논문 데이터베이스에서의 LDA 기반 텍스트 유사도 계산 방안

엄태환, 윤석호, 배덕호, 김상욱
한양대학교 전자컴퓨터통신공학과
e-mail: {mgdjaxo, bogely, dhbae, wook}@agape.hanyang.ac.kr

LDA-based Text Similarity in Scientific Literature Databases

Tae-Hwan Eom, Seok-Ho Yoon, Duck-Ho Bae, Sang-Wook Kim
Department of Electronics and Computer Engineering, Hanyang University

요 약

본 논문에서는 기존의 LDA 기반 유사도 계산 방안의 논문 데이터에 대한 적합성을 검증한다. 실제 논문 데이터를 이용해 기존텍스트 유사도 계산 방안과 LDA 기반 유사도 계산 방안의 정확도를 비교함으로써 논문 데이터베이스에서의 LDA 기반 텍스트 유사도 계산 방안의 유용성을 검증한다.

1. 서론

문서들 간의 유사도는 문서 검색, 문서 군집화 등 다양한 연구에 기반 정보로 활용될 수 있다. 최근 들어 문서 내의 단어들을 이용하여 문서들을 모델링하는 방안들이 연구되어져 왔다. 대표적으로 LSI (Latent Semantic Indexing), PLSA (Probabilistic Latent Semantic Analysis), LDA (Latent Dirichlet Allocation)[1] 등이 있다. 특히, LDA는 단어들의 분포와 단어들 간의 co-occurrence를 동시에 고려할 수 있기 때문에 문서들 간의 유사도를 계산하는데 활용될 수 있다.

본 논문에서는 기존 LDA 기반 텍스트 유사도 계산 방안이 논문 데이터에 적합한지에 대해 검증해보고자 한다.

2. 관련연구

2.1. 텍스트 유사도 계산 방안

문서에 출현하는 단어들을 이용해서 두 문서 사이의 유사도를 계산하는 대표적인 텍스트 유사도 계산 방안으로는 cosine measure가 있다. Cosine measure는 문서가 가지고 있는 단어와 단어들의 빈도를 벡터로 표현하고, 두 벡터 간의 cosine 값을 두 논문 간의 유사도로 간주한다[2]. 이 때, TF-IDF 방안을 통해 단어들의 중요도에 따라 단어에 가중치를 다르게 부여할 수 있다.

2.2. LDA (Latent Dirichlet Allocation)

LDA는 단어 집단 (text corpus)과 같은 이산형 데이터의 집합을 위한 generative probabilistic model이며 기계학습 분야에서 문서들의 주제 분석을 위해 많이 사용되어 왔다[1]. LDA에서 하나의 문서는 여러 개의 단어들로 구성되며, 하나의 단어는 여러 개의 잠재적인 주제에 속할 수 있다고 가정한다. 또한, 하나의 주제는 여러 단어들의 확률 분포로 구성되며, 각각의 주제는 서로 다른 단어들의 분포를 가진다고 가정한다.

LDA는 전체 문서들의 집합과 문서 집합에 잠재된 주제 개수 (k)를 입력 받아 k개의 주제와 각 주제의 단어 분포를 자동적으로 학습한다.

3. LDA 기반 유사도 계산

최근 LDA 기반 유사도 계산 방안들이 연구 되어 왔다[3]. 본 논문에서는 기존의 LDA 기반 유사도 계산 방안을 논문 데이터베이스에 적용하고, 해당 방안의 유용성을 검증해보고자 한다.

LDA 기반 유사도는 다음과 같은 과정을 통해 계산된다.

(1) LDA를 통해 각 주제의 단어 분포를 학습한다.
(2) 학습된 단어 분포를 이용하여 각 논문의 주제 벡터를 계산한다. 주제 벡터는 해당 논문에 각각의 주제에 속할 확률의 분포를 나타낸다.
(3) 두 주제 벡터 간의 cosine 값을 계산한다.

논문 데이터베이스에서 LDA를 이용한 유사도 계산 방안은 다음과 같은 장점을 가진다. 첫째, LDA는 주제의 단어 분포를 학습할 때, 한 문서에 출현한 단어의 빈도뿐만 아니라, 전체 문서들의 집합에 출현한 단어의 빈도도 함께 고려하기 때문에 각 단어의 중요도를 반영할 수 있다.

둘째, 단어들 간의 co-occurrence를 반영한다. 예를 들어, '알고리즘'이라는 단어와 '순서도'라는 단어가 동일한 분야에 속하는 문서들에서 빈번히 출현한다면 두 단어의 연관성이 높다고 말할 수 있다. 두 논문 간에는 일치하는 단어가 없이 논문 A에는 '알고리즘'이라는 단어만 출현하고, 논문 B에는 '순서도'라는 단어만 출현한다고 가정하자. cosine measure로 계산한 두 논문 간의 유사도는 0이 된다. 반면, LDA를 통해 학습 과정을 거치면, 두 단어가 동일한 주제에 높은 확률로 속하게 되므로, 논문 A와 논문 B는 높은 유사도를 가지게 된다.

셋째, 한 편의 논문을 벡터로 표현할 때 차원 축소 효과를 볼 수 있다. 기존의 텍스트 유사도 계산 방안의 경우

한 편의 논문은 논문 전체 집합에서 사용된 모든 단어의 수 (N)와 동일한 크기를 갖는 벡터로 표현된다. 최악의 경우, 해당 논문의 실제 크기보다 표현된 벡터의 크기가 더 커질 수도 있다. 반면, LDA에서 한 편의 논문은 크기가 k ($\ll N$)인 주제 벡터로 표현된다. 이로 인해, 주제 벡터를 계산, 저장하는 오버헤드를 크게 줄일 수 있으며, 이로 인한 유사도 계산 성능의 향상을 기대할 수 있다.

4. 실험

본 논문에서는 실험을 위해 DBLP[1]와 Libra[2]에서 수집한 논문 데이터베이스를 구축하였다. 구축된 데이터베이스의 모든 논문들의 유사도를 계산하는 것은 매우 많은 시간이 필요하기 때문에 전체 논문을 이용하지 않고 데이터베이스와 관련된 논문들 중 일부만을 대상으로 실험을 수행한다. 총 논문 수는 106편이며, 총 단어 수는 1,394개, 문서에 출현하는 평균 단어 수는 34개이다.

먼저, LDA를 통해 주제의 단어 분포를 학습하였다. 이 때, 주제 개수는 10으로 하였다. 표 1은 각 주제와 주제에 속할 확률이 가장 높은 단어 5개를 나타낸다. 주제 9의 'mine'은 주제 1과 주제 7 등에서 주요 단어로 선정되었다. 이를 통해 한 단어가 의미가 다른 여러 주제에서 동시에 주요한 역할을 할 수 있음을 알 수 있다.

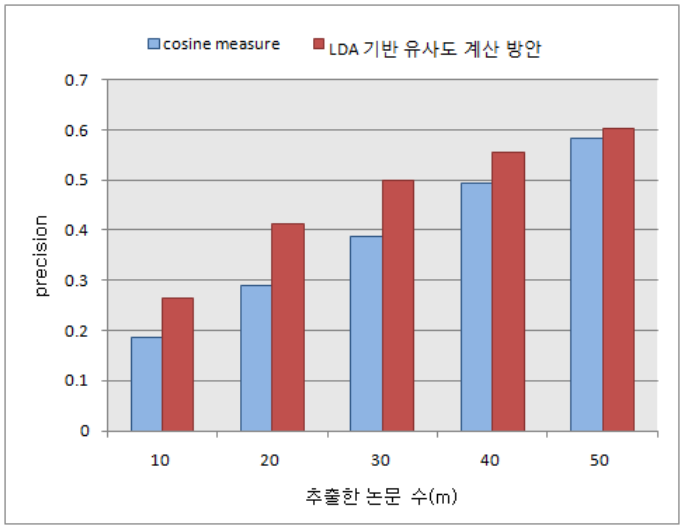
<표 1> LDA 학습 결과

주제1 (Mining Frequent Patterns)	주제2 (Cluster Analysis)	주제4 (Outlier Analysis)	주제7 (Mining Association Rules)	주제9 (Graph Mining)
itemset	spatial	large	rule	mine
frequent	cluster	find	associate	graph
problem	constraint	outlier	mine	frequent
close	queries	dataset	databas	subgraph
find	set	database	interest	pattern
mine	frequent	base	data	data
search	constrain	study	discover	algorithm
algorithm	optimize	identify	larg	structure
rule	presence	deal	user	set
associate	variable	cluster	measur	efficient

둘째, 학습된 주제를 기반으로 각 논문의 주제 벡터를 계산하였다. 마지막으로 논문 데이터베이스에서 LDA 기반 유사도 계산 방안의 적합성을 검증하기 위해 cosine measure와의 정확도를 비교하였다. 유사도의 정확도를 측정하기 위해 [4]에서 사용한 정답 셋을 이용하였다. 정답 셋에는 20개의 분야에 대해 각 분야에 속하는 논문들의 리스트가 저장되어 있다.

정확도 측정을 위해, 정답 셋에 속한 논문 한 편을 시드로 선택한 뒤, 두 유사도 계산 방안으로 계산된 유사도 각각을 기준으로 시드 논문과 가장 유사한 m편의 논문을 추출하였다. 그 후, 추출된 논문들 중 정답 셋에 속한 논문의 precision을 측정하였다. 이 때, cosine measure는 TF-IDF를 적용하였으며, LDA를 이용한 유사도 계산 방안에서 주제 개수는 10으로 설정하였다.

그림 1은 측정된 precision을 나타낸다. x축은 추출한 논문의 수를 나타내고 y축은 precision을 나타낸다. LDA 기반 유사도 계산 방안의 precision이 cosine measure에 비해 높으며, 추출한 논문의 수가 10, 20, 30일 때 큰 차이를 보였다. 이는 앞에서 언급한 LDA의 특징들이 논문 간의 유사도를 정확하게 계산하는데 유용하다는 것을 의미한다.



(그림 1) 각 유사도 계산 방안의 precision.

5. 결론

본 논문에서는 기존의 LDA 기반 유사도 계산 방안의 논문 데이터에 대한 적합성을 검증하였다. 실험 결과, LDA 기반 유사도 계산 방안은 cosine measure에 비해 높은 precision을 보였다.

감사의 글

"이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2008-0061006)."

참고 문헌

[1] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, 2003.

[2] R. Baeza-Yates and B. Riberio-Neto, *Modern Information Retrieval*, ACM Press/Addison-Wesley, 1999.

[3] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-Link LDA: Joint Models of Topic and Author Community," In *Proc. of Int'l. Conf. on Machine Learning*, ICML, pp. 665 - 672, 2009.

[4] S. Yoon, S. Kim, and S. Park, "A Link-Based Similarity Measure for Scientific Literature," In *Proc. of Int'l. Conf. on World Wide Web*, WWW, pp. 1213-1214, 2010.

1) http://www.informatic.uni-trier.de/~ley/db/
2) http://academic.research.microsoft.com