 은닉층으로부터의 생성모델(generative model)을 위한 연결선을 나타내고, 점선은 입력층으로부터의 인식모델(recognition mdel)을 위한 연결선을 나타낸다. Φ와 θ는 각각 생성모델과 인식모델의 매개변수 집합을 의미한다. 텍스트 문서에 대한 분석시, 각 입력노드는 하나의 단어에 해당하며, 각 은닉노드는 텍스트에서의 하나의 토픽 또는 잠재의미 자질(latent semantic feature)로서 이는 해당 노드로부터의 연결 가중치가 높은 단어들의 집합으로 정의한다.

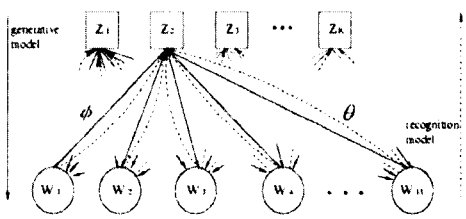

그림 1. 은닉층이 하나인 헬름홀츠 머신

헬름홀츠 머신은 다중요인모델의 학습에서 각 은닉변수 조합적 상태 $z = (z_1, z_2, \ldots, z_K)$에 대한 사후확률분포 $P(z|\, d_n)$의 추정을 용이하게 하기 위해 인식 네트워크에 의한 다음과 같은 근사화 방법을 제공한다.

$$Q(z|\, d_n) = \prod_{k=1}^{K} Q(z_k|\, d_n) \quad (3)$$

$Q(z_k|\, d_n)$은 시그모이드(sigmoid) 함수를 통해 계산된다.

헬름홀츠 머신에서의 생성모델과 인식모델의 매개변수 추정을 위해 온라인 학습 알고리즘인 wake-sleep 알고리즘[2, 5]을 이용한다. 단어 벡터로 표현된 하나의 문서 벡터가 주어지면, wake-단계에서는 인식 모델을 이용하여 각 잠재의미(토픽)들이 활성화될 확률을 구하고 이 확률에 따른 샘플링(1/0)을 통하여 실제 잠재 의미들의 활성화를 결정한다. 각 잠재의미에 대한 샘플링결과와 입력 문서 벡터를 이용하여 생성모델의 매개변수 값을 갱신한다. Sleep 단계에서는 wake-단계와는 반대로, 앞서 수정된 생성모델의 매개변수를 이용하여 가상의 잠재의미 조합을 샘플링하고 그 결과를 기초로 하나의 가상 문서데이터를 생성해 낸다. 이렇게 결정된 값들을 이용하여 인식모델의 매개변수 값을 갱신한다. 이러한 wake-단계와 sleep 단계를 모든 데이터에 대해 반복적으로 수행하면서 모델의 매개변수를 추정한다.

헬름홀츠머신 기반 문서 분석에서 최종적으로 활용되는 것은 은닉노드에서 입력노드에 이르는, 즉 각 토픽이 단어들에 대해 가지는 가중치들의 집합이며, 이는 $M \times K$ 행렬 R로 표현할 수 있다. M은 문서집합을 구성하는 어휘 개수, K는 은닉공간의 차원, 즉 설정된 토픽의 개수를 의미하며 R의 각 열은 M개의 어휘에 대한 가중치 집합으로 정의되는 하나의 잠재의미 자질이다. 보통 $K \ll M$이므로 이는 LSA와 같이 차원 감소의 효과가 있으며, 하나의 토픽 내 가중치가 높은 단어들은 의미나 관련 주제면에서 연관도가 높다고 가정할 수 있다. 따라서 $k_{HM}(d_1, d_2) = < R^T d_1, R^T d_2 >$은 두 문서간의 유사도 계산 시 단어들 간의 의미관계를 고려할 수 있으며, 이를 통해 변환행렬 $P = R$인 의미 커널이 구축된다.

4. 실 험

문서 데이터 집합 2 개 (MEDLINE, CACM)에 대해 문서 검색 실험을 하였으며, 두 문서 집합의 구성은 다음과 같다.

- MEDLINE: 미국 국립 의학도서관으로부터의 초록 1,033개로 구성되어 있으며 질의어 수는 30개이다.
- CACM: CACM 저널의 초록 3,204개로 구성되어 있으며 질의어 수는 51개이다.

불용어 제거와 스테밍(stemming) 과정을 거친 후, 각 문서의 어휘 크기는 각각 7,014와 5,056이다. 각 질의어에 대해 문서와의 유사도를 측정하기 위해 기본 벡터 공간에서의 코사인 유사도 측정치와 은닉공간상에서의 유사도 측정치의 평균값을 취하였다. 질의문 q와 문서 d 간의 유사도 $sim(q, d)$는 다음과 같이 정의된다.

$$sim(q, d) = \frac{1}{2}(q^T d + q^T R R^T d) = \frac{1}{2} q^T (I + RR^T) d$$

그림 2는 두 문서 집합에 대한 precision-recall 그래프를 보인다. 기본 벡터공간 모델에서의 코사인 유사도 측정치에 의한 결과(word-space)를 기본성능으로 하고 의미 커널에서의 변환행렬 P를 유도하기 위한 방법으로서 generalized VSM(doc-index), k-means 알고리즘[2], LSA, 헬름홀츠 머신(HM)

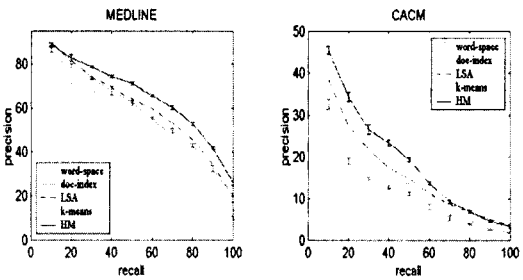

그림 2. MEDLINE과 CACM 문서 집합에 대한 문서 검색 성능.

에 의한 방법을 적용하였다. 그리고 k-means 알고리즘, LSA, 헬름홀츠 머신의 경우에는 은닉 공간 차원 K(k-means 알고리즘의 경우 클러스터의 수)를 $K=16, 32, 48, 64, 80, 96$으로 변화시켜 나가면서 그 중 가장 좋은 성능을 보인 경우를 선택하였다. 결과를 보면, MEDLINE 데이터에 대해서는 generalized VSM, k-means, LSA, 다중요인모델 모두 기본 VSM보다 향상된 성능을 보임을 알 수 있다. 하지만 CACM 데이터에 대해서는 LSA와 다중요인모델은 성능 향상이 있었지만, generalized VSM과 k-means 알고리즘에 기반한 의미 커널은 오히려 기본 VSM에 비해 성능이 저하되는 것을 확인할 수 있었다. 이러한 결과로 볼 때, 단순한 단어 공기 정보 이용이 문서간 유사도 측정에서 반드시 도움이 되지 않음을 알 수 있다. 표 1은 두 문서 집합에 대한 각 방법들의 11-point 평균 정확도 면에서의 성능을 요약한다.

	MEDLINE	CACM
word-index	52.7%	16.9%
doc-index	59.8%	15.2%
k-means	60.9±1.11%	13.3± 0.54%
LSA	62.3%	17.4%
HM	**65.3±0.77%**	**21.2± 0.64%**

표 1. 두 문서 집합에 대한 11-point 평균정확도. k-means와 HM의 경우에는 학습 초기 값에 따라 학습결과가 달라지므로 10번 시행 후 평균 성능과 표준편차를 표시함.

그림 3은 은닉변수모델의 차원 수에 따른 11-point 평균 정확도 값을 보인다. 헬름홀츠 머신 기반의 의미 커널을 사용한 경우에는 초기부터($K=16$) 눈에 띌만한 정확도 향상을 보이고 이후 차원 수에 따라 성능 차의 변화가 적은데 반해, LSA의 경우 초기에는 별 성능 향상이 없다가 $K=80$일 때 최고 성능을 보이며 그 이후에는 거의 변화가 없음을 알 수 있다. 비록 MEDLINE 문서 집합에 국한된 경우긴 하지만, 헬름홀츠 머신에 의한 방식이 훨씬 작은 차원만으로 LSA에 의한 최고 성능보다도 약간 우수한 성능을 보였다는 점에서 의미있는 결과라고 할 수 있다.

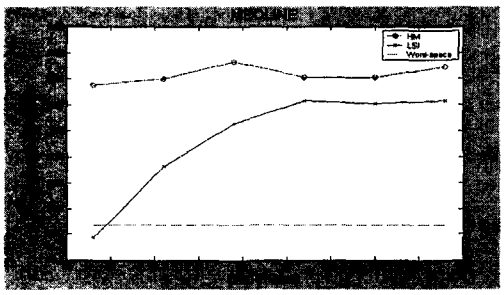

그림 3. MEDLINE 문서 집합에 대한 문서 검색에서 은닉 공간 차원에 따른 성능 측정. 16, 32, 48, 64, 80, 96 차원에 대한 11-point 평균정확도를 보인다.

5. 결 론

본 논문에서는 의미적으로 관련 있거나 적어도 동일 주제와 관련된 단어 집합을 헬름홀츠 머신 학습에 의해 텍스트 문서로부터 추출하고 이를 기반으로해서 의미 커널을 구축하고 이를 이용하여 문서 유사도를 측정하는 방법을 제시하였다. 2개의 문서 집합에 대한 실험에서, 기본 벡터공간모델에서의 단순 코사인 유사도에 의한 방법에 비해 단어들의 의미 관계 정보를 활용한 방법이 문서 검색 성능 면에서 더 우수함을 확인할 수 있었다. 또한 헬름홀츠 머신에 기반한 의미 커널 구축 방식이 generalized VSM이나 k-means 방식에 비해 평균 정확도 면에서 우수하였으며, 특히 LSA 기법과 비교할 때 은닉공간의 차원이 상대적으로 작을 경우에도 좋은 성능을 달성할 수 있었다.

향후 보다 대규모 문서 데이터에 대한 실험 및 성능 평가와 더불어, 추출된 단어 집합들(잠재의미 자질) 간의 의미관계를 파악하고 이를 문서 유사도 측정에 반영하기 위한 방법론에 대해 연구를 진행할 것이다. 또한 Support vector machine(SVM)과 같은 커널 분류기를 이용한 문서 분류에서 제안된 방법을 활용하고자 한다.

감사의 글

본 연구는 과학기술부 뇌신경정보학 사업(BrainTech)과 교육부 BK21-IT 프로그램에 의하여 일부 지원되었음. 이 연구를 위해 연구 장비를 지원하고 공간을 제공한 서울대학교 컴퓨터신기술 공동연구소에 감사드립니다.

참고문헌

[1] 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.

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

[3] 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.

[4] Dhillon, I. S. and Modha, D. S., Concept decomposition for large sparse text data using clustering, *Machine Learning*, 47(1), pp. 143-175, 2001.

[5] Frey, B. J.,. *Graphical models for machine learning and digital communication*, The MIT Press, 1998.

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

[7] Landauer, T. K., Foltz, P. W., and Laham, D., "An Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, pp. 259-284, 1998.

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

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

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

2) [4]에서 제시된 *spherical k-means* 알고리즘을 적용하였다.