논문 데이터베이스에서 링크 기반 유사도 계산을 위한 정규화 방안

김지수, 윤석호, 김상욱
한양대학교 전자컴퓨터통신공학과
e-mail: kimjisu29@agape.hanyang.ac.kr

Normalization for Link-Based Similarity Measures in Scientific Literature

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

요 약

본 논문에서는 기존 링크 기반 유사도 계산 방안에 사용되는 두 가지 정규화 방안들을 설명하고, 두 정규화 방안 중에서 논문 데이터베이스에 적합한 정규화 방안을 선정한다. 또한, 실제 논문 데이터베이스에 두 가지 정규화 방안을 적용한 기존 링크 기반 유사도 계산 방안의 정확도를 측정함으로써 선정된 정규화 방안이 다른 정규화 방안보다 우수하다는 것을 규명한다.

1. 서론

학술 정보에 대한 사용자들의 관심이 증가하면서 DBLP, Google Scholar, Citeseer와 같은 논문 검색 서비스를 이용하는 사용자의 수가 증가하고 있다. 대표적인 논문 검색 서비스 중 하나는 사용자가 관심 있는 논문과 유사한 논문들을 사용자에게 제공하는 서비스이다. 이러한 서비스를 제공하기 위해서는 논문들 간의 유사도를 계산하는 방안이 필요하다. 본 논문에서는 논문들 간의 유사도를 계산하는 기존 방안과 유사도 계산 방안에 적용되는 정규화 방안에 대해서 논의하고, 논문 데이터베이스에 적합한 정규화 방안이 무엇인지 규명한다.

2. 관련 연구

2.1 기존 유사도 계산 방안

논문들 간의 유사도 계산은 주로 논문을 노드로 논문들 간의 참조 정보를 링크로 간주하고 링크 기반 유사도 계산 방안을 이용했다. 기존 링크 기반 유사도 계산 방안으로는 Bibliographic Coupling[1], Co-Citation[2], Amsler[3], SimRank[4], rvs-SimRank[5], 그리고 P-Rank[5]가 있다. Bibliographic Coupling은 두 객체가 공통적으로 가리키는 객체들의 수를 이용하여 유사도 계산한다[1]. Co-Citation은 Bibliographic Coupling과 반대로 두 객체를 공통적으로 가리키는 객체들의 수를 이용하여 유사도 계산한다[2]. Amsler는 Bibliographic Coupling으로 계산된 유사도와 Co-Citation으로 계산된 유사도를 가중치 합해서 유사도를 계산한다[3].

Bibliographic Coupling, Co-Citation, 그리고 Amsler는 두 객체와 링크로 직접 연결되어 있는 객체들만을 이용하여 유사도를 계산한다. 반면 SimRank, rvs-SimRank, 그리고 P-Rank는 두 객체와 직접적으로 연결되어 있는 객체들뿐만 아니라 간접적으로 연결되어 있는 객체들도 이용하여 유사도를 계산한다.

2.2 정규화 방안

기존 링크 기반 유사도 계산 방안은 논문들이 가지고 있는 링크의 수가 많을수록 유사도가 증가하는 문제점이 있다. 따라서 이러한 문제를 해결하기 위해 유사도를 정규화 하는 과정이 필요하다. 대표적인 정규화 방안으로는 자카드 계수(Jaccard Coefficient)가 있다. 자카드 계수는 두 논문들이 공통적으로 참조하거나 두 논문들을 공통적으로 참조하는 논문들의 수를 두 논문이 참조하거나 두 논문들을 참조하는 모든 논문들의 수로 나누어주는 정규화 방안이다. 또 다른 정규화 방안으로는 두 논문들이 참조하거나 두 논문들을 참조하는 모든 논문들 간의 평균 유사도를 두 논문들이 참조하거나 두 논문들을 참조하는 모든 논문들의 쌍의 수로 나누어 주는 방안이다. 이러한 정규화 방안을 본 논문에서는 pairwise 정규화 방안으로 명명한다.

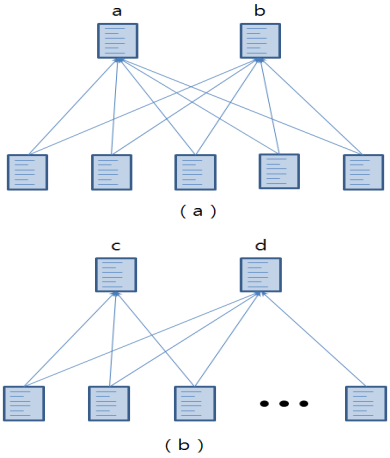
기존의 링크 기반 유사도 계산 방안 중 비재귀적 유사도 계산 방안들(Bibliographic Coupling, Co-Citation, 그리고 Amsler)은 두 객체가 공통적으로 가리키거나 두 객체를 공통적으로 가리키는 객체들의 수를 이용하여 유사도를 계산하기 때문에 주로 자카드 계수를 적용하였다. 그러나 재귀적 유사도 계산 방안들(SimRank, rvs-SimRank, 그리고 P-Rank)은 두 객체가 공통적으로 가리키거나 두 객체를 공통적으로 가리키는 객체들의 평균 유사도를 이용하여 유사도를 계산하기 때문에 주로 pairwise 정규화 방안을 적용하였다.

3. 논문 데이터베이스를 위한 정규화 방안

Pairwise 정규화 방안은 비교하고자 하는 두 논문이 참조하거나 참조 받는 논문들이 많으면 많을수록 두 논문의 유사도가 낮아지는 문제점을 가지고 있다. 논문 데이터베이스에는 일반적인 논문들에 비해서 다른 논문들에게 참조를 많이 받는 저명한 논문들이 존재한다. 저명한 논문일수록 논문의 질이 높기 때문에 논문 검색 서비스를 이용하는 사용자들은 유사한 논문들 중에서 저명한 논문들을 선호한다. 그러나 사용자가 관심 있는 논문과의 유사도 계산에서 일반적인 논문에 비해 참조를 많이 받는 저명한 논문과의 유사도는 낮게 계산된다. (그림 1)은 pairwise 정규화 방안으로 인해서 참조를 많이 받는 두 논문의 유사도가 낮게 계산되는 경우의 예이다.

(그림 1)에서 사각형은 논문을 나타내고 화살표는 참조 관계를 나타낸다. (그림 1)의 (a)에서 논문 a와 논문 b를 참조하고 있는 논문들의 수는 5개이며, (그림 1)의 (b)에서 논문 c와 논문 d를 참조하고 있는 논문들의 수는 100

개이다. 논문 *c*와 논문 *d*를 공통적으로 참조하는 논문들의 수가 많기 때문에 링크 기반 유사도 계산에서는 논문 *c*와 논문 *d*의 유사도가 높아야 한다. 그러나 pairwise 정규화 방안을 적용하여 유사도를 계산하면 논문 *a*와 논문 *b*의 유사도는 1/5(5/(5*5))인 반면, 논문 *c*와 논문 *d*의 유사도는 1/100(100/(100*100))이 된다. 즉, 공통적으로 참조를 많이 받는 논문 *c*와 논문 *d*의 유사도가 pairwise 정규화에 의해 논문 *a*와 논문 *b*의 유사도보다 낮게 계산된다. 물론, 재귀적으로 유사도를 계산할 때마다 유사도는 높아진다. 그러나 재귀적으로 계산할 때마다 동일한 문제가 발생하기 때문에 근본적으로 이러한 문제가 해결되지 않는다. 자카드 계수 정규화 방안을 적용할 경우 논문 *a*와 논문 *b*의 유사도는 1(5/5)이 되고, 논문 *c*과 논문 *d*의 유사도는 1(100/100)로 동일하게 적용되어 pairwise 정규화 방안의 문제점이 발생하지 않는다. 웹 그래프를 대상으로 하는 연구에서도 이와 유사한 결과를 도출한 적이 있다[6].



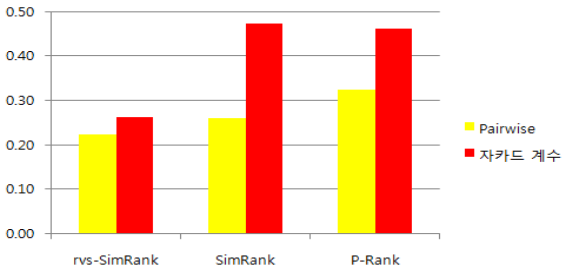
(그림 1) Pairwise 정규화 방안의 문제의 예

4. 실험

본 실험에서는 논문 데이터베이스에 기존 링크 기반 유사도 계산 방안들을 대상으로 pairwise 정규화 방안과 자카드 계수 정규화 방안을 적용하여 어느 정규화 방안이 논문 데이터베이스에 적합한지 실험을 통해 확인한다.

실험 데이터는 DBLP[1]에 있는 논문들을 사용했으며, 논문들 간의 참조 정보는 Libra[2]에서 크롤링하여 논문 데이터베이스를 구축하였다. 구축된 논문의 수는 총 448,000개 이며, 링크의 수는 126,281개 이다. 실험에 사용된 링크 기반 유사도 계산 방안들은 SimRank, rvs-SimRank, 그리고 P-Rank이며, 각 링크 기반 유사도 계산 방안에 자카드 계수 정규화 방안[6]과 pairwise 정규화 방안을 적용하여 정확도를 비교한다. 본 논문에서는 두 정규화 방안의 정확도를 측정하기 위해서 [7]에서 사용한 평가 방법을 사용한다.

(그림 2)는 자카드 계수를 적용한 유사도 계산 방안과 pairwise 정규화 방안을 적용한 유사도 계산 방안의 정확도 측정 결과이다. (그림 2)에서 알 수 있듯이 모든 링크 기반 유사도 계산 방안에서 자카드 계수를 적용한 방안이 pairwise를 적용한 방안 보다 정확도가 더 높음을 알 수 있다. 자카드 계수를 적용한 방안이 pairwise를 적용한 방안보다 rvs-SimRank의 경우 정확도가 약 17%, SimRank의 경우 약 81%, 그리고 P-Rank의 경우 약 42%로 높게 측정되었다.



(그림 2) 자카드 계수와 pairwise 정규화 방안의 정확도 비교

4. 결론

본 논문에서는 링크 기반 유사도 계산 방안에 사용되는 자카드 계수와 pairwise 정규화 방안을 설명하고 논문 데이터베이스에 적합한 정규화 방안이 자카드 계수라는 것을 설명하였다. 또한, 실제 논문 데이터베이스에서 기존 링크 기반 유사도 계산 방안들에 두 가지 정규화 방안을 적용하여 실험을 수행하였다. 실험 결과, 논문 데이터베이스에서 자카드 계수를 적용한 방안이 pairwise를 적용한 방안보다 rvs-SimRank의 경우 약 17%, SimRank의 경우 약 81%, 그리고 P-Rank의 경우 약 42% 정확도가 높게 측정되었다.

감사의 글

본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음. (NIPA-2010-(C1090-1011-0009))

참고문헌

[1] M. Kessler, "Bibliographic Coupling between Scientific Papers," *Journal of the American Documentation*, Vol. 14, No. 1, pp. 10-25, 1963.

[2] H. Small, "CoCitation in the Scientific Literature: A New Measure of the Relationship between Two Documents," *Journal of the American Society for Information Science*, Vol. 24, No. 4, pp. 265-269, 1973.

[3] R. Amsler, Application of Citation-Based Automatic Classification, Technical Report 72-14, *The University of Texas at Austin Linguistics Research Center*, 1972.

[4] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity," In *Proc. Int'l. Conf. on Special Interest Group on Knowledge Discovery and Data*, pp. 538-543, 2002.

[5] P. Zhao, J. Han, and Y. Sun, "P-Rank: a Comprehensive Structural Similarity Measure over Information Networks," In *Proc. of Int'l. Conf. on Information and Knowledge Management*, pp. 553-562, 2009.

[6] D. Fogaras and B. Ra'cz, "Scaling Link-Based Similarity Search," In *Proc. of Int'l. Conf. on World Wide Web*, pp. 641-650, 2005.

[7] 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*, pp. 1213-1214, 2010.

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