제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

텍스트 기반 논문 유사도 계산 방안

윤석호, 황원석, 김상욱
한양대학교 전자컴퓨터통신공학과
e-mail: bogely@agape.hanyang.ac.kr

A Text-based Similarity Measure for Scientific Literature

Seok-Ho Yoon, Won-Seok Hwang, Sang-Wook Kim
Department of Electronics and Computer Engineering, Hanyang University

요 약

본 논문에서는 텍스트 기반 유사도 계산 방안을 이용하여 논문들 간의 유사도를 계산하는 방안을 제안한다. 논문 데이터베이스에는 논문의 본문이 거의 저장되어 있지 않다. 따라서 논문 데이터베이스에 저장되어 있는 논문의 제목과 요약글들의 키워드들을 이용하여 기존 텍스트 기반 유사도 계산 방안으로 논문들 간의 유사도를 계산할 수 있다. 그러나 논문의 제목과 요약글은 논문의 본문이 가지고 있는 키워드들에 비해서 너무나도 적은 수의 키워드들을 가지고 있기 때문에 해당 키워드들만으로 논문들 간의 유사도를 계산하면 정확도가 낮을 수 있다. 따라서 본 논문에서는 논문을 표현하는 키워드의 수를 증가시키기 위해서 새로운 논문 유사도 계산 방안을 제안한다. 실험을 통하여 제안하는 방안의 우수성을 검증한다.

1. 서론

최근 들어, 학술정보에 대한 사용자들의 관심이 증가하면서 학술 정보 검색 서비스가 발달하기 시작했다. 대표적인 학술 정보 검색 서비스 중에 하나는 사용자가 관심을 가지는 논문과 유사한 주제의 논문들을 찾아주는 유사 논문 검색 서비스이다. 이러한 서비스를 개발하기 위해서는 논문들 간의 유사도를 계산하는 방안이 필요하다. 본 논문에서는 기존 텍스트 유사도 계산 방안을 이용한 논문 유사도 계산 방안에 대해서 논의하고자 한다.

2. 관련 연구

문서들 간의 유사도를 계산하는 기존 연구로는 텍스트 기반 유사도 계산 방안이 있다[1]. 텍스트 기반 유사도 계산 방안은 문서들을 문서 내에 포함하고 있는 키워드들의 집합들로 표현하고 키워드들의 집합들을 서로 비교해서 얼마나 공통적인 키워드들이 많은가로 두 논문의 유사도를 계산한다. 텍스트 기반 유사도 계산 방안은 크게 세 가지 모델로 분류되는데 불리언 모델[1], 벡터 모델[2], 그리고 확률 모델이다[3]. 이들 중에서 가장 많이 사용하는 모델은 벡터 모델이다.

벡터 모델은 일반적으로 다음과 같은 과정을 통해서 문서들 간의 유사도를 계산한다. 먼저, 문서 내에 존재하는 키워드들을 추출한다. 이 때 어떠한 키워드들을 추출할 것인지에 따라서 최종 유사도가 달라질 수 있는데 본 논문에서는 모든 키워드를 추출하여 불필요한 요소들(조사, 특수 기호, stopword 등)을 제거한 후에 해당 키워드들을 사용한다[1]. 문서에서 추출한 키워드들의 집합은 벡터로 표현한다. 이 때 벡터의 각 항은 문서 내에 존재하는 키워드들의 빈도로 결정된다. 따라서 일반적으로 문서 내에 존재하는 키워드들이 많아야 벡터로 표현된 문서들 간의 유사도가 정확하게 계산된다. 마지막으로 벡터들 간의 유사도를 계산해서 문서들 간의 유사도로 사용하는데 벡터들 간의 유사도 계산은 일반적으로 수식 1의 cosine measure를 이용한다. 벡터를 구성하는 방법과 벡터들 간의 유사도를 계산하는 방안 역시 다양한 연구가 진행 되었으나 본 논문에서는 위에서 설명한 기본 방안만을 이용한다[1].

$$Sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} \quad \text{(수식 1)}$$

논문 데이터베이스에는 논문에 대한 다양한 정보가 저장되어 있다. 그러나 논문의 본문은 크롤링(crawling)과 파싱(parsing)에 어려움이 있기 때문에 논문 데이터베이스 내에 텍스트로 저장되어있지 않은 경우가 많다. 따라서 본 논문에서는 논문의 본문 없이 한정된 텍스트 정보만을 활용하여 논문들 간의 유사도를 계산할 수 있는 새로운 텍스트 기반 유사도 계산 방안을 제안하고자 한다.

3. 제안하는 방안

3.1. 기존 방안의 적용

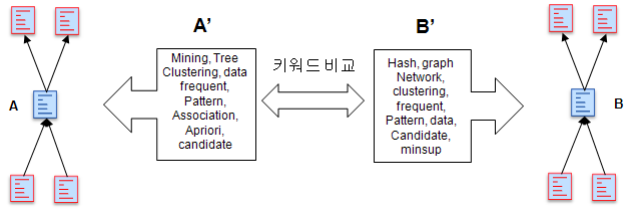
논문 데이터베이스 내에 논문의 본문은 텍스트로 저장되어 있는 경우가 별로 없지만 논문의 제목과 요약글은 일반적으로 저장되어 있다. 논문의 제목과 요약글은 그 논문의 특성을 나타내는 키워드들로 구성되어 있다. 따라서 논문의 제목과 요약글에 있는 키워드들을 대상으로 기존 텍스트 기반 유사도 계산 방안을 적용해서 논문들 간의 유사도를 계산할 수 있다. 그림 1은 논문의 제목과 요약글의 키워드들을 가지고 기존 텍스트 기반 유사도 계산 방안으로 논문들 간의 유사도를 계산하는 방안을 그림으로 나타낸 것이다.

그러나 논문의 제목과 요약글은 논문의 본문이 가지고 있는 키워드들의 수에 비해서 너무나도 적은 수의 키워드들을 가지고 있기 때문에 해당 키워드들만으로 논문들 간의 유사도를 계산하면 정확도가 낮을 수 있다. 따라서 본 논문에서는 논문을 표현하는 키워드의 수를 증가시키기 위해서 새로운 논문 유사도 계산 방안을 제안한다.

(그림 1) 기존 텍스트 유사도 계산 방안을 이용한 논문 유사도 계산 방안.

3.2. 제안하는 방안

논문의 저자는 자신이 작성하는 논문과 관련이 있는 논문들을 참고 문헌에 기록한다. 따라서 해당 논문이 참조하는 논문들과 해당 논문들을 참조하는 논문들은 서로 유사한 주제의 내용을 가지고 있을 확률이 높다. 본 논문에서 제안하는 방안은 논문들 간의 유사도를 해당 논문들이 가지고 있는 제목과 요약글의 키워드들뿐만 아니라 해당 논문이 참조하거나 해당 논문을 참조하는 논문들의 제목과 요약글을 모두 이용하여 벡터를 생성하고 유사도를 계산한다. 그림 2는 제안하는 방안을 그림으로 나타낸 것이다.



(그림 2) 제안하는 논문 유사도 계산 방안.

4. 실험

제안하는 방안을 검증하기 위해서 본 논문에서는 DBLP[1] 에 있는 논문들을 사용했으며 논문들 간의 참조 정보는 Libra[2]에서 크롤링해서 사용하였다. 실험 방법은 본 논문에서 기존 텍스트 유사도 계산 방안을 그대로 적용한 방안과 제안하는 방안을 각각 이용해서 주어진 시드 논문과 가장 유사한 논문 5개를 추출하고 추출된 논문들이 주어진 시드 논문과 주제가 유사한지 살펴보고자 한다. 시드 논문은 데이터베이스 분야에서 저명한 논문[4]으로 선정했다.

<표 1> 각 방안을 이용해서 추출한 시드 논문 [4]와 유사한 상위 5편의 논문들

	기존 방안		제안하는 방안	
순위	논문 제목	유사도	논문 제목	유사도
1	Termination Detection for Diffusing Computations	0.425	Maximizing the spread of influence through a social network	0.632
2	Information Technology Diffusion A Review of Empirical Research	0.423	Summarization and Visualization of Communication Patterns in a Large Scale Social Network	0.604
3	Measuring User Involvement A Diffusion of Innovation Perspective	0.321	A framework for community identification in dynamic social networks	0.601
4	Adaptive Load Diffusion for Multiway Windowed Stream Joins	0.319	Latent Friend Mining from Blog Data	0.572
5	Practical and value compatibility their roles in the adoption diffusion and success of telecommuting	0.248	A framework for analysis of dynamic social networks	0.540

<표 1>은 각 방안을 이용해서 주어진 각각의 시드 논문과 유사한 상위 5개의 논문들을 나타낸다. <표 1>에서 기존 방안으로 추출한 논문들은 시드 논문의 제목에 있는 'diffusion'이라는 키워드가 공통적으로 들어가 있다. 그러나 동일한 키워드가 들어가 있을 뿐 시드 논문의 주제와는 다소 거리가 있는 논문들이다. 제안하는 방안으로 추출한 논문들은 비록 'diffusion'라는 키워드를 제목에 포함하고 있지는 않지만 시드 논문과 유사한 주제인 블로그와 사회연결망과 관련된 논문들이다. 기존 방안은 해당 논문을 설명하는 키워드의 수가 너무 적기 때문에 두 논문이 가지고 있는 키워들 중에 한 두 개의 키워드만 일치해도 유사도가 높게 계산될 수 있다. 따라서 논문들 간의 주제가 유사하지 않는 경우에도 유사도가 높게 계산되는 문제가 발생한다. 그러나 본 논문에서 제안하는 방안은 해당 논문을 설명하는 키워드의 수가 충분하기 때문에 기존 방안보다 유사도 계산 결과가 정확하다.

5. 결론

본 논문에서는 논문의 제목과 요약글을 이용해서 논문들 간의 유사도를 계산하는 방안을 제안했다. 본 논문에서는 논문 데이터베이스에 일반적으로 저장되어 있지 않은 논문의 본문 대신 논문의 제목과 요약글을 이용해서 논문들 간의 유사도를 계산한다. 제안하는 방안은 해당 논문들이 가지고 있는 제목과 요약글의 키워드들뿐만 아니라 해당 논문이 참조하거나 해당 논문을 참조하는 논문들의 제목과 요약글을 모두 이용해서 유사도를 계산한다. 기존 방안과 제안하는 방안들 각각을 이용해서 시드 논문과 유사하다고 판단되는 논문들을 직접 살펴봄으로서 본 논문에서 제안하는 방안의 우수성을 검증했다.

감사의 글

본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업(NIPA-2010-(C1090-1011-0009))의 부분적인 지원과 NHN(주)의 지원을 받았습니다. 그러나 본 논문에서 제시된 의견이나 결론, 또는 권고 등은 온전히 저자(들)의 것이며, 반드시 지원회사의 입장을 대변하는 것은 아닙니다.

참고문헌

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[2] G. Salton and M. Lesk, "Computer Evaluation of Indexing and Text Processing," *Journal of the ACM*, Vol. 15, No. 1, pp. 8-36, 1968.

[3] S. Robertson and K. Jones, "Relevance Weighting of Search Terms," *Journal of the American Society for Information Sciences*, Vol. 27, No. 3, pp. 129-146, 1976.

[4] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information Diffusion through Blogspace." In *Proc. Int'l. Conf. on World Wide Web*, pp. 491-501, 2004.

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