정보 네트워크를 위한 새로운 노드 유사도 계산 방안

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

A New Method for Computing Node Similarities in Information Networks

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 방안의 문제점을 제시하고, 이러한 문제를 해결하는 방안을 제안한다. 실험을 통하여 제안하는 방안과 기존 링크 기반 유사도 계산 방안의 정확도를 비교함으로써 제안하는 방안의 우수성을 검증한다.

1. 서론

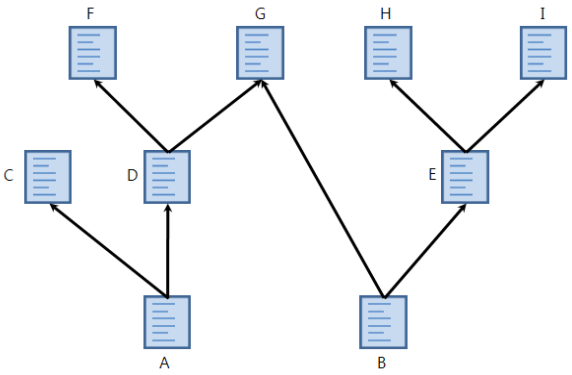
객체들 간의 유사도는 클러스터링, 추천, 분류, 그리고 랭킹 등 다양한 분야의 기반 정보로 제공될 수 있기 때문에 중요하다. 객체들 간의 유사도를 계산하는 대표적인 방안 중 하나는 객체들 간의 관계 정보를 이용하는 링크 기반 유사도 계산 방안이다. 본 논문에서는 기존 링크 기반 유사도 계산 방안의 문제점을 제시하고, 이러한 문제를 해결하는 방안을 제안한다.

2. 기존 연구의 문제점

기존 링크 기반 유사도 계산 방안은 유사도를 계산하는 두 객체가 공통으로 참조하는 객체가 많을수록 또는 두 객체를 공통으로 참조하는 객체가 많을수록 유사하다는 개념을 사용한다. 대표적인 방안으로는 rvs-SimRank[1], SimRank[2], 그리고 P-Rank[1]가 있다. Rvs-SimRank는 유사도를 계산하고자 하는 두 객체가 직접 또는 간접적으로 참조하는 객체들을 이용하여 유사도를 계산한다. SimRank는 유사도를 계산하고자 하는 두 객체를 직접 또는 간접적으로 참조하는 객체들을 이용하여 유사도를 계산한다. P-Rank는 rvs-SimRank의 유사도 계산 결과와 SimRank의 유사도 계산 결과의 가중치 합을 유사도로 사용한다.

그림 1은 객체들 간의 관계를 나타낸 예제로 사각형은 객체를 나타내고, 화살표는 객체들 간의 관계를 나타낸다. 그림 1에서 객체 A와 객체 B의 유사도를 rvs-SimRank로 계산할 경우 객체 A가 참조하는 객체 C와 D, 객체 B가 참조하는 객체 E와 G의 모든 객체 쌍 간에 유사도를 계산하여야 한다. 참조 객체 C와 D, 참조 객체 E와 G의 모든 객체 쌍의 유사도가 객체 A와 객체 B의 유사도를 결정한다. 그러나 참조 객체 C와 D, 참조 객체 E와 G의 모든 객체 쌍의 유사도가 계산되어 있지 않다. 참조 객체 C와 D, 참조 객체 E와 G의 모든 객체 쌍의 유사도를 계산하기 위해서는 각 객체 쌍에 해당하는 객체들 간의 유사도를 재계산하여야 한다. 객체 A의 참조 객체 C와 D, 객체 B의 참조 객체 E와 G는 서로 동일한 단계에 존재하는 객체이다. 그리고 객체 D의 참조 객체 F와 G, 객체 E의 참조 객체 H와 I는 서로 동일한 단계에 존재하는 객체이다.

[1]에서는 rvs-SimRank와 SimRank, 그리고 P-Rank를 수식 1과 같이 하나의 수식으로 나타낸다. k는 재귀적 단계를 의미한다. 수식 1에서 $k=\infty$일 때 λ=1이면 SimRank 수식으로 표현되고 λ=0이면 rvs-SimRank 수식으로 표현된다. P-Rank는 일반적으로 λ=0.5를 사용한다. 기존 링크 기반 유사도 계산 방안은 재귀적인 형태로 k단계에 존재하는 객체들 간의 유사도를 계산하여 k+1단계에 존재하는 객체들 간의 유사도를 계산하는데 이용된다.



(그림 1) 객체들 간의 관계를 나타낸 예제.

$$R_0(a,b) = \begin{cases} 0 & (\text{if } a \neq b) \\ 1 & (\text{if } a = b) \end{cases},$$

$$R_{k+1}(a,b) = \lambda \times \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} R_k(I_i(a), I_j(b)) + (1-\lambda) \times \frac{C}{|O(a)||O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} R_k(O_i(a), O_j(b)) \quad \text{(수식 1)}$$

기존 링크 기반 유사도 계산 방안은 동일한 단계에 존재하는 객체들 간의 유사도만을 계산하기 때문에 단계가 다른 객체들 간의 유사도는 고려하지 않는다. 예를 들어 객체 B는 객체 G를 직접적으로 참조하고 있으며, 객체 A는 객체 D를 통해 객체 G를 간접적으로 참조한다. 따라서 객체 G는 객체 A와 객체 B가 공통적으로 참조하는

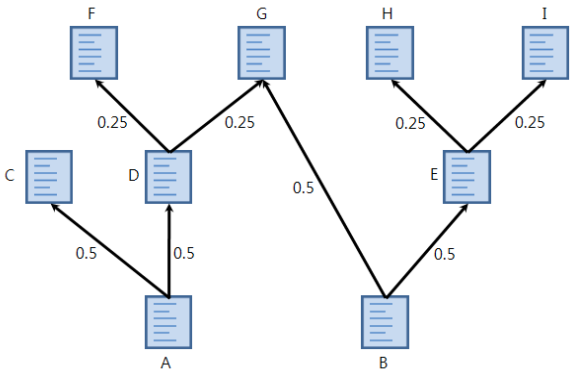
객체이다. 그러나 객체 B는 객체 D를 직접적으로 참조하지만 객체 A는 객체 D를 간접적으로 참조하는 객체이다. 따라서 객체 A와 객체 B의 유사도 계산에서 객체 G는 동일한 단계에 존재하는 객체가 아니기 때문에 객체 G를 공통으로 참조하는 객체라 판단하지 않는다. 이러한 문제는 유사도를 계산하고자 하는 두 객체가 유사한 객체를 참조하더라도 유사도 계산에 이용할 수 없기 때문에 유사도 계산의 정확도에 영향을 미친다.

3. 제안하는 방안

본 논문에서는 유사도를 계산하고자 하는 두 객체가 k 단계까지 참조하는 모든 객체들을 이용하여 유사도를 계산하는 방안을 제안한다. 제안하는 방안은 해당 객체와 참조 객체의 거리가 가까울수록 유사하다는 개념을 이용하며, 해당 객체와 참조 객체들 간의 도달 확률을 가중치로 부여한다.

그림 2는 참조 객체들에 가중치를 설정한 예제이다. 그림 2에서 사각형은 객체를 나타내고, 화살표는 객체들 간의 관계를 나타낸다. 각 링크에 설정된 가중치는 해당 객체에서 참조 객체로 도달 가능 확률을 의미하며, 도달 가능 확률은 링크의 수에 의해 균등하게 분배된다. 예를 들어, 객체 A에서 객체 C와 객체 D에 도달 가능 확률은 균등하게 각 0.5로 설정된다. 객체 A에서 객체 F 또는 객체 G에 도달 가능 확률은 객체 D를 경유하기 때문에 객체 D의 도달 가능 확률을 다시 균등하게 분배하여 각 0.25로 설정된다. 제안하는 방안의 유사도 계산은 유사도를 계산하고자 하는 두 객체의 k단계까지 존재하는 모든 객체들 중 공통적으로 존재하는 객체들의 가중치 곱으로 계산된다. 그림 2에서 객체 A와 객체 B의 유사도는 객체 G를 공통으로 참조하고 있기 때문에 객체 A에서 객체 G로 도달 가능 확률 0.25와 객체 B에서 객체 G로 도달 가능 확률 0.5의 곱인 0.115이다. 하나 이상의 공통 객체가 존재할 경우 각 공통 객체에 대한 유사도 합으로 계산된다.

링크 기반 유사도 계산 방안은 이론적으로 k=∞까지 이용하여 유사도를 계산하지만 일반적으로 작은 임의의 상수 단계에서 대부분 유사도가 수렴하게 된다[1]. 따라서 본 논문에서는 k=2로 설정하여 이용한다.



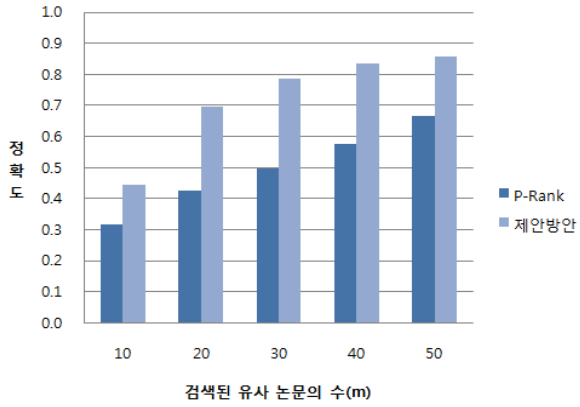
(그림 2) 제안하는 방안의 가중치 설정 예제.

4. 실험

본 논문에서는 기존 링크 기반 유사도 계산 방안과 제안하는 방안의 정확도를 비교하여 제안하는 방안의 우수성을 검증한다.

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 방안은 P-Rank를 사용하였으며, P-Rank와 제안하는 방안의 정확도를 측정하기 위해 [3]에서 사용한 평가 방법을 사용한다.

그림 3은 P-Rank와 제안하는 방안의 정확도 측정 결과이다. 그림 3에서 알 수 있듯이 제안하는 방안이 P-Rank보다 모든 구간(m)에서 정확도가 향상되었으며, 최대 약 60% 정확도가 향상되었다. 이는 제안하는 방안이 동일한 단계에 존재하는 객체들 간의 유사도뿐만 아니라 단계가 다른 객체들 간의 유사도를 고려하기 때문에 기존 링크 기반 유사도 계산 방안에 비해 객체들 간의 유사도를 더 정확하게 계산할 수 있었다.



(그림 3) P-Rank와 제안 방안의 정확도 비교.

5. 결론

본 논문에서는 링크 기반 유사도 계산 방안의 문제점을 제시하고 이를 해결하여 유사도 계산 결과의 정확도를 향상하는 방안을 제안하였다. 제안하는 방안이 유사도 계산 결과의 정확도를 향상하였음을 검증하기 위해 실제 논문 데이터베이스를 대상으로 실험을 통해 정확도를 비교하였다. 실험 결과 제안하는 방안이 링크 기반 유사도 계산 방안에 비해 최대 약 60% 정확도가 향상되었다.

감사의 글

본 연구는 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단 (No.2008-0061006) 및 지식경제부 및 정보통신산업진흥원의 'IT융합 고급인력과정 지원사업' (NIPA-2011-C6150-1101-0001)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] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity," In *Proc. of Int'l. Conf.* on Special Interest Group on Knowledge Discovery and Data, pp. 538-543, 2002.

[2] 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.

[3] 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/