http://dx.doi.org/10.3745/KIPSTD.2011.18D.5.317

논문 데이터베이스를 위한 텍스트 기반 유사도 계산 방안

윤 석 호[†] · 김 상 욱[††]

요 약

본 논문에서는 기존 텍스트 기반 유사도 계산 방안을 이용해서 논문들 간의 유사도를 계산하는 방안에 대해서 논의한다. 먼저, 실험을 통해서 논문의 제목, 요약, 그리고 본문 중에서 어떤 부분이 유사도를 계산하는데 더 유용한지 확인하고 적절한 가중치를 부여한다. 두 번째로 논문의 텍스트 정보가 불완전한 상황에서 논문들 간의 유사도를 보다 정확하게 계산할 수 있는 키워드 확장 방안을 제안한다. 실제 논문 데이터베이스를 이용해서 제안하는 방안의 우수성을 검증한다.

키워드 : 텍스트 기반 유사도, 키워드 확장, 논문 데이터베이스

A Text-based Similarity Measure for Scientific Literature

Seok-Ho Yoon[†] · Sang-Wook Kim[††]

ABSTRACT

This paper addresses computing of similarity among papers using text-based measures. First, we analyze the accuracy of the similarities computed using different parts of a paper, and propose a method of *Keyword-Extension*, which is very useful when text information is incomplete. Via a series of experiments, we verify the effectiveness of *Keyword-Extension*.

Keywords : Text-based Similarity Measure, Keyword-extension, Scientific Literature

1. 서 론

최근 들어, 학술 정보에 대한 사용자들의 관심이 증가하면서 학술 정보를 대상으로 하는 많은 연구가 진행되고 있다[1][2]. 학술 정보를 대상으로 하는 연구들 중 대표적인 연구 중 하나는 논문들 간의 유사도를 계산하는 것이다. 논문들 간의 유사도는 학술 정보를 대상으로 하는 클러스터링, 분류, 추천, 그리고 랭킹 등의 기술에 기반 정보로 제공될 수 있기 때문에 중요하다[3].

논문과 같은 문서에 대한 유사도를 계산하는 기존 유사도 계산 방안들은 크게 텍스트 기반 유사도 계산 방안과 링크 기반 유사도 계산 방안으로 분류될 수 있다[4]. 본 논문에서는 텍스트 기반 유사도 계산 방안을 이용해서 논문들 간의 유사도를 계산하고자 한다.

※ 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것입니다(2008-0061006). 또한, 지식경제부 및 정보통신산업진흥원의 IT융합 고급인력과정 지원사업(NIPA-2011-C6150-1101-3001)과 두뇌한국 21 사업의 부분적인 지원을 받아 수행됨.
† 준 회 원 : 한양대학교 전자컴퓨터통신공학과 박사과정
†† 종신회원 : 한양대학교 정보통신대학 정보통신학부 교수
논문접수 : 2011년 6월 15일
수 정 일 : 1차 2011년 8월 2일
심사완료 : 2011년 8월 9일

논문은 제목, 요약, 그리고 본문 세 개의 부분으로 구성되어 있다. 각 부분에 포함된 단어의 종류와 수가 서로 상이하기 때문에 각 부분을 이용해서 논문들 간의 유사도를 계산하게 되면 계산된 유사도가 서로 다르다. 따라서 세 부분 중에서 어떤 부분이 유사도 계산에 유용한지 확인하고 각 부분에 적절한 가중치를 부여해야한다.

논문 데이터베이스를 제공하는 대표적인 웹 사이트인 CiteSeer, Google Scholar, and MS Libra은 논문의 텍스트 정보를 웹 크롤링과 논문 파일의 파싱을 통해서 수집한다. 그러나 논문의 본문은 저작권 문제로 인해서 텍스트의 형태가 아닌 일반적으로 파일을 형태로 사용자에게 제공한다. 또한, 논문 데이터베이스에 저장된 논문의 요약 정보는 크롤링과 파싱의 기술적 한계로 인해서 원래 논문의 요약에 포함된 단어들 중 일부가 소실되어 저장되어 있지 않다. 결국 실제적으로 논문의 텍스트 정보는 불완전한 상태이다. 따라서 이러한 텍스트 정보를 이용해서 논문들 간의 유사도를 계산할 경우 유사도 계산의 정확도가 낮아질 가능성이 높다.

본 논문에서는 실험을 통해서 논문의 세 부분 중에서 어떤 부분이 유용한지 확인하고 각 부분에 적절한 가중치를

부여한다. 또한 논문의 텍스트 정보가 완전하지 않은 상황에서 논문들 간의 유사도를 정확하게 계산할 수 있는 키워드 확장 방안을 제안한다.

본 논문에서는 실제 논문 데이터베이스를 이용해서 다음과 같은 실험을 수행했다. 먼저, 각 부분을 이용해서 논문들 간의 유사도를 계산하고 [4]에서 사용한 평가 방법을 이용해서 해당 유사도의 정확도를 측정한다. 측정 결과 요약을 이용한 논문들 간의 유사도 계산 결과의 정확도가 가장 높았다. 또한, 두 부분 이상을 동시에 이용하였을 때 제목과 요약의 가중치가 0.3:0.7일 때, 가장 정확도가 높았다.

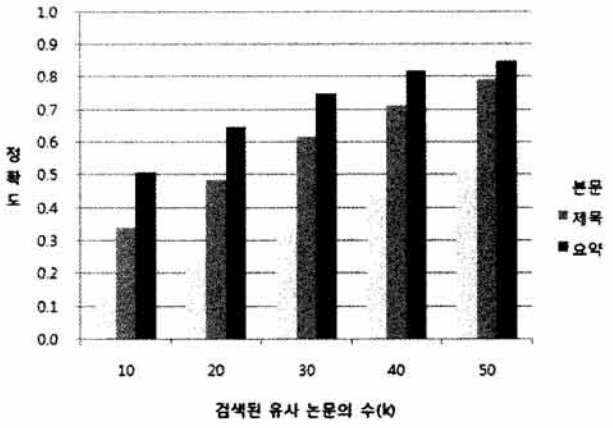
일반적으로 논문의 저자는 해당 논문과 유사한 주제의 논문을 참조한다. 따라서 해당 논문이 참조하는 논문들은 해당 논문의 주제와 관련이 높은 단어들을 가지고 있다. 본 논문에서는 논문이 참조하는 논문들과 해당 논문을 참조하는 논문들의 단어들을 해당 논문의 단어들로 이용하는 키워드 확장 방안을 제안한다. 다양한 실험을 통하여 키워드 확장 방안이 논문 데이터가 불완전한 상황에서도 논문들 간의 유사도를 정확하게 계산할 수 있다는 것을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 실험을 통하여 논문의 세 부분의 가중치를 설정한다. 3장에서는 본 논문에서 제안하는 키워드 확장 방안을 설명한다. 4장에서는 실험을 통하여 제안하는 방안의 우수성을 검증한다. 끝으로 5장에서는 본 논문의 내용을 요약한다.

2. 가중치 설정

본 절에서는 실험을 통해서 논문의 각 부분을 이용해서 계산한 유사도의 정확도를 정량적으로 평가한다. 본 실험에서는 DBLP[1]에 있는 논문들을 사용했으며 논문들 간의 참조 정보는 Libra[2]에서 크롤링해서 사용하였다. 해당 데이터의 논문의 수는 1,071,973 편이고 참조 정보의 수는 2,473,636개이다. 실험에서 사용하는 텍스트 기반 유사도 계산 방안은 가장 널리 알려진 TF·IDF를 이용한 cosine similarity이다[3][5]. 본 실험에서 사용하는 정확도 측정 방안은 다음과 같다. 먼저, 유명한 데이터 마이닝 교재[3]에서 21개의 소챕터를 선택하고 해당 챕터에서 모든 참고 논문을 추출한다. 각 소챕터 내의 각 참고 논문을 질의 논문을 차례대로 선택하고 해당 논문과 가장 유사한 k개의 논문을 각 부분을 통해서 계산된 유사도를 이용해서 찾는다. 그런 후에 찾은 k개의 논문들이 질의 논문이 포함되어 있는 소챕터에 얼마나 많이 포함되어 있는지 계산한다. 소챕터 내에 있는 모든 논문이 질의 논문이 될 때까지 위 과정을 반복한다. 이 계산 방안은 응답도 계산 방법과 유사하다[4].

(그림 1)은 각 부분을 통해서 계산한 유사도의 정확도 측정 결과를 나타낸다. 요약을 이용해서 계산한 유사도가 가장 정확했고 제목을 이용해서 계산한 유사도가 본문을 이용



(그림 1) 각 부분을 이용한 유사도 계산 결과의 정확도

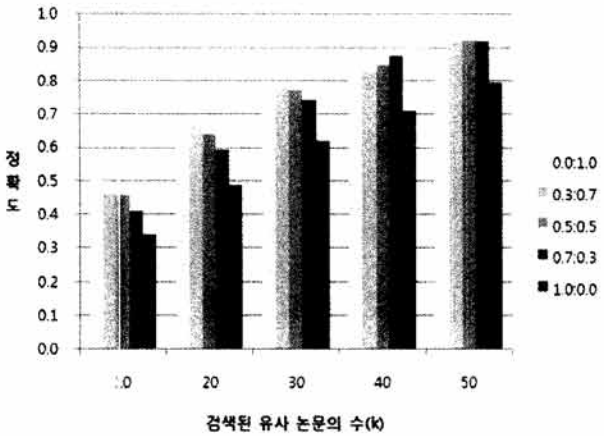
해서 계산한 유사도보다 정확했다. 이는 본문이 많은 단어들을 가지고 있지만 해당 논문의 주제와 관련이 없는 단어들도 많이 포함되어 있기 때문이다. 제목은 비록 해당 논문의 주제와 관련된 단어들을 주로 가지고 있지만 단어의 수가 너무 적다. 요약은 제목보다 많은 수의 단어들을 가지고 있으며 본문이 가지고 있는 단어들보다 해당 논문의 주제와 관련성이 높은 단어들을 가지고 있다. 결국, 논문의 주제와 관련이 있는 단어들을 충분히 가지고 있어야 논문들 간의 유사도를 정확하게 계산할 수 있다.

위의 실험 결과가 오직 요약만을 이용해서 논문들 간의 유사도를 계산하는 것이 가장 정확하다는 것을 나타내지 않는다. 따라서 두 부분 이상을 동시에 이용해서 유사도를 계산하고 계산된 유사도의 정확도를 판정하고자 한다. 그러나 본문은 세 부분 중에서 유사도 계산 결과의 정확도가 가장 낮았고 본문에서 논문의 주제와 관련성이 적은 단어를 제거하는 것이 어렵다. 또한 서론에서 설명하였듯이 논문 데이터베이스에는 본문의 텍스트 정보를 제공하지 않기 때문에 본 논문에서는 본문을 제외한 제목과 요약의 단어들만을 대상으로 가중치를 적절하게 부여해서 논문들 간의 유사도를 계산하고자 한다.

실험 방법은 논문의 제목과 요약의 단어들에 부여된 가중치를 0.0:1.0, 0.3:0.7, 0.5:0.5, 0.7:0.3, 그리고 1.0:0.0으로 변경하면서 논문들 간의 유사도를 계산하고 계산된 결과의 정확도를 측정한다. 가중치 설정이 1.0:0.0일 경우 제목의 단어들의 가중치가 1.0으로 제목의 단어들만을 이용해서 유사도를 계산한 것을 의미하며 0.0:1.0인 경우 반대로 요약의 단어들만을 이용해서 유사도를 계산한 것을 의미한다.

(그림 2)은 제목과 요약의 가중치 변화에 따른 유사도 계산 결과의 정확도를 나타낸다. (그림 2)에서 알 수 있듯이 제목의 단어 가중치를 0.3으로 요약의 단어 가중치를 0.7로 부여했을 때 유사도 계산 결과의 정확도가 전반적으로 높았다. 제목은 단어 수가 적지만 해당 논문을 대표하는 단어들로 구성되어 있기 때문에 논문들 간의 유사도 계산의 정확도를 높이는데 기여한다. 또한, 요약의 텍스트 정보가 없는

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



(그림 2) 제목과 요약의 가중치 변화에 따른 유사도 계산 결과의 정확도

유사한 논문들 간의 유사도를 요약만을 이용하는 경우 유사도가 0으로 계산되지만, 제목을 동시에 이용하는 경우 어느 정도 적절한 유사도를 계산할 수 있다. 따라서 제목과 요약의 단어들을 함께 이용하여 유사도를 계산하는 경우가 가장 정확하다고 할 수 있다. 본 논문의 이후 실험에서는 제목의 단어들의 가중치를 0.3으로 요약의 단어들의 가중치를 0.7로 부여한다.

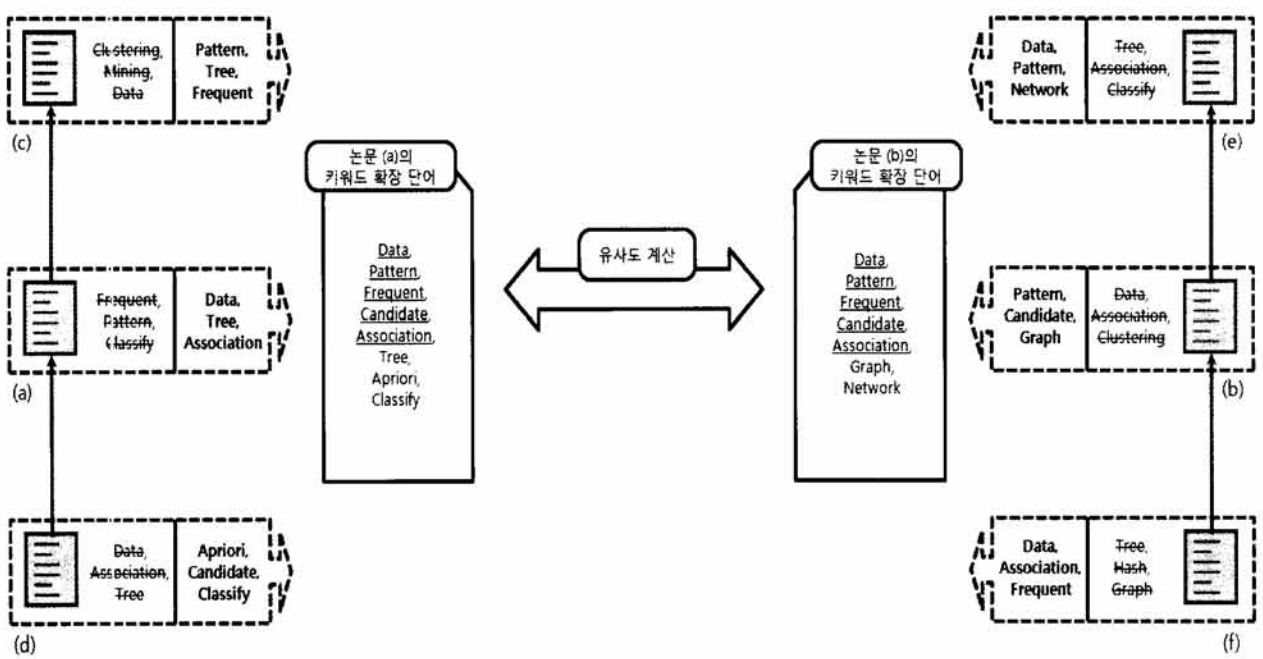
3. 키워드 확장 방안

2장에서는 요약에 있는 단어들이 유사도를 계산하는데 유용하다는 것을 실험을 통해서 보였다. 그러나 크롤링과 파싱의 기술적 한계로 인해서 실제 논문 데이터베이스에 저장된 요약의 텍스트 정보가 완전하지 않은 경우가 많다. 따라서 불완전한 요약의 텍스트 정보를 이용해서 논문들 간의 유사도를 계산하게 되면 계산된 유사도의 정확도가 낮아지게 된다. 따라서 소실된 단어들을 보충할 수 있는 방안이 필요하다.

논문의 저자는 자신이 작성한 논문의 주제와 관련이 있는 논문들을 참조한다. 따라서 두 논문이 참조 관계일 때 두 논문이 가지고 있는 단어들이 유사할 가능성이 높다. 본 논문에서는 이러한 관찰을 바탕으로 논문 P가 참조하는 논문들의 단어들과 논문 P를 참조하는 논문들의 단어들을 논문 P의 단어들로 이용하는 키워드 확장 방안을 제안한다. 예를 들어, 논문 B가 C를 참조하고 논문 A가 논문 B를 참조하면 논문 A와 C의 단어들을 논문 B에 단어 집합에 포함시킨 후에 해당 단어들을 모두 이용해서 논문 B와 다른 논문들 간의 유사도를 계산한다.

(그림 3)는 논문의 참조 정보를 이용하여 유사도를 계산하는 키워드 확장 방안을 그림으로 나타낸 것이다. 사각형은 논문을 의미하며, 점선 사각형 내 존재하는 단어들은 해당 논문을 설명하는 단어를 의미한다. 굵은 글씨의 단어들은 논문 데이터베이스 존재하는 단어들이며, 중앙에 밑줄 친 얇은 글씨의 단어들은 소실된 단어들이다. 밑줄 친 단어들은 제안하는 방안을 통해 두 논문이 공통적으로 가지고 있는 단어들을 의미한다. 화살표는 논문들의 참조 관계를 나타낸다.

논문 (a)와 논문 (b)의 유사도를 계산하는 경우, 단어의 소실로 인해 공통 단어가 존재하지 않아 두 논문의 유사도는 낮게 계산된다. 그러나 논문 (a)과 논문 (b)의 단어가 소실되지 않았을 경우 여러 공통 단어들이 존재하기 때문에 두 논문의 유사도는 높게 계산된다. 키워드 확장 방안은 해



(그림 3) 참조 관계에 있는 논문의 텍스트 정보를 이용한 키워드 확장 방안

당 논문이 참조하는 논문과 해당 논문을 참조하는 논문의 단어들을 이용하여 논문들 간의 유사도를 계산한다. 예를 들어, 논문 (a)의 참조 논문들은 논문 (c)과 논문 (d)이며, 논문 (b)의 참조 논문들은 논문(e)과 논문 (f)이다. 논문 (a)과 논문 (b)의 유사도 계산 시 논문 (a)는 논문 (c)와 논문 (d)의 단어를 같이 이용하며, 논문 (b)는 논문 (e)와 논문 (f)의 단어들을 같이 이용한다. 따라서 논문 (a)와 논문 (b)의 유사도가 논문 (a)와 논문 (b)가 완전한 텍스트 정보를 가지고 있을 때의 유사도에 근접하게 계산될 수 있다.

4. 실 험

4.1 정성적 방법을 통한 키워드 확장 방안의 성능 평가

본 실험에서는 기존 텍스트 유사도 계산 방안과 키워드 확장 방안을 이용한 텍스트 기반 유사도 계산 방안을 이용해서 주어진 질의 논문과 가장 유사한 논문 5편을 각각 추출하고, 추출된 논문들이 주어진 질의 논문과 주제가 유사한지 살펴보고자 한다. 질의 논문은 데이터베이스 분야에서 저명한 [6] 그리고 [7]을 선정했다.

〈표 1〉 각 방안을 이용해서 추출한 질의 논문 [6]과 유사한 상위 5편의 논문들

	기존 방안		키워드 확장 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 2〉 각 방안을 이용해서 질의 논문 [7]과 유사한 상위 5편의 논문들

	기존 방안		키워드 확장 방안	
순위	논문 제목	유사도	논문 제목	유사도
1	Efficient Search in Very Large Databases	0.347	A Distribution Based Clustering Algorithm for Mining in Large Spatial Databases	0.862
2	Very Large Databases How Large How Different	0.345	An Efficient Approach to Clustering in Large Multimedia Databases with Noise	0.824
3	Constraint based clustering in large databases	0.339	CURE An Efficient Clustering Algorithm for Large Databases	0.822
4	An Efficient Cell Based Clustering Method for Handling Large High Dimensional Data	0.330	WaveCluster A Multi Resolution Clustering Approach for Very Large Spatial Databases	0.816
5	A two way visualization method for clustered data	0.324	A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise	0.806

<표 1>과 <표 2>는 각 방안을 이용해서 주어진 각각의 질의 논문과 유사한 상위 5개의 논문들을 나타낸다. <표 1>에서 기존 방안으로 추출한 논문들은 질의 논문의 제목에 있는 'diffusion'이라는 단어가 공통적으로 들어가 있다. 그러나 동일한 단어가 들어가 있을 뿐 질의 논문의 주제와는 다소 거리가 있는 논문들이다. 제안하는 방안으로 추출한 논문들은 비록 'diffusion'라는 단어를 제목에 포함하고 있지는 않지만 질의 논문과 유사한 주제인 블로그와 사회연결망과 관련된 논문들이다. <표 2>에서는 제안하는 방안으로 추출한 논문들 모두가 질의 논문과 같은 클러스터링에 관한 내용인 반면에 기존 방안으로 추출한 논문들 중 질의 논문과 첫 번째, 두 번째로 유사하다고 판단된 논문들이 클러스터링과 관련이 오히려 적었다.

기존 방안은 해당 논문을 설명하는 키워드의 수가 너무 적기 때문에 두 논문이 가지고 있는 키워들 중에 한 두 개의 단어만 일치해도 유사도가 높게 계산될 수 있다. 따라서 논문들 간의 주제가 유사하지 않는 경우에도 유사도가 높게 계산되는 문제가 발생한다. 그러나 키워드 확장 방안을 이용하면 해당 논문을 설명하는 단어의 수가 충분하기 때문에 기존 방안보다 유사도 계산 결과가 정확하다.

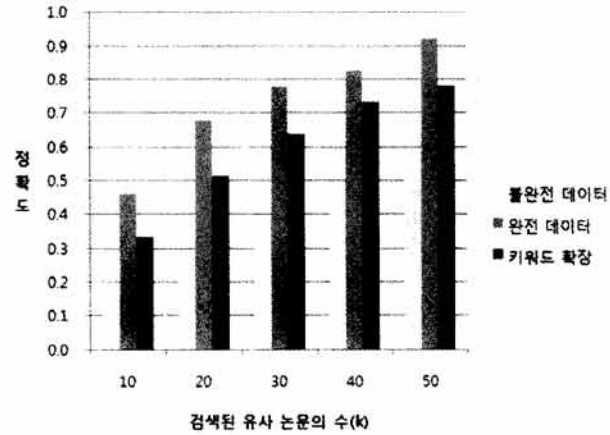
4.2 정량적 방법을 통한 키워드 확장 방안의 성능 평가

본 실험에서는 키워드 확장 방안의 성능을 정량적으로 평가하기 위해서 2장에서 사용한 평가 방법을 이용한다. 실험 방법은 키워드 확장 방안을 이용한 텍스트 기반 유사도 계산 방안과 키워드 확장 방안을 이용하지 않은 텍스트 기반 유사도 계산 방안, 그리고 요약 정보가 완전한 상태에서의 텍스트 기반 유사도 계산 방안의 정확도를 비교한다. 키워드 확장 방안에서 해당 논문, 해당 논문이 참조하는 논문, 그리고 해당 논문을 참조하는 논문들의 가중치는 1:1:1로 설정한다.

(그림 4)는 해당 실험의 결과를 나타낸다. '불완전 데이터'는 불완전 데이터를 가지고 기존 텍스트 기반 유사도 계산 방안으로 유사도를 계산한 결과의 정확도이고 '완전 데이터'는 본래로 완전한 데이터를 가지고 유사도를 계산할 결과이다 '키워드 확장'은 키워드 확장 방안을 이용해서 텍스트 기반 유사도 계산 방안으로 유사도를 계산한 결과의 정확도이다.

완전한 요약 정보를 이용한 텍스트 기반 유사도 계산 방안보다는 키워드 확장 방안을 이용한 텍스트 기반 유사도의 정확도가 낮았다. 그러나 키워드 확장 방안을 이용한 텍스트 기반 유사도 계산 방안의 결과가 키워드 확장 방안을 이용하지 않은 텍스트 기반 유사도 계산 방안의 결과보다 정확도가 3.3배 높았다. 이는 키워드 확장 방안이 불완전한 논문 데이터베이스에 논문들 간의 유사도를 정확하게 계산하는데 유용하다는 것을 나타낸다.

5. 결 론



(그림 4) 키워드 확장 방안과 키워드 확장 방안을 이용하지 않는 두 방안의 정확도 비교

본 논문에서는 논문의 세 부분 중에서 어떤 부분이 논문들 간의 유사도를 정확하게 계산하는데 유용한지 실험을 통해서 확인하고 각 부분에 적절한 가중치를 부여하였다. 또한, 논문의 텍스트 정보가 완전하지 않을 때 논문들 간의 유사도를 정확하게 계산할 수 있는 키워드 확장 방안을 제안하였다. 본 논문에서는 실제 논문 데이터를 이용해서 제안하는 키워드 확장 방안의 유용성을 검증하였다. 실험 결과 키워드 확장 방안이 논문들 간의 유사도를 정확하게 계산하는데 유용했다.

참 고 문 헌

[1] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and Mining of Academic Social Networks," In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp.990-998, 2008.

[2] X. Liu, S. Yu, Y. Moreau, B. Moor, and W. Glanzel, "Hybrid Clustering of Text Mining and Bibliometrics Applied to Journal Sets," In *Proc. of SIAM Int'l Conf. on Data Mining*, pp.49-60, 2009.

[3] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[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*, pp.1213-1214, April, 2010.

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

[6] 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.

[7] T. Zhang, R. Ramakrishnam, and M. Livny, "BIRCH: an Efficient Data Clustering Method for Very Large Databases," In *Proc. Int'l. Conf. on Management of Data*, pp.103-114, 1996.

윤 석 호

e-mail : bogely@agape.hanyang.ac.kr
2005년 성결대학교 컴퓨터공학과(학사)
2007년 한양대학교 전자통신컴퓨터공학과 (공학석사)
2007년~현 재 한양대학교 전자컴퓨터 통신공학과 박사과정
관심분야 : 사회연결망분석, 인터넷 포탈 데이타 분석, e-비지니스, 데이타 마이닝

김 상 욱

e-mail : wook@agape.hanyang.ac.kr
1989년 서울대학교 컴퓨터공학과(학사)
1991년 한국과학기술원 전산학과(석사)
1994년 한국과학기술원 전산학과(박사)
1991년 7월~1991년 8월 미국 Stanford University, Computer Science Department, 방문 연구원
1994년 3월~1995년 2월 KAIST 정보전자연구소 전문 연구원
1999년 8월~2000년 8월 미국 IBM T.J. Watson Research Center, Post-Doc.
1995년 3월~2003년 2월 강원대학교 정보통신공학과 부교수
2009년 1월~2010년 2월 미국 Carnegie Mellon University, Visiting Scholar
2003년 3월~현 재 한양대학교 정보통신대학 정보통신학부 교수
관심분야 : 데이터베이스 시스템, 저장 시스템, 트랜잭션 관리, 데이터 마이닝, 멀티미디어 정보 검색, 공간 데이터베이스/GIS, 주기억장치 데이터베이스, 이동 객체 데이터베이스/텔레매틱스, 사회 연결망 분석, 웹 데이터 분석