ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 43, No. 9, pp. 1043-1051, 2016. 9
http://dx.doi.org/10.5626/JOK.2016.43.9.1043

개체 중의성 해소를 위한 사용자 유사도 기반의 트윗 개체 링킹 기법

(Tweet Entity Linking Method based on User Similarity for Entity Disambiguation)

김 서 현 † 서 영 덕 † 백 두 권 ††
(SeoHyun Kim) (YoungDuk Seo) (Doo-Kwon Baik)

요 약 트위터 문서는 웹 문서에 비해 길이가 짧기 때문에 웹 기반의 개체 링킹 기법을 그대로 적용시킬 수 없어 사용자 정보나 집단의 정보를 활용하는 방법들이 시도되고 있다. 하지만, 트윗의 개수가 충분하지 않은 사용자의 경우 데이터 희소성 문제가 여전히 발생하고 관련이 없는 집단의 정보를 사용할 경우 링킹의 결과에 악영향을 미칠 수 있다. 본 논문에서는 기존 연구의 문제를 해결하기 위해 단일 트윗 내의 의미 관련도 뿐만 아니라 사용자의 트윗 집합과 다른 사용자들의 트윗 집합까지 고려하여 데이터 희소성을 해결하고, 관련성이 높은 사용자들의 트윗 정보에 가중치를 주어 트윗 개체 링킹의 성능을 높이고자 한다. 실제 트위터 데이터를 활용한 실험을 통해 제안하는 트윗 개체 링킹 기법이 기존의 기법에 비해 높은 성능을 가지며, 유사도가 높은 사용자의 정보를 사용하는 것이 트윗 개체 링킹에서 데이터 희소성 해결과 링킹 정확도 향상에 연관성이 있음을 보였다.

키워드: 소셜 네트워크 서비스, 마이크로블로그, 트윗 개체 링킹, 개체 링킹, 개체 중의성 해소, 사용자 유사도

Abstract Web based entity linking cannot be applied in tweet entity linking because twitter documents are shorter in comparison to web documents. Therefore, tweet entity linking uses the information of users or groups. However, data sparseness problem is occurred due to the users with the inadequate number of twitter experience data; in addition, a negative impact on the accuracy of the linking result for users is possible when using the information of unrelated groups. To solve the data sparseness problem, we consider three features including the meanings from single tweets, the users' own tweet set and the sets of other users' tweets. Furthermore, we improve the performance and the accuracy of the tweet entity linking by assigning a weight to the information of users with a high similarity. Through a comparative experiment using actual twitter data, we verify that the proposed tweet entity linking has higher performance and accuracy than existing methods, and has a correlation with solving the data sparseness problem and improved linking accuracy for use of information of high similarity users.

Keywords: social network service, microblog, tweet entity linking, entity linking, entity disambiguation, user similarity

· 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단-차세대정보·컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No.2012M3C4A7033346)

† 학생회원 : 고려대학교 컴퓨터학과
tjgus3253@korea.ac.kr
seoyoungd@korea.ac.kr
†† 종신회원 : 고려대학교 컴퓨터학과 교수(Korea Univ.)
baikdk@korea.ac.kr
(Corresponding author임)

논문접수 : 2016년 6월 7일
(Received 7 June 2016)
논문수정 : 2016년 7월 5일
(Revised 5 July 2016)
심사완료 : 2016년 7월 6일
(Accepted 6 July 2016)

정보과학회논문지 제43권 제9호(2016. 9)

1. 서 론

트위터는 사용자 자신의 관심사나 경험을 실시간으로 다른 사람과 공유할 수 있는 마이크로블로그 소셜 네트워크 서비스이다. 트위터상에서 사용자들이 작성한 문서를 분석하는 작업을 통해 사용자의 관심사를 추출하거나 사용자에 관한 중요한 정보 자원을 얻을 수 있다. 이때 대량의 데이터를 모두 사람들이 직접 읽고 분석하기에는 한계가 있으므로 기계가 자동으로 문서를 해독하고 사용자의 관심사를 추출하는 방법이 필요하다. 이를 위한 방법으로 개체 링킹이 사용되며, 트윗의 개체를 파악하여 지식 베이스 항목과 연결해주는 작업을 트윗 개체 링킹(tweet entity linking)이라 한다.

개체 링킹(entity linking)은 구조화되지 않은 텍스트와 구조화된 지식 베이스 사이를 연결하여 기계가 코퍼스 상의 개체의 의미를 찾아낼 수 있도록 하는 작업이다. 개체 링킹에서는 동일한 개체표현(named entity mention)을 공유하는 서로 다른 개체가 두 개 이상 존재할 수 있어 개체표현의 중의성을 해소하는 절차(entity disambiguation)가 필수적으로 요구된다[1]. 따라서 개체 링킹에 대한 연구들은 지식 베이스의 지식과 의미적인 관련도를 통해 중의성을 해소하는 데 중점을 두고 있고, 대다수의 연구는 웹 문서를 기반으로 한다[2-5]. 트위터 문서는 웹 문서에 비해 길이가 짧고 노이즈가 많아 의미 있는 개체가 부족하기 때문에 웹 문서 기반의 개체 링킹 기법을 트위터 문서에 그대로 적용할 경우 링킹의 성능이 떨어지는 문제가 발생한다. 예를 들어, 트위터에서는 약어를 많이 사용하기 때문에 중의성이 더욱 빈번하게 발생하고, 비공식적인 언어와 신조어 등의 사용은 지식 베이스의 항목과 연결되지 않는 개체가 많이 발생하는 결과를 초래한다. 또한, 문서를 작성할 때에는 140자 이내로 적어야 하기 때문에 개체 의미 파악에 필요한 데이터가 희소해지는 문제가 발생하게 되며, 이는 결과적으로 개체 링킹 정확도를 떨어뜨리는 요인이 된다.

트윗 개체 링킹 연구에서는 이러한 트위터 문서의 특성을 고려하고자 사용자의 정보와 집단의 정보를 활용하여 데이터 희소성(data sparseness)으로 인한 문제를 해결하고자 한다. 사용자의 정보를 활용한 방법은 사용자가 작성했던 트윗의 기록을 통해 사용자의 관심사를 반영하여 데이터 희소성 문제를 해결한다[6,7]. 사용자가 직접 작성했던 트윗 집합은 사용자의 관심사를 파악하는 데 중요한 정보가 될 수 있으며, 링킹 정확도 향상에 도움이 된다. 하지만 사용자가 작성했던 트윗이 없거나 트윗의 개수가 적은 경우에는 데이터 희소성 문제가 여전히 발생하는 단점이 있다. 사용자 개인의 트윗 집합 크기에 영향을 받는 문제를 보완하고자 집단의 정보를 이용한 기법이 시도되고 있다[8]. 이러한 방식은 사용자 개인이 가지는 트윗 데이터 집합의 크기와 상관없이 집단의 트윗 데이터가 있다면 데이터 희소성 문제를 해결할 수 있다. 기존의 집단 정보를 이용한 트윗 개체 링킹 연구들은 단일 트윗을 하나의 문서로 보고, 문서 간의 문맥적인 관련성을 통해 사용자 간 유사성을 측정한다. 하지만 트위터의 특성상 단일 트윗에 포함되는 단어의 수가 적으므로 트윗 문서 내의 개체들만을 비교하여 사용자들이 작성한 트윗 집합의 유사성을 정확하게 판단하기는 어려우며, 관련성이 낮은 사용자들의 트윗을 사용하게 되어 링킹의 정확도가 떨어질 수 있다. 또한, 사용자가 작성한 트윗 집합의 단일 트윗 각각을 모두 비교하는 방법은 계산 횟수의 증가로 오버헤드가 발생하여 실제 시스템에 적용하기에는 성능상으로도 문제가 있다.

본 논문에서는 트윗 개체 링킹에서 발생하는 데이터 희소성 문제를 해결하고 링킹의 정확성을 향상시키고자 사용자의 정보와 집단의 정보를 통합하여 활용한다. 특히, 집단의 정보 활용 시 사용자의 유사도를 측정하여 유사성에 따라 사용자의 정보에 가중치를 주어 트윗 개체 링킹에서 개체 중의성을 해소하는 기법을 제안한다. 이때 사용자가 작성한 모든 트윗 집합을 트윗 코퍼스라고 하고, 트윗 코퍼스를 비교하여 문맥적인 유사도, 즉 사용자 간 유사도를 측정한다. 제안하는 방식은 단일 트윗 간의 비교를 통해 사용자 유사도를 측정하는 기존의 기법에 비해 시간적인 비용이 절약되며, 코퍼스의 크기가 단일 트윗에 비해 크기 때문에 사용자들 간의 유사성을 비교적 정확하게 측정하여 기존의 기법보다 높은 링킹 정확성을 갖는다.

이 논문의 구성은 다음과 같다. 제2장에서는 관련 연구로 개체 링킹 기법 중 특히 트윗 개체 링킹 기법에 관한 연구에 대해 언급한다. 제3장에서는 본 논문에서 제안하는 사용자 유사도에 대해 정의하고, 사용자 유사도 기반의 개체 중의성 해소를 위한 트윗 개체 링킹 기법 대해 자세히 서술한다. 제4장에서 제안하는 기법의 성능을 평가하기 위한 실험에 관해 설명하고, 기존의 개체 링킹 연구와 비교 평가하여 제안하는 트윗 개체 링킹 기법의 우수성을 검증한다. 마지막으로 제5장에서는 결론 및 향후 연구에 대해 제시한다.

2. 관련 연구

2.1 개체 중의성 해소를 위한 개체 링킹 기법

개체 링킹은 개체명 인식, 개체 중의성 해소, 링크 개체 선택의 세 가지 문제를 다루고 있다. 지식 베이스 기반 개체 링킹에서의 개체명 인식(named entity recognition)은 텍스트 내 모든 가능한 지식 베이스에 존재하는

개체들의 정규화 된 명칭들의 후보들을 추출하는 것이다[10]. 초기 개체 링킹의 모델에서는 텍스트 문서 내 모든 가능한 n-gram 용어 중 개체명 사전에 해당하는 것들만 추출하는 방식이 시도되었고[2], 최근에는 개체명과 정답 의미가 부여된 문서를 분류기로 학습하여 개체명을 인식하는 방법이 등장하였다[3].

개체명 인식 후에는 개체명에 대한 중의성 해소 과정을 거친다. 개체 중의성 해소와 관련된 기존 연구에서는 크게 개체 선험 확률(entity prior probability), 공기 용어 정보(co-occurrence term information), 공기 개체 의미 관련도(co-occurrence entity semantic relatedness)가 활용되고 있다. 개체 선험 확률은 위키피디아와 같은 지식 베이스의 기사들에 얼마나 많이 등장하는가에 대한 확률로써, 최빈(the most frequent) 의미를 찾는다. [11]은 개체표현의 각 후보 개체에 대해 비중의성 공기 개체들과의 평균 의미 관련도와 개체 선험 확률을 통해 개체 중의성을 해소하기 위한 시도를 제안하였다. 공기 용어 정보는 개체표현이 포함된 문서 내 다른 용어들을 의미하는데, 이러한 용어 정보를 이용해 문맥적인 유사도를 구할 수 있다. 공기 용어 정보를 활용한 연구로 [12]에서는 개체표현과 후보 개체 매칭률 계산을 위해 TF-IDF (Term Frequency-Inverse Document Frequency)를 가중치로 한 코사인 유사도(cosine similarity)를 사용한다. 의미 관련도는 위키피디아의 개체 중의성 해소를 위한 대부분의 방법론에서 활용되고 있다. 공기 개체 의미 관련도 지표인 NGD(Normalized Google Distance)는 용어의 hit count를 이용하여 두 용어의 의미적 거리를 계산하기 위해 제안되었고[13], 최근에는 개체 중의성 해소를 위해 용어의 hit count 대신 의미의 링크 집합을 활용한다. 웹 문서 기반의 연구는 비교적 적은 지표를 이용하여 간단한 수식으로도 개체 링킹이 가능하며, 개체 링킹의 성능을 향상시키기 위한 다양한 연구가 진행되고 있다. 하지만 웹 문서 기반의 개체 링킹 기법을 트위터 문서에 그대로 적용하는 것은 적합한 방식이 아니다.

2.2 트윗 개체 링킹 기법

트위터 문서는 노이즈가 많고 길이가 짧은 텍스트를 다루기 때문에 불충분한 문맥 정보와 불완전한 개체를 제공한다. 다시 말해, 이러한 트위터 문서의 특성상 문서에 두 개 이상의 개체표현을 갖는 경우가 극히 드물어 개체표현 간 의미적인 관련도를 구할 수 없는 경우가 빈번히 발생한다. 이를 데이터 희소성 문제라 하며, 트위터 문서 기반의 개체 링킹에 대한 연구들은 이러한 데이터 희소성을 해결하기 위해 웹 문서 기반의 접근 방법과 함께 뉴스 스트림을 이용하거나 사용자 정보, 집단 정보 내의 코퍼스를 활용한다.

데이터 희소성 문제를 해결하기 위한 트윗 개체 링킹 기법에서는 사용자 정보를 활용하는 방법이 주로 사용된다[6,7]. [6]은 사용자 관심사 모델링을 통한 개체 링킹 기법을 제안하였다. 웹 문서 기반의 개체 링킹 기법과 같은 방식인 트윗 내의 로컬 정보(intra-tweet local information)를 고려할 뿐만 아니라 사용자가 작성했던 트윗을 통해 얻어진 사용자 관심사 정보(inter-tweet user interest information)를 결합하는 방식을 시도하였다. 이는 개체 선험 확률, 공기 용어정보, 공기 개체 의미 관련도를 모두 이용하여 개체 순위화를 위한 수식을 제안한다. [7]은 사용자 정보를 활용해 다른 트윗 내의 개체표현을 비교할 때 Wikipedia-Miner API를 통해 측정한 의미 관련도에 두 트윗 사이의 문맥적인 유사도를 가중치로 둔 수식을 제안하였다. 이와 같이 사용자 정보를 이용한 방법은 사용자의 트윗 기록이 많을 경우 트윗 개체 링킹에서 발생하는 데이터 희소성 문제를 해결할 수 있고, 트윗 내의 로컬 정보만 활용한 기법보다 더 좋은 성능을 보인다[6,7]. 하지만, 사용자가 가지는 트윗이 적을 경우에는 의미 있는 개체의 수도 줄어들기 때문에 여전히 데이터 희소성 문제가 발생하고 링킹의 정확도가 떨어지는 문제점이 발생한다.

사용자 정보만을 활용한 연구의 한계를 극복하기 위해 집단의 정보를 활용하거나 외부 자원을 통해 정보를 확장하는 트윗 개체 링킹기법들이 등장하였다[8,9]. [9]는 사용자 모델과 지식 베이스만 사용하는 개체 링킹의 한계를 극복하기 위해 최신 이슈를 포함하는 뉴스 스트림을 통해 개체 링킹을 수행하는 모델을 제안하였다. 하지만 이 기법은 사용자의 관심사와 관련이 없는 뉴스 스트림이 사용될 경우 링킹의 정확성이 떨어질 수 있다는 문제점이 있다. [8]은 단순히 불특정 다수의 정보를 이용하는 것이 아닌 다른 트윗과의 관련성을 고려한 집단 추론 방법 기반의 개체 링킹 기법을 제안하였다. 이는 집단 정보에서 추출한 후보 개체들 각각의 개체 선험 확률, 트윗 사이의 공기 개체 정보를 이용한 문맥 유사도 그리고 후보 개체 사이의 의미적인 관련도를 이용하고, 다른 트윗의 집합을 반영할 때 동일한 계정에서 작성된 문서인지 아닌지에 따라 스코어를 다르게 부여한다. 또한, 트윗 간의 문맥적인 유사도를 구하기 위해 코사인 유사도, 공통적인 해시태그의 발생 여부와 편집거리(edit distance)를 고려한다. 집단 정보를 사용하면 트위터 문서를 다룰 때 발생하는 데이터 희소성 문제를 해결할 수 있다. 하지만, [8]에서처럼 사용자들 간 트윗의 문맥적인 유사도를 구하기 위해 사용자의 단일 트윗 각각을 모두 비교한다면 계산 시간적인 측면에서 성능이 좋지 않다. 또한 일반적으로 개체 링킹에서 길이가 짧은 문서를 비교하여 문맥의 유사성을 측정하는 것은 링킹의 정확성이 떨어지기 때문에 사용자가 작성한 단

일 트윗들을 각각 비교하여 유사성을 측정하는 것은 실제 트윗 개체 링킹 시스템에 적용하기에 성능상으로 문제가 있다.

사용자 정보와 집단 정보를 함께 이용하는 것은 트윗 개체 링킹에서 중요한 역할을 한다. 본 논문에서 사용자 정보와 집단 정보를 통합하여 활용하고, 집단 정보를 사용함에 있어 사용자가 작성한 모든 트윗 집합의 문맥적인 유사도를 측정하여 높은 관련성이 있는 사용자의 정보에 가중치를 주어 높은 성능을 갖는 트윗 개체 링킹 기법을 제안한다. 또한 사용자 유사도 계산 시 오버헤드가 적게 발생하도록 하기 위해 각 사용자들이 작성한 모든 트윗 집합의 문맥적인 유사도를 측정한다.

3. 사용자 유사도 기반 개체 링킹 기법

본 논문에서 제안하는 방법은 세 가지 특성으로 구성되며, 이는 각각 로컬 특성(local features), 사용자 특성(user features), 글로벌 특성(global features)으로 이루어진다. 그중에서 글로벌 특성에 초점을 맞춰 사용자가 작성한 트윗 집합 문서의 유사도를 측정하여 유사도가 높은 사용자에 가중치를 주는 트윗 개체 링킹 방식을 제안한다.

제안하는 트윗 개체 링킹 방식의 전체적인 프레임워크는 그림 1과 같다. 우선 위키피디아를 통해 트위터 문서의 개체표현에 대한 후보 개체가 포함되어 있는 dictionary D를 생성한다. 그다음은 중의성 해소 프로세스의 필요성 여부를 판단하는 작업을 수행한다. 만약 후보 개체의 개수가 하나인 비중의성 개체인 경우라면 중의성 해소 프로세스를 거치지 않고 첫 번째 후보 개체 $e_{i,1}$로 바로 링크한다. 중의성 개체의 경우는 마지막으로 중의적인 의미를 갖는 개체표현에 대한 명확화 과정을 수행한다. 개체표현 m_1의 중의성을 해소하기 위해 단일 트윗(single-tweet) 내의 다른 개체 정보들을 이용한 로컬 정보, 동일한 사용자가 작성한 트위터 문서 정보를 이용한 사용자 특성 그리고 다른 사용자들의 트위터 문서를 사용한 글로벌 특성을 통합하여 최댓값을 갖는 후보 개체 e_i^*로 트윗 개체 링킹을 수행한다.

3.1 후보 개체 집합 생성

개체표현에 대한 후보 개체 집합을 구하는 것은 개체 중의성 해소를 위해 필수적인 과정으로, 사용하는 지식 베이스와 집합을 구성하는 방법이 다양하다. 본 논문에서는 후보 개체를 생성하는 데 필요한 지식 베이스로 위키피디아(Wikipedia)를 사용한다. 링크할 개체에 대한

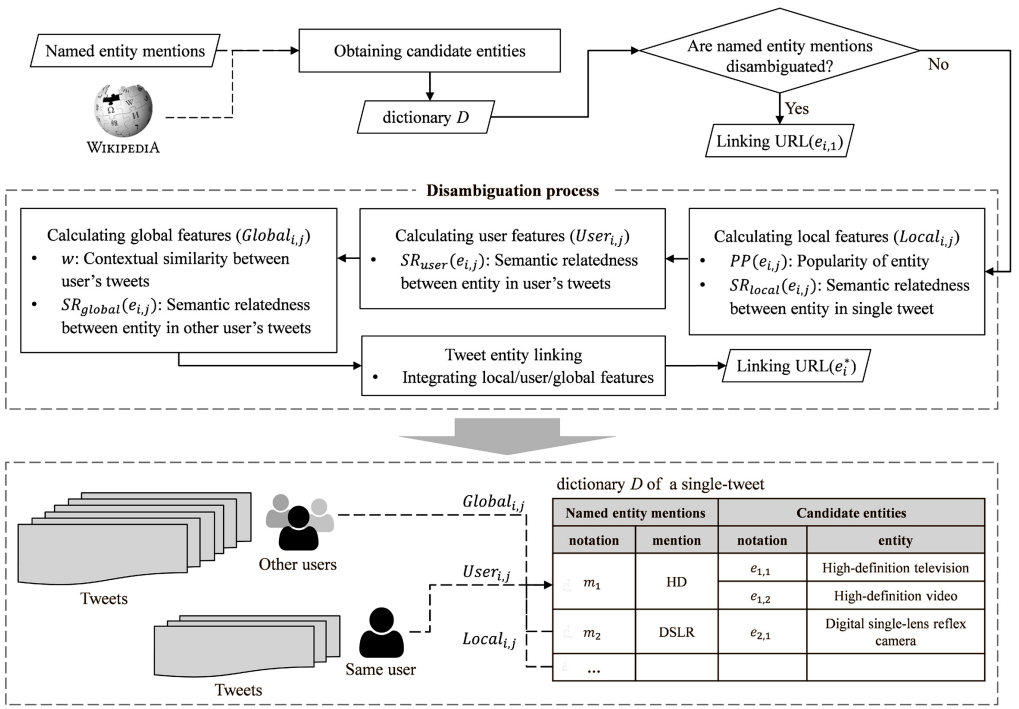

그림 1 트윗 개체 링킹 프레임워크

Fig. 1 Tweet Entity Linking Framework

후보 집합은 위키피디아의 disambiguation page를 근거로 하며, 찾아낸 후보 개체 집합을 dictionary D에 저장한다. D는 멘션과 후보 개체 그리고 각 후보 개체가 위키피디아에 링크된 페이지 수(link count)로 구성된다. 멘션 $M=\{m_1, m_2, \dots, m_i\}$은 트윗에서 추출된 개체표현을 나타내고, 멘션 m_i에 링크되는 후보 개체 집합을 $E_i=\{e_{i,1}, e_{i,2}, \dots, e_{i,j}\}$로 정의한다. 링크 수는 각각의 후보 개체와 링크되는 위키피디아 기사의 수를 의미한다. 링크 수가 작은 후보 개체는 드물게 언급되는 개체이므로 일부 제외하고 구성한다. 앞의 표 1은 구축한 D의 일부분을 나타낸다.

표 1 사전 예시
Table 1 A part of the dictionary D

Surface form	Candidate Entity	Link Count
UPS	United Parcel Service	343
	Uninterruptible power supply	45
	UPS	28
Hacker News	Hacker News	7
HD	High-definition television	311
	High-definition video	176
	HD Radio	24
	Henry Draper Catalogue	23
	HD	20
DSLR	Digital single-lens reflex camera	95

3.2 사용자 정보를 활용한 특성

사용자 정보를 활용한 특성에는 단일 트윗 내 개체표현들과의 의미적인 관련도를 통해 측정하는 로컬 특성, 사용자가 작성한 다른 트윗 정보를 사용해 의미적인 관련도를 구하는 사용자 특성이 있다. 의미 관련도는 두 의미의 상호 관련된 정도를 수치화한 것으로 개체 링킹에서는 지식 베이스의 전체 문서 집합을 코퍼스로 고려하여 두 개체의 의미 관련도를 계산하는 것이 일반적이다. 의미 관련도 중 하나인 *NGD*는 위키피디아 지식 베이스를 통해 개체 중의성을 해소하기 위한 대부분의 방법론에서 활용되고 있다. 두 개체 e_i, e_j 사이의 의미 관련도 $NGD(e_i, e_j)$는 다음 식으로 정의된다[13]. L_i, L_j는 각각 e_i, e_j로의 링크를 포함하는 위키피디아 페이지들의 집합이며, N은 전체 위키피디아 개체의 개수이다.

$$NGD(e_1, e_2)=\frac{\max(\log(|L_1|), \log(|L_2|))-\log(|L_1\cap L_2|)}{\log(N)-\min(\log(|L_1|), \log(|L_2|))} \quad (1)$$

그러나 식 (1)은 두 의미 간의 거리를 표현하고 있으므로 의미 관련도를 사용하기 위해 기존 개체 중의성 해소 방법에서는 $1-NGD$의 형태로 변경하여 활용되었다[4]. 또한 *NGD* 지표는 이론적으로 0에서 $+\infty$ 사이의 값으로 *NGD*를 0과 1 사이의 의미 관련도로 정규화하여, 최종적으로 아래의 식 (2)와 같이 의미 관련도를 구할 수 있다. NGD_n[14]는 정규화를 거친 의미 관련도를 의미하며, 두 개체 e_i, e_j 사이의 의미 관련도를 $SR(e_i, e_j)$[15]로 나타낸다.

$$SR(e_i, e_j)=1-NGD_n(e_i, e_j) \quad (2)$$

로컬 특성은 개체 선험 확률과 의미 관련도를 통해 계산한다. 하나의 트윗 내 개체 선험 확률은 dictionary D에서 구했던 각 후보 개체의 link count 값에 모든 후보 개체들의 link count 값의 합을 나누어 구한다. 이는 위키피디아에 등장하는 빈도 즉, 개체의 인기성을 의미한다[1]. 로컬 의미 관련도는 개체표현이 언급된 단일 트윗 내의 모든 개체표현에 대한 후보 개체들 사이에서 구한 의미 관련도의 평균값이다. m_i에 대한 j번째 후보 개체인 $e_{i,j}$의 개체 선험 확률을 $PP(e_{i,j})$라 하고, 로컬 의미 관련도는 $SR_{local}(e_{i,j})$로 나타낸다. 따라서, 로컬 스코어 $Local_{i,j}$는 개체 선험 확률과 의미 관련도를 통해 아래의 식 (3)과 같이 구한다[6]. E_L는 단일 트윗에서 m_i가 아닌 다른 개체표현에 대한 모든 후보 개체를 포함하는 집합을 의미한다.

$$Local_{i,j}=\frac{PP(e_{i,j})+SR_{local}(e_{i,j})}{2} = \frac{PP(e_{i,j})+\frac{1}{|E_L|}\sum_{\forall e_{a,b}\in E_L} SR(e_{i,j}, e_{a,b})}{2} \quad (3)$$

$E_L=\{e_{a,b}|e_{a,b}$는 m_i가 발생한 트윗 내 다른 개체표현에 대한 모든 후보 개체$\}$

사용자 특성은 m_i를 작성한 사용자의 다른 트윗 정보를 사용하여 트윗 집합 내의 모든 개체표현에 대한 후보 개체 집합과의 의미 관련도의 평균값으로 구해진다. 후보 개체 $e_{i,j}$의 사용자 스코어 $User_{i,j}$는 다음의 식 (4)로 계산한다[6]. E_{U_k}는 m_i가 발생한 트윗을 제외한 사용자의 다른 트윗 집합 내의 모든 후보 개체 집합을 의미한다.

$$User_{i,j}=SR_{user}(e_{i,j})=\frac{1}{|E_U|}\sum_{\forall e_{a,b}\in E_U} SR(e_{i,j}, e_{a,b}) \quad (4)$$

$E_{U_k}=\{e_{a,b}|e_{a,b}$는 사용자 U_k의 트윗 정보에 나타나는 모든 후보 개체$\}$

3.3 집단 정보를 활용한 특성

집단 정보를 활용한 글로벌 특성에서는 사용자가 작성한 트윗 집합을 트윗 코퍼스라 하고 트윗 코퍼스 간의 유사도를 측정한다. 사용자들 간 트윗 정보 사이의 의미적인 관련도를 구하고, 측정한 사용자 유사도를 가중치로 두어 비슷한 관심사를 가지는 사용자들의 트윗 정보에 높은 영향력을 부여한다. 기존 트위터상의 사용자 유

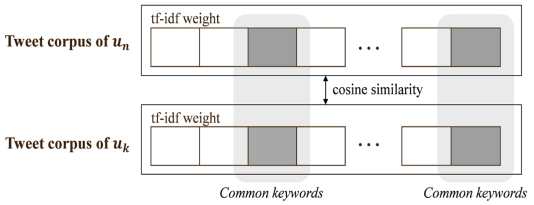

그림 2 사용자의 트윗 코퍼스를 통한 사용자 유사도
Fig. 2 User similarity through user's tweet corpus

사도를 구하는 방법은 일반적으로 사용자 정보나 활동, 영향력에 기반을 둔다[16-18]. 하지만, 제안하는 방법에서의 사용자 유사도는 사용자들이 서로 얼마나 관련 있는 문서를 작성하는지에 대해 측정해야 하므로 기존 방법과 달리 우선시해야 할 것은 사용자들이 작성했던 트윗 코퍼스 간의 문맥적인 유사도를 측정하는 것이다.

트윗 코퍼스 사이의 문맥적인 유사도를 구하기 위해 TF-IDF 벡터를 가중치로 하는 코사인 유사도를 측정한다. 트윗 문서의 경우 특정 단어열을 포함하는 트윗의 코퍼스의 크기가 일반 문서에 비해 작으므로 단일 트윗을 비교하는 것은 의미 있는 값을 얻기 어렵다. 따라서 제안하는 사용자 유사도는 그림 2와 같이 각 사용자의 트윗 집합을 하나의 코퍼스로 보고, 두 사용자 간의 유사도를 구할 때에는 사용자의 트윗 코퍼스 간 코사인 유사도를 통해 측정된다.

이를 위해 사용자별로 트윗 집합 전체의 내용을 word-gram 용어로 토큰화하고, 그중 불용어 사전에 등록된 단어들을 제거한다. 다음으로 코퍼스의 TF-IDF 벡터를 구한 후 사용자 간의 코사인 유사도를 통한 사용자 유사도 w는 식 (5)로 나타낼 수 있다. $\vec{v}(u_n)$는 사용자 u_n이 작성한 용어에 대한 TF-IDF 벡터값을 의미한다. 정규화를 통해 사용자 유사도는 0에서 1 사이 값을 가지고, 같은 사용자 간의 사용자 유사도는 1로 가정한다. 사용자 유사도 값이 클수록 사용자 간의 유사도가 크며, 유사도가 클수록 트윗이 개체 중의성 해소에 반영되는 영향력이 높다.

$$w = sim(u_n, u_k) = \frac{\vec{v}(u_n)\vec{v}(u_k)}{|\vec{v}(u_n)||\vec{v}(u_k)|} \quad (5)$$

글로벌 의미 관련도는 집단의 정보 즉, 다른 사용자의 트윗 집합을 사용한다. 이는 트윗 내 후보 개체들과의 의미 관련도 평균값으로 구한다. 사용자에 따라 유사한 관심사를 갖거나 혹은 아예 관련성이 없는 내용을 언급할 수도 있다. 관련성에 따라 다른 가중치를 부여하기 위해 사용자별로 구한 사용자 유사도를 의미 관련도에 반영한다. 후보 개체 $e_{i,j}$의 글로벌 스코어 $Global_{i,j}$는 아래의 식 (6)과 같이 나타낼 수 있다. u_n은 $e_{i,j}$를 가지는 사용자를 나타내며, E_G는 집단 정보에 나타나는 모든 개체표현에 대한 후보 개체 집합으로써 모든 사용자 $U=\{U_1, U_2, \ldots, U_n\}$가 가지는 후보 개체 집합을 의미한다.

$$Global_{i,j} = w \times SR_{global}(e_{i,j}) = \frac{1}{|E_G|} \sum_{U_k \in U} \sum_{\forall e_{a,b} \in E_{U_k}} sim(u_n, u_k) \times SR(e_{i,j}, e_{a,b}) \quad (6)$$

$$E_G = \{e_{a,b} | e_{a,b} \in E_{U_1} \cup E_{U_2} \ldots \cup E_{U_n}\}$$

3.4 트윗 개체 링킹

트윗 개체 링킹을 위한 후보 개체 순위화는 계산된 로컬 스코어, 사용자 스코어와 글로벌 스코어를 모두 적용하여, 구한 결과 중 최댓값을 갖는 개체로 결정된다. m_i의 후보 개체 집합 중 제안하는 트윗 개체 링킹 기법을 통해 결정된 최종 개체인 e_i^*는 다음과 같이 구할 수 있다. 상수 α, β, γ는 $\alpha+\beta+\gamma=1$을 만족하고, 실험을 통해 각각 $0.6, 0.2, 0.2$로 설정하였다.

$$e_i^* = \arg max(\alpha Local_{i,j} + \beta User_{i,j} + \gamma Global_{i,j}) \quad (7)$$

4. 실험 및 평가

본 논문에서의 실험은 트위터 데이터[6]을 이용하여 제안하는 개체 링킹 기법의 성능을 실험을 통해 증명하였다. 지식 베이스는 위키피디아에서 제공하는 2015년 12월 1일에 등록된 데이터를 사용하였다. 제안 방법의 링킹 성능을 증명하기 위해 웹 문서 기반 개체 링킹 방법인 'Unambiguous only', MFS (the Most Frequent Sense), WLM (Wikipedia Link-based Measure)과 트위터 문서 기반의 KAURI (Knowledge bAse via UseR Interest modeling), KAURI (global), CIM (Collective Inference Method)과 비교평가 한다. 다음은 제안하는 트윗 개체 링킹 기법과의 성능 비교를 위한 기존의 개체 중의성 해소 기법을 간략히 설명한다.

- 'Unambiguous only': 입력 문서 내 개체표현 중 비중의성 개체표현만을 고려한다.
- MFS[11]: 입력 내 개체표현의 후보 개체에 대해 가장 빈번하게 사용되는 의미를 고려한다.
- WLM[15]: MFS 방법과 Wikipedia-Miner API를 통한 의미적인 관련도를 고려한다.
- KAURI[6]: 사용자 정보를 사용하여 MFS 방법, 문맥적인 정보와 비중의성 공기 개체 기반의 의미적인 관련도를 고려한다.
- KAURI (global)[6]: 사용자 정보만을 고려하는 KAURI 시스템을 확장하여 집단 정보를 통해 코퍼스의 크기를 증가시켜 개체 중의성 해소를 수행한다.
- CIM[8]: 집단 정보를 사용하여 MFS 방법, 의미 관

련도와 문맥적인 유사성을 검사하기 위해 트윗 사이의 코사인 유사도(TF-IDF vectors, topic distribution vectors), 계정 동일 여부(1 또는 0), 공통된 해시태그의 포함 여부(1 또는 0), 편집 거리(edit distance) 등을 고려한다.

4.1 데이터 집합

본 논문에서는 [6]에서 제공한 트위터 데이터를 사용하고, 이는 표 2와 같다. 데이터는 20명의 사용자와 사용자가 작성한 최대 200개까지의 트윗을 제공한다. 총 3,818개의 트윗 중 적어도 하나의 개체가 언급된 트윗은 1,721개이고, 개체표현(named entity mentions)은 2,918개이다. 이 중 불확실한(uncertain) 개체 245개를 제외한 2,653개의 개체표현으로 테스트하였다. 테스트 개체표현 중 위키피디아로부터 후보 개체에 대한 정보를 받아왔을 때 후보 개체가 없는(unlinkable) 개체가 410개가 발생하였고, 이를 제외한 총 2,263개의 개체표현에 대해 실험하였다. 개체표현 2,263개에 대한 후보 개체는 8,260개로써 하나의 개체표현 당 평균 약 3.7개의 후보 개체를 가진다. 또한, 2,263개의 개체표현 중 비중의성 개체표현은 1,058개이며 중의성을 갖는 개체표현의 평균 후보 개체 집합의 개수는 약 6개이다.

4.2 정확도 측정

성능을 평가하기 위한 지표는 iAcc(micro Accuracy)와 aAcc(macro Accuracy)를 사용한다. iAcc는 개별 개체표현 단위의 정확률이며, aAcc는 트윗 단위로 계산된 iAcc들의 평균 정확률로 아래와 같은 방법으로 구할 수 있다[10].

$$iAcc = \frac{\text{정답 의미로 결정된 개체표현 개수}}{\text{테스트셋 내 개체표현 전체 개수}}$$

$$aAcc = \text{테스트셋 내 트윗별 } iAcc\text{들의 평균}$$

그림 3은 기존의 개체 링킹 방법들과 제안하는 방법을 정량적으로 비교평가한 결과를 나타낸다. 'Unambiguous only' 방법은 비중의성 개체표현들에 대해서만 개체 링킹을 수행하므로 비중의성 개체표현 1,058개 중 정답 의미와 일치하는 개체표현이 995개로써 비중의성 개체표현 내에서의 정확도는 94.04%이나, 전체 개체표현에 대해서는 43.93%로 절반에도 못 미치는 결과 값이 나오게 된다.

웹 문서 기반의 개체 링킹 방법인 MFS와 WLM 방법은 개체 중의성 해소를 위한 과정을 통해 개체 링킹의 성능을 높였다. MFS와 WLM 방법의 차이점은 WLM 방식에서는 두 개체 사이의 의미 관련도를 추가하여 측정한다는 점이다. 웹 문서를 사용했을 때에는 문서 내

표 2 실험데이터

Table 2 Data Set

Twitter user	Tweets		Named entity mentions		Test named entity mentions		
	All #	Test #	All #	Uncertain #	All #	Linkable #	Unlinkable #
20	3,818	1,721	2,918	245	2,673	2,263	410

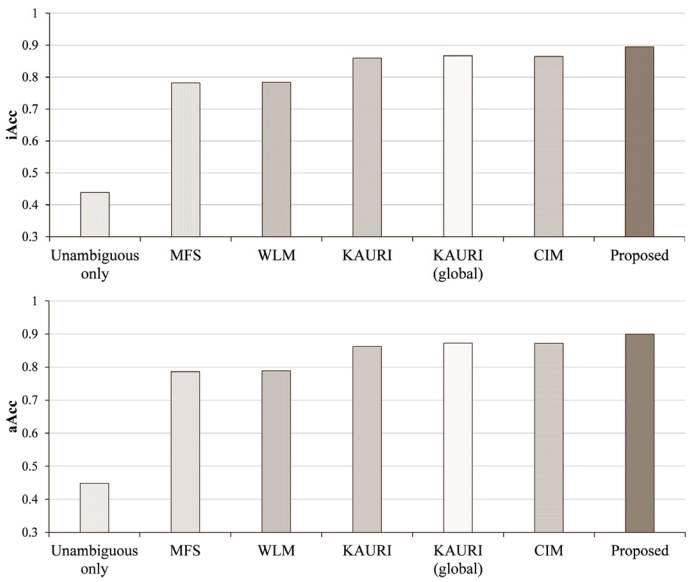

그림 3 정량적 비교평가 결과

Fig. 3 Result of comparison evaluation

개체와의 관련성을 고려하는 WLM 방식의 정확률이 월등히 높다. 하지만, 트위터 문서의 경우에는 길이가 짧아 의미적인 관련도를 구하는 데 필요한 개체표현이 없는 경우가 빈번하게 발생한다. 실험한 데이터 집합의 경우를 예로 들면 1,721개의 트윗 중 개체표현이 두 개 이상인 트윗이 556개로, 오직 556개의 트위터 문서에서만 의미 관련도를 구할 수 있다. 그렇기 때문에 트위터 문서 기반으로 실험한 결과 성능 차이가 0.23% 밖에 차이가 나지 않았다. 이를 통해 단일 문서에 나타나는 개체표현을 통한 의미 관련도는 트위터 문서에서는 중의성 해소에 거의 영향을 주지 못한다는 것을 결과적으로 확인할 수 있다.

트위터 문서 기반의 개체 중의성 해소를 위한 개체 링킹 방법인 KAURI, KAURI(global)과 CIM은 트위터 문서의 특성상 불충분한 문맥 정보를 다루므로 다른 트윗 집합을 통해 이를 보완하고자 하였다. KAURI는 트윗 내의 정보와 사용자 정보를 활용하였고, KAURI (global)과 CIM은 집단 정보도 모두 고려한 점이 다르다. 사용자 정보를 활용한 개체 링킹 방법에서는 사용자 당 180~200개 사이의 트위터 문서를 이용할 경우 가장 높은 성능을 보이는데, 사용하는 트위터 데이터도 마찬가지로 사용자 정보를 활용한 방법에 유리한 데이터 집합이다. 사용자가 작성한 트위터 문서들은 비슷한 토픽을 언급할 가능성이 높고, 반대로 집단의 트위터 문서에는 다양한 토픽들이 언급되어있다. 따라서, KAURI (global)은 KAURI 시스템에 비해 활용한 코퍼스의 크기가 20배 이상이 늘어났음에도 불구하고 0.82%의 성능 향상률을 보였다. 또한, CIM은 KAURI 시스템과 비교했을 때 0.57%의 적은 성능 차이를 보이는데, 이를 통해 제안하는 사용자 유사도가 트윗과 트윗을 비교한 문맥 유사도를 가중치로 두는 CIM의 방식보다 더 효율적이라는 것을 확인할 수 있다. 하지만 집단 정보를 이용한 KAURI (global)이나 CIM이 사용자 정보만을 이용한 KAURI 시스템보다 좋은 성능을 보이는 것으로 보아 트위터 문서를 사용할 때 사용자 정보와 집단 정보 모두 반영하는 것이 더 정확한 링킹이 가능하다는 것을 알 수 있다.

마지막으로 제안하는 방법과 기존의 개체 링킹 기법과 비교했을 때, 'Unambiguous only' 방법보다 103%가 향상된 것으로 보아 트윗 개체 링킹에서도 개체의 중의성을 해소하는 작업은 매우 중요하다는 것을 알 수 있다. 또한, MFS 방법에 비해 약 14.4% 상승하였으며, WLM 방법과 비교하였을 때에도 14.2%의 비슷한 상승률을 보였다. 이는 트위터 문서를 기반으로 할 때에는 웹 문서 기반의 개체 링킹 방법을 그대로 쓸 때의 문제점을 보여주며 웹 문서 기반의 개체 링킹과는 다른 접근 방법이 필요함을 의미한다. KAURI와 비교했을 때에는 4.05%가 증가하였고, KAURI (global), CIM과는 각각 3.21%, 3.47%의 성능 향상을 보였다. 결과적으로, 본 연구에서 제안하는 개체 링킹 기법은 트윗 기록이 많은 사용자를 대상으로 한 실험 데이터에 대해서도 정확도가 향상된 것으로 보아 집단 정보를 이용하는 것은 데이터 희소성 문제와 부정확한 링킹 문제 역시 해결할 수 있음을 실험을 통해 보였다. 더불어 계산시간적인 측면에서도 사용자들간의 유사도를 측정하는 CIM 방식에 비해 효율적이다. n명의 사용자가 가지는 각각의 트윗 집합의 크기가 k라고 할 때, 문맥적인 유사도를 구하는 데에 걸리는 시간을 시간 복잡도로 나타내면 CIM의 경우 $O(n^2k^2)$인 반면 제안하는 사용자 유사도는 $O(n^2)$의 시간이 소요되므로 기존의 기법에 비해 비용이 줄어들기 때문에 효과성(efficiency)이 더 높다.

5. 결론 및 향후 연구

이 논문은 개체 중의성 해소를 위한 사용자 유사도 기반의 트윗 개체 링킹 기법을 제안한다. 제안하는 트윗 개체 링킹 기법은 데이터 희소성 문제를 해결하기 위해 사용자 정보와 집단 정보를 모두 사용한다. 또한, 집단 정보를 이용할 때에 사용자 유사도를 측정하여 유사도가 높은 사용자 정보에 가중치를 주어 후보 개체를 순위화하기 때문에 데이터 희소성 문제 해결과 동시에 링킹의 정확도를 향상시킬 수 있다. 제안 방법의 성능을 측정하기 위해서 실제 트위터 데이터를 사용하였고, 지식 베이스 내의 지식을 기반으로 후보 개체 추출과 두 개체 간의 의미 관련도를 측정하여 실험을 진행하였다. 실험을 통해 제안하는 기법이 기존의 웹 문서 기반의 기법보다 성능이 월등히 좋음을 보였고, 기존 트윗 개체 링킹 기법보다도 높은 성능을 보임을 확인하였다. 또한, 실험결과를 통해서 제안하는 사용자들 간 유사도를 구하는 방식이 기존 기법에서 사용하는 트윗 간의 문맥적인 관련성을 구하는 방식에 비해 더 정확하고, 비용적 측면에서도 효과적인 것을 확인할 수 있다.

본 논문에서는 트위터 문서에 대한 개체 링킹 기법을 다뤘지만, 트위터 문서와 비슷한 속성을 갖는 다른 마이크로블로그 문서에도 적용이 가능할 것이다. 이를 검증하기 위해 제안 방법이 트위터 문서에 대해서만 국한되는 것이 아니라 다양한 마이크로블로그 환경에서도 높은 성능을 갖는다는 것을 실험을 통해 보이고자 한다. 또한, 제안하는 트윗 개체 링킹 프레임워크를 실제 트위터 데이터를 실시간으로 크롤링하여 자동화된 링킹 시스템으로 구현하고자 한다. 마지막으로 실험대상이 되는 사용자의 수를 늘려 유사한 속성을 가진 사용자 집단을 클러스터링 하고, 그룹 기반의 개체 링킹 기법에 대해 연구하고자 한다.

References

[1] W. Shen and J. Wang, "Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions," *Journal of IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, pp. 443-460, 2015.

[2] R. Mihalcea and A. Csomai, "Wikify!: linking documents to encyclopedic knowledge," *Proc. Of the 16th Conference on Information and Knowledge Management*, pp. 233-242, 2007.

[3] D. Milne and I. H. Witten, "Learning to Link with Wikipedia" *Proc. Of the 18th Conference on Information and Knowledge Management*, pp. 509-518, 2008.

[4] X. Han, L. Sun and J. Zhao, "Collective Entity Linking in Web Text: A Graph-Based Method," *Proc. Of the 34th International ACMSIGIR Conferenceon Research and Development in Information Retrieval*, pp. 765-774, 2011.

[5] S. Kulkarni et al., "Collective annotation of Wikipedia entities in web text," *Proc. Of the 15th ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 457-466, 2009.

[6] W. Shen, J. Wang, P. Luo and M. Wang, "Linking Named Entities in Tweets with Knowledge Base via User Interest Modeling," *Proc. Of the 19th SIGKDD international conference on Knowledge Discovery and Data Mining (KDD)*, pp. 68-76, 2013.

[7] R. Bansal et al., "EDIUM: Improving Entity Disambiguation via User Modeling," *Journal of Advances in Information Retrieval*, pp. 418-423, 2014.

[8] X. Liu et al., "Entity Linking for Tweets," *Proc. Of the 51th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1304-1311, 2013.

[9] S. Jeong, Y. Park, S. Kang and J. Seo, "Entity Linking for Tweets using User Model and Real-time News Stream," *Journal of Cognitive Science*, pp. 435-452, 2015. (in Korean)

[10] I-S. Kang, "An Effect of Semantic Relatedness on Entity Disambiguation: Using Korean Wikipedia," *Journal of Korean Institute of Intelligent Systems*, pp. 111-118, 2015. (in Korean)

[11] O. Medelyan, I. H. Witten, D. Milne, "Topic indexing with Wikipedia," *Proc. Of the Wikipedia and AI workshop at AAAI-08*, 2008.

[12] S. Dill. Eiron et al., "Semtag and seeker: bootstrapping the semantic web via automated semantic annotation," *Proc. Of the 12th international conference on World Wide Web (WWW)*, pp. 178-186, 2003.

[13] R. L. Cilibrasi and Paul M. B. Vitanyi, "The Google Similarity Distance," *Proc. Of IEEE Transation on Knowledge and Data Engineering*, Vol. 19, No. 3, pp. 370-383, 2007.

[14] J. Gracia et al., "Querying the Web: A Multiontology Disambiguation Method," *Proc. Of the 6th International Conference on Web Engineering*, 2006.

[15] D. Milne and I. H. Witten, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," *Proc. Of the AAAI Workshop on WIKIAI*, pp. 25-30, 2008.

[16] Y-D. Seo, J-D. Kim and D-K. Baik, "PReAmacy: A Personalized Recommendation Algorithm considering Contents and Intimacy between Users in Social Network Services," *Journal of KIISE*, Vol. 41, No. 4, pp. 209-226, 2014. (in Korean)

[17] K-S. Seol, J-D. Kim, H-N. Shin, D-K. Baik, "Intimacy Measurement Method and Experiment between Social Network Service Users," *Journal of KIISE*, Vol. 39, No. 4, pp. 335-341, 2012. (in Korean)

[18] H. Park, H. Kwak, M. Cha and S. B. Moon, "Influentials Ranking in Social Networks," *Journal of KIISE*, Vol. 28, No. 3, pp. 24-30, 2010. (in Korean)

김 서 현

2015년 금오공과대학교 컴퓨터공학과(학사). 2015년~현재 고려대학교 컴퓨터학과 석사과정. 관심분야는 소셜 네트워크 서비스, 빅 데이터, 자연어처리, 자가적응 소프트웨어

서 영 덕

2012년 고려대학교 컴퓨터학과(학사). 2012년~현재 고려대학교 컴퓨터학과 석·박통합과정. 관심분야는 소셜 네트워크 서비스, 빅 데이터, 추천 시스템, 온톨로지, 자가적응 소프트웨어

백 두 권

1974년 고려대학교 수학과 학사. 1977년 고려대학교 산업공학과 석사. 1983년 Wayne State Univ. 전산학과 석사. 1985년 Wayne State Univ. 전산학과 박사. 현재 고려대학교 정보대학 컴퓨터학과 교수. 1989년~1991년 고려대학교 전산과학과 학과장 1990년~1991년 미국 Arizona 대학교 객원 교수. 1991년~2013년 ISO/IEC JTC1/SC32 전문위원회 위원장. 1993년~1999년 한국과학기술원 객원책임연구원. 1993년~1999년 한국DB진흥센터 표준연구위원. 1996년~1997년 고려대학교 컴퓨터과학기술연구소 초대소장. 1997년~1998년 고려대학교 정보전산원 원장. 1998년~1999년 한국정보과학회 전산교육연구회 운영위원장. 1999년~2001년 정보통신진흥협회 데이터기술위원회 의장. 2002년~2004년 고려대학교 정보통신대학 초대학장. 2002년~2003년 한국시뮬레이션학회 회장. 2003년~현재 정보통신부 컴퓨터프로그램보호위원회 위원. 2004년~2005년 한국정보처리학회 부회장. 2005년~2008년 한국소프트웨어진흥원 이사. 2009년~2010년 고려대학교 정보통신대학 학장. 관심분야는 메타데이터, 소프트웨어공학, 데이터공학, 컴포넌트 기반 시스템, 메타데이터 레지스트리, 프로젝트 매니지먼트