지능정보연구 제19권 제1호 2013년 3월(pp.111~123) http://dx.doi.org/10.13088/jiis.2013.19.1.111

의미간의 유사도 연구의 패러다임 변화의 필요성 - 인지 의미론적 관점에서의 고찰

최영석
서울대학교 경영대학 박사과정
(aquinas9@snu.ac.kr)

박진수
서울대학교 경영전문대학원
(jinsoo@snu.ac.kr)

개념간의 의미적 유사도 및 관계도(Semantic Similarity/Relatedness)를 구하는 연구는 고전적인 연구에서는 데이터베이스 통합이나 시스템 통합, 그리고 현대의 연구에 있어서는 태그 및 키워드 추출, 연관 단어 추천 등에 걸쳐 다양한 분야에서 활용되어 온 연구이다. 그 연구는 역사가 오래되었을 뿐만 아니라, 경영정보와 컴퓨터 공학, 계산 언어학에 걸쳐 여러 분야에서도 많은 관심을 가져왔던 연구 분야라고 할 수 있다.

그러나, 지금까지의 개념간의 관계도 계산 방식은 미리 만들어진 사전이나 참조할 수 있는 다른 시맨틱 네트워크(Semantic Network)를 이용하여 계산하는 방법이 주를 이루었다. 이러한 접근 방법의 경우, 개념간의 의미적 관계가 변화에 대한 가능성을 고려하지 않는 것이 일반적이다. 하지만, 정보 기술의 발달과 빠른 사회변화는 개념간의 의미관계 등에 변화를 가져오고 있는 것이 현실이다. 사회적으로 일어나는 사건이나, 문화적 변화 등이 개념간의 의미관계를 변화시키는 것은 물론이며, 이러한 변화가 정보 통신 기술의 도움으로 빠르게 공유되고 있다. 이렇게 개념간의 의미 관계가 시간이나 맥락에 따라 빠르게 변화할 수 있는 가능성이 있음에도 불구하고, 기존의 개념간 의미적 유사도 및 관계도에 대한 연구들은 이러한 '의미관계의 변화'에 대한 새로운 문제에 대해 해답을 제시하지 못한 것이 사실이다.

따라서, 본 연구에서는 개념간의 유사도 연구에 있어 지금까지 있어왔던 '정적인 의미간 관계도 패러다임'에서 '동적인 의미간 관계도 패러다임'으로의 전환의 필요성과 그 당위성을 인지 의미론적(Cognitive Semantics)의 관점에서 역설하고자 한다. 인간이 인지하는 개념간의 의미관계가 변화할 수 있는 이론적 근거를 인지 의미론에서 찾아봄으로써, 패더다임 변화의 방향을 구체적으로 제시하였다. 또한 이러한 패러다임의 변화에 맞추어 개념간의 의미적 유사도 및 관계도에 대한 연구가 어떠한 방향으로 나아가야 할지 구체적인 연구 방향을 제시함으로써 관련 연구자들에게 새로운 연구의 가이드라인을 제시하였다.

논문접수일 : 2013년 03월 07일 논문수정일 : 2013년 03월 12일 게재확정일 : 2013년 03월 13일
투고유형 : 학술대회우수논문 교신저자 : 박진수

1. 서론

의미간의 유사도 및 관계도를 계산하는 연구는 그 활용 범위도 넓을 뿐만 아니라 오랜 역사를 갖는 연구이다. 데이터베이스나 시스템 통합 과정에서 일어나는 의미간의 비교 및 병합 과정 등을 진행하기 위해서는 반드시 거쳐야 한다. 이를 위해 다양한 개념간의 유사도를 계산하는 방식들이 연구되었으며 경영 정보 연구 분야뿐만 아니라, 컴퓨터 공학과 계산 언어학(Computational Linguistics) 분야에 걸쳐 많은 연구가 이루어졌다.

근래에는 웹 서비스가 활성화 되면서, 다양한 분

야에서 개념간의 유사도 및 관계도를 계산하는 방법이 필요하게 되었다. 웹 커뮤니티에서의 키워드 추천이나 연관어 자동 제공, 시맨틱 검색, 유사한 글 목록 추천 등 다양한 범위에서 시맨틱 네트워크 등을 이용해 의미나 쿼리간의 유사도를 이용해 특정 기능을 수행하는 빈도가 급증하고 있다(Kim et al., 2011; Cho and Kim, 2011; Park et al., 2011).

반면 최근까지의 개념간의 유사도를 구하는 연구는 과거의 연구 범위에서 크게 벗어나지 못하고 있는 것이 현실이다. 참고가 될 수 있는 사전이나 시맨틱 네트워크 등을 이용해 과거 연구들을 부분적으로 개량해 기존의 접근 방식들의 단점을 개선해 나아가는 연구들이 주를 이루고 있다.

그러나, 개념간의 의미관계를 계산하는 것은 인간이 두뇌를 이용해 두 가지 개념을 비교하는 인지 과정을 일련의 계산과정으로 대체하는 것이라고 할 때 기존의 연구들의 진부함은 되새겨 볼 만한 가치가 있다고 할 수 있겠다. 인간의 두뇌에 있는 인지 구조는 환경의 영향에 의해 빠른 속도로 변화하고 있고, 정보 통신 기술의 발달은 다양한 사건이나 문화적 변화를 통해 이루어지는 새로운 지식들을 보다 빠르게 많은 사람들이 공유할 수 있게 하였다. 이런 일련의 과정으로 벌어지고 있는 인간의 인지 범위의 역동적 확장은 개념간의 유사도를 도출하는 연구에도 반영될 수 있어야 보다 정확하게 다양한 분야에 적용될 수 있을 것이다.

따라서, 본 연구에서는 '정적인 의미간 관계도 패러다임'에서 벗어나 '동적인 의미간 관계도 패러다임'으로의 연구 패러다임 전환의 필요성을 역설해 보고자 한다. 이러한 패러다임 전환의 당위성을 살펴보기 위해 인지 의미론(Cognitive Semantics)적 관점에서 인간의 개념에 대한 인지 과정을 자세히 살펴보고자 한다.

2. 문헌연구-정적인 의미간 유사도 패러다임하의 연구들

의미적 유사도 및 관계도 관련 많은 연구들은 인간의 지식 체계를 네트워크 형태로 나타낸 시맨틱 네트워크를 이용해 개념간의 유사도를 계산하였다. 이런 이유로, 네트워크 형태로 잘 표현된 인간의 지식을 이용해 의미간의 유사도를 구하는 여러 방법들을 위상적 방법(Topological method)라고 부르기도 한다. 이 범주에 들어가는 대부분의 방법론들은 정적인 의미간 유사도 패러다임에 근거해 개념간의 유사도를 계산했다고 간주할 수 있다.

위상적 방법은 다양한 형태의 시맨틱 네트워크(위계적 택소노미(Hierarchical taxonomy)도 포함)를 기반으로 개념간의 관계도 및 유사도를 계산한다. 네트워크상의 노드(Node)는 개념을 나타내며, 이 노드간의 의미적 유사도를 측정하는 방법은 노드 기반의 접근(Node-based Approach)과 엣지 기반의 접근(Edge-based Approach)으로 나뉠 수 있다. 이 두 방법을 각각 정보 및 컨텐츠 기반의 접근(Information Content Approach)과 개념적 거리에 의한 접근(Conceptual Distance Approach) 방법으로 불리기도 한다.

2.1 노드 기반의 접근 방식

노드 기반의 접근 방식는 시맨틱 네트워크나 택소노미에서 두 개념이 얼마나 많은 정보를 공유하는지에 기반해 개념간의 유사도를 계산한다. 이것이 노드 기반의 접근 방식을 정보 및 컨텐츠 기반의 접근 방식이라고도 부르는 이유이다. 대부분의 노드 기반의 접근 방식은 정보 이론을 기반으로 한 Resnik의 아이디어를 기반으로 하고 있다(Resnik,

1995). 정보 이론에 따르면, 개념 C가 포함하고 있는 정보 및 컨텐츠(Information Contents, IC)는 아래와 같이 정량화 될 수 있다.

$$IC(C) = \log^{-1}P(C),$$

여기서 $P(C)$는 개념 C가 발견될 확률을 지칭한다. 이 개념에 기반해 두 개념간의 유사도는 다음의 공식으로 정의된다 :

$$sim(c_1, c_2) = \max [IC(c)], c \in Sup(c_1, c_2),$$

여기서 $Sup(c_1, c_2)$는 두 개념 c_1과 c_2를 모두 포함하는 포섭자(Subsumer)의 집합을 의미한다. Resnik은 오직 최저 공통 포섭자(lowest common subsume, lcs)가 포함하고 있는 정보 컨텐츠만을 고려했다는 특징이 있다. 최저 공통 포섭자는 두 개념간의 최단 경로 거리를 계산할 수 있는 어휘적 택소노미(lexical taxonomy)를 제공하는 역할을 하게 되는 것이다.

Resnik의 정보 및 컨텐츠 기반 접근 방식을 토대로, Lin은 개념간의 유사도를 측정하는 유사도 이론(Similarity Theorem)을 제안했다(Lin, 1998). 두 개념 A와 B의 유사도는 두 개념의 공통성을 기술하는데 필요한 총 정보, 그리고 A와 B를 완벽하게 기술하는 데 필요한 정보의 비유로서 정의하게 된다.

$$sim(A, B) = \frac{\log \ P(common(A, B))}{\log \ P(description(A, B))}$$

Couto et al.(2007)이 제안한 GraSM(Graph-based Similarity Measure)은 유전자 온톨로지 GO (Gene Ontology)에서 노드 기반의 접근 방식을 이용해 개념간의 의미적 유사도를 계산하는 방식이다. Couto et al.(2007)은 단순히 트리구조를 분석하는 것이 아니라, 유전자 온톨로지에 있는 모든 정보를 이용해서 개념간 유사도를 구한다는 점이 다르다.

2.2 엣지 기반의 접근 방식

엣지 기반의 접근 방식은 시맨틱 네트워크를 이용하는 방식 중에 가장 직관적이고 자연스러운 접근 방식이라고 할 수 있다. 이 접근 방식은 두 개념을 상징하는 시맨틱 네트워크상의 노드간의 거리를 측정하여 개념간의 유사도를 측정하는 방식이다.

Rada는 이러한 접근 방식을 검색 시스템인 Medline의 정보를 이용해 만든 위계적 네트워크(Hierarchy Network)인 MeSH(Medical Subject Headings)에 적용하였다(Rada et al., 1989; Rada and Bicknell, 1989). 그들의 주된 가정은 시맨틱 네트워크 상에서 두 개념간의 거리는 그 두 개념의 유사도를 상징한다는 엣지 기반 접근 방식의 대전제를 따르고 있다. 이러한 전제를 기반으로, 그들은 문서 Doc와 쿼리 Q와의 거리는 다음과 같이 정의된다.

$$DISTANCE(Doc, Q) = \frac{1}{mn} \sum_{t_i \in Doc} \sum_{t_j \in Q} d(t_i, t_j),$$

여기서 $d(t_i, t_j)$ 는 t_i에서 t_j까지 연결하는 최단의 엣지 수를 지칭한다.

2.3 시맨틱 네트워크 기반 두 접근방식의 공통점

의미적 유사도를 구하는 대부분의 방식은 시맨틱 네트워크를 기반으로 거리를 측정하거나 혹은 정보 공유량을 측정하는 방식으로 유사도를 계산하고 있음을 알 수 있다. 하지만, 이 방식들은 이미 만들어져 있는 시맨틱 네트워크를 기반으로 의미적인 유사도를 계산하는 방식이다. 또 많은 연구들이

이용하는 WordNet이나 유의어 사전(Thesaurus) 등은 의미간의 관계가 변화하는 것을 반영하기에는 많은 한계가 있다.

이를 이용한 대부분의 연구들은, 어떻게 주어진 시맨틱 네트워크를 이용해 합리적으로 의미적 유사도를 구할지에 초점이 맞춰져 있다. 이용하는 시맨틱 네트워크가 인간의 인지구조를 가장 잘 표현하고 있다는 가정 하에 계산 방식의 개선에만 초점을 맞춰온 것이 사실이다. 하지만 더 중요한 것은 인간의 인지구조를 잘 반영하는 시맨틱 네트워크 혹은 이와 동일한 기능을 하는 레퍼런스를 어떻게 구현하고 더 나아가 이 구현된 구조화된 지식을 어떻게 잘 활용할 수 있을 지를 고민하는 것이 필요하다.

더욱이 정보통신기술의 발전과 지식 공유체계의 다양화로 의미간의 관계 변화가 일어나게 된 경우 이러한 변화가 빠른 속도로 대중들에게 전파되고 실질적인 인지구조로 자리잡게 되었다(의미간 관계 변화가 일어나게 되는 기전과 원리는 다음 장에서 기술하기로 한다). 따라서 이런 상황을 모두 반영할 수 있는 연구 패러다임의 대두가 필요한 상황이라고 할 수 있겠다.

3. 의미 관계의 역동성 - 인지 의미론적 관점

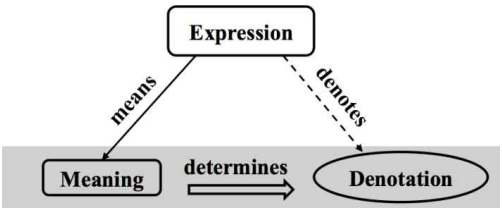
앞서 언급했듯이, 개념간의 의미적인 유사도를 계산하는 것은 인간이 개념간의 내포된 의미(intended meaning)를 비교하는 사고과정을 계산적인 프로세스를 이용해 모사하는 작업으로 볼 수 있다. 의미간의 관계도 및 유사도를 계산하는 가장 전형적인 방법은 앞서 언급되었듯이 시맨틱 네트워크를 기반으로 한 연구이다.

많은 연구들이 시맨틱 네트워크를 이용하는 이유는, 시맨틱 네트워크가 인간의 사고에 있는 개념간의 관계를 표현하기 때문이다. 시맨틱 네트워크의 대표적인 예는 온톨로지나 택소노미 등을 들 수 있겠다. 이들은 세상에 존재하는 개념과 그들간의 관계를 네트워크 형태로 표현하여 인간의 인지구조를 모사하고 있다. 이 시맨틱 네트워크 내에서의 두 개념의 관계도를 계산하는 작업은 결과적으로 인간이 두 개념을 비교하는 과정을 모사한 것이다.

그런데, 인간의 인지가 변화하는 상황에서는 기존에 인지했던 개념간의 관계도 변화할 수 있다. 이를 설명하기 위해 의미론에 대한 인지적 접근, 즉 인지 의미론을 도입할 필요가 있다.

3.1 인지 의미론에서의 개념의 인식 방법

인지 의미론은 언어활동에 사용되는 인간의 언어 지식이 인지의 한 부분이라는 대 전제에서 출발한다(Saeed, 2003). 다시 말해, 언어 지식은 인간의 일반적인 사고인 인지와 분리할 수 없다는 것이다. 따라서, 인지 의미론은 낱말의 의미가 어떻게 인간의 인지 영역 내에서 기술되는지를 설명하며, 그에 따라 특정 대상이 어떠한 인지 과정을 통해 특정한 표현으로 기술되는지를 설명한다(Löbner, 2002). Löbner는 그의 책인 '의미론의 이해(Understanding Semantics)'에서 인지 의미론(Cognitive Semantics)의 초점을 '의미 삼각형(semiotic triangle)'(Ogden and Richards, 1923)을 이용해 아래 <Figure 1>과 같이 표현했다.



<Figure 1> The Focus of Cognitive Semantics

인간의 개념에 대한 인지는 인간이 가지고 있는 경험과 지식 등으로 구축된다. 인간은 경험과 지식을 기반으로 개념의 내포 및 외연 영역을 사고하며, 이 영역의 포함관계 등을 기반으로 두 개념의 의미 연관을 추리한다. 만약 인간이 가지고 있는 경험과 지식이 변화하고 또 확장된다면 인간의 개념에 대한 인지에는 어떠한 변화가 일어날까? 바로 이점이 인간이 인지하고 있는 개념간의 의미적 관계도가 변할 수 있음을 암시한다.

전통적 의미론에서는 '의미적 지식(semantic knowledge)'을 '세계 지식(world knowledge, 개인적 지식과 문화적 지식을 포함)'과 구분 가능하다고 판단하는 반면, 인지 의미론의 관점에서는 이러한 구분이 어렵다고 판단하고 있다.[1] 개념의 의미를 인지하기 위한 의미적 지식과 후천적으로 습득된 세계 지식을 구분하기 어렵다고 생각하는 탓이다. 인지 의미론에서는 '인지'라고 하는 일반적인 사고 영역을 통해 의미를 판단한다고 가정한다. 이들 중 어느 관점을 택하더라도, 인간이 사물을 인지하는 데 필요한 지식은 수시로 변하고 추가될 수 있다는 가능성을 보게 된다.

인지 의미론에서 '의미'는 관습화된 개념 구조에 기초한다고 주장한다. 다시 말해, 인간의 두뇌 속에 의미 구조는 세상에서의 성장과 활동을 통한 경험을 통해 형성된 정신적 범주를 반영한다는 것이다. 이는 개인의 성장과 경험으로 생긴 인지 구조의 변화가 결국 의미 구조의 변화로 이어질 수 있다는 것이다.

전통적인 의미론의 관점에서도 이와 유사한 논의를 확인할 수 있다. 특정한 문화적 경험을 공유하는 집단이 갖는 문화적 지식은 개인이 세상을 이해함에 있어 매우 중요한 역할을 한다(Holland and Quinn, 1987). 이러한 문화적 지식은 다양한 사회 현상에 의해 생성되고 공유될 수 있기 때문에 집단의 인지 체계 변화에 영향을 줄 수 있다. 더욱이 IT의 발전으로 인해, 지식을 공유하는 데에 있어 효율성이 높아지고 있을 뿐 아니라 보다 다양한 지식 공유의 수단이 등장하고 있다. 따라서 문화적 지식은 과거에 비해 빠른 속도로 공동체 구성원에게 공유되고 있다.

3.2 세계 지식을 통한 개념의 인지와 문화적 지식

세상에 대한 이해를 돕는 세계 지식은 앞서 언급한 바와 같이 문화적 지식과 개인적 지식으로 구성된다. 이 문화적 지식과 개인적 지식으로 세상을 인지하는 방식을 다음과 같이 예를 들어볼 수 있다. 우리 모두는 '포도'에 대해 몇 가지 사실들을 알고 있다. 최근에 먹었던 포도가 어떠한 맛이었는지, 또 이 포도를 어디서 구입할 수 있는지, 자신의 과일에 대한 기호에서 포도는 어느 정도로 선호되는지 등의 사실을 알고 있을 것이다. 이러한 사실들은 대부분 개인의 경험을 통해 습득하게 된 개인적 지식이라고 할 수 있다.

반면, 포도는 과일의 일종이라는 사실과 비타민이 함유되어 있다는 사실, 그리고 포도의 알려진 효능, 그리고 포도는 어디에서 주로 팔리는 지 등 개인의 경험이 아닌 사회 구성원 모두가 공감하는 사실이 있다. 바로 이러한 종류의 지식을 문화적 지식이라고 할 수 있다.

다시 말해, '포도'라고 하는 개념은 개인적인 경험에 의한 지식과 문화적으로 습득된 지식 모두로서 개인에게 인지되는 것이다. 개인적인 지식은 개인의 경험을 통해 각기 다른 방식으로 형성되는

1) 인지 의미론에서도 '세계 지식'이라는 용어를 사용하지만, '의미적 지식'과 구분되는 개념으로 사용하지 않는다.

반면, 문화적 지식은 공동체에 의해 형성되며 그 공동체의 특징을 반영하게 된다. 동일한 동물이나 사물이 종교 공동체나 언어 공동체에 따라 각기 다른 의미를 가지고 있는 것이 이에 해당한다.

문화적 지식은 공동체가 공유하는 사회적 경험과 사건, 그리고 관습 등에 의해 형성된다. 새로운 사건이나 공동체의 문화적 변화는 인간의 문화적 지식의 변화를 일으키게 될 것이다. 결과적으로 인지 의미론적 관점에서 이러한 변화는 사물을 인식하는 인지 방식에도 변화를 가져올 수 밖에 없는 것이다.

3.3 원형 이론과 사물의 범주화

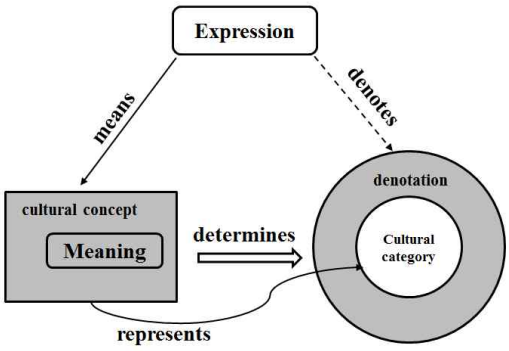
인지 언어학(Cognitive Linguistic)에서 지적하는 인간의 개념 범주화 과정을 살펴볼 때도 개념간의 의미적 관계가 변할 수 있는 가능성을 알 수 있다. 원형 이론(prototype theory)은 특정 개념의 외연들을 어떻게 인간이 인지적으로 범주화(semantic categorization)하고 있는지에 대한 이론이다(Rosch, 1973; Rosch, 1975). 앞서 논의한 인지 의미론적 관점은 사물을 어떻게 인식하고 이해하는지를 설명하고 있다면, 원형 이론은 이렇게 인식된 여러 사물들이 인간의 사고 과정에서 어떻게 범주화하는지를 논의하는 것이다. 일반적으로 두 개념간의 '의미적 관계(Semantic Relation)'를 주어진 개념간의 상하 관계나 포함 관계 등을 통해 살피는 점을 감안할 때, 원형 이론은 의미적 관계의 변화에 있어 보다 직접적인 근거를 제공할 수 있다. 원형 이론의 범주화의 결과가 결국 의미적 관계로 이어질 수 있기 때문이다.

범주화의 과정은 원래 가지고 있던 지식 이외에 새롭게 생성된 지식에 의해서 새로운 범주화를 가능하게 하는 경우가 많다. 예를 들어, 'MP3 Player'의 범주에 포함되는 사물을 선별한다고 가정해 보자. 과거 휴대전화가 MP3 Player의 기능을 하지 않았던 시대에는, 대부분의 사람들이 휴대폰을 MP3 Player의 범주라고 생각하지 않았을 것이다. 하지만, 기술의 발전으로 휴대전화가 MP3 Player의 기능을 갖게 되어 휴대폰 역시 MP3 Player의 기능적 범주로 포함시킬 수 있게 되었다.

Löbner는 이러한 개념의 외연을 범주화함에 있어서 모호함(fuzziness)을 야기할 수 있는 원인 중 하나를 '언어를 사용하는 공동체 내에서 단어의 의미 변화(variation of word meanings within a language community)'라고 지적하고 있다. 집단에 따라 공유하는 경험이 각기 다르게 되면, 사물의 분류 방식이나 개념간의 관계에 대한 인식이 집단에 따라 달라질 수 있다고 지적하고 있다(Löbner, 2002).

3.4 의미론과 개념의 의미 구조 변화

앞선 논의들을 바탕으로 인지 의미론이 문화적 지식의 통합을 통해 일어나는 변화를 의미 삼각형을 이용하여 표현하면 <Figure 2>와 같이 표현할 수 있다.



<Figure 2> The Semiotic Triangle Integrating Cultural Knowledge(Löbner, 2002)

문화적인 지식으로 인해 새롭게 생기거나 변동되는 인간의 인지가 새로운 방식의 범주화 과정을 야기하게 되며, 이로 인해 개념 간의 범주화 방식 및 의미적인 관계가 변화할 수 있다는 것이다. 다시 말해, 전통적인 의미론적 관점에서뿐만 아니라 의미론의 관점에서 인간의 인지의 변화에 의해 개념의 의미 구조가 변화할 수 있다는 가능성을 확인할 수 있다. 또한 개념간의 관계를 판단하게 하는 인간의 인지, 특히 문화적 지식의 영역이 IT 환경을 통해 빠르게 공유되고 있는 상황이기 때문에, 변화하는 의미간의 관계를 반영할 수 있는 새로운 접근 패러다임이 필요하다.

4. 동적인 의미간 유사도 패러다임으로의 전환과 연구 방향

앞 장에서 논의된 바와 마찬가지로, 개념간의 유사도나 연관성 등 관계에 대한 논의를 위해서는 그 관계가 가질 역동성에 대한 부분을 보다 중요하게 생각해야 한다. 기존 시맨틱 네트워크를 이용하는 연구들은 주어진 시맨틱 네트워크을 의심 없이 받아들이는 경향이 다분하다. 다시 말해, 주어진 네트워크가 당연히 인간의 인지를 잘 반영하고 있다고 가정하고 연구가 시작되기 때문에 시맨틱 네트워크 이용 방법론에 초점이 맞추어 진 채로 많은 연구들이 알고리즘적 보안에만 초점을 맞추어 왔다. 또 많지는 않지만 시맨틱 네트워크를 이용하지 않는 접근 방식의 경우에도 지식 베이스(Knowledge Base)가 실제 인간의 인지를 정확하게 표현하고 있는지 크게 신경 쓰지 않는 것이 일반적이다.

따라서 본 논문에서는 이러한 기존의 의미간 유사도에 대한 연구가 동적인 의미간 유사도 패러다임으로 전환하기 위해서 나아가야 할 연구 방향을 제시해 보고자 한다. 여기서 제시하는 동적인 의미간 유사도 패러다임이란, 의미간의 유사도나 관계가 시간이나 문맥적 변화에 의해 변화될 수 있다는 전제를 염두에 둔 연구 방향으로서, 의미간 유사도의 역동성을 감안한 연구 방식을 의미한다.

4.1 코퍼스(Corpus) 기반의 통계적 접근(Statistical Approach)을 통해 의미간의 유사도를 구하는 연구

새롭게 생겨나고 인간에 의해 공유되는 정보들을 가장 빠르게 읽어낼 수 있는 부분은 인간이 발화하거나 쓰는 언어로 나타나기 마련이다. 이런 이유로 언어학에서는 코퍼스를 인지 언어학의 주요 관찰 분야로 간주해 왔으며 '코퍼스 언어학(Corpus Linguistics)'라는 분야가 새로이 자리잡았을 정도로 관심이 있는 분야이다.

코퍼스의 학문적인 정의는 명시한 예가 없으나 일반적으로 사람들이 실제로 사용하는 문장들을 수집해 놓은 대용량의 문서 뭉치를 '코퍼스'라고 지칭한다. 이러한 코퍼스들을 분석하면 실제 동시대에서 동일한 문화 및 언어를 사용하는 사람들의 인지구조를 파악하는데 큰 도움이 될 뿐만 아니라, 특정 기간의 코퍼스들의 분석을 통해 그 기간에 국한되어 있는 의미 구조들의 생성과 소멸을 이해할 수도 있다.

특히 IT와 모바일 환경에서 쓰여진 문장이나 글들을 코퍼스로 수집하는 방법은 앞서 언급했던 인간의 경험적 지식이나 문화적 지식을 가장 확실하게 수집하는 방법이라 할 수 있다. 특정 사건이나 문화적 격변으로 인해 개념간의 새로운 의미 관계가 생기거나 변질되는 것이 실제 사람들의 언어 사용에서 발견되기 때문에, 모바일 환경이나 웹 사이트에 쓰여진 문장들을 수집하여 분석하는 것은 최신의 사회적 인지현상을 분석하는 방법이 될 수

있을 것이다.

전통적으로 언어학에서 코퍼스를 분석하는 방식들은 통계적 접근을 활용하는 것이 일반적이었다. 특정한 명사나 술어, 혹은 특정한 두 명사가 동시에 한 문장에서 발화되거나 인접하여 발화되는 경우들을 산정해 이를 통계적으로 분석하는 방식들을 의미한다.

개념간의 유사도를 구하는 연구에 있어서도 이와 같이 코퍼스를 수집하고 이를 통계적으로 분석해 관계도를 계산해 내는 방식이 매우 유용할 수 있다. 앞서 논의했던 시맨틱 네트워크를 이용하는 경우보다는 훨씬 더 융통성 있는 방법론을 개발할 수 있을 뿐만 아니라, 동적으로 변화하는 의미간의 관계들을 확인하는 가장 확실한 방법이 될 수 있을 것이다. 특히 대부분의 사람들이 웹을 사용하고 또 이를 통해 정보를 습득하고 지식을 쌓아가고 있는 현실을 감안할 때, 코퍼스는 인간의 인지구조를 모사하기에 더 없이 좋은 원천이 될 수 있을 것이다.

4.2 동적으로 생성된 시맨틱 네트워크를 기반으로 의미간의 유사도를 구하는 연구

시맨틱 네트워크를 이용하는 의미간 유사도 계산 연구에 있어 시맨틱 네트워크를 업데이트하는 방식이나 버져닝하여 최신의 정보로 관리하는 방식보다 더 근본적인 방법은 최신의 지식과 정보를 이용하여 시맨틱 네트워크를 새롭게 구성하는 방식이 될 수 있다. 기존에 구성되어 있는 시맨틱 네트워크를 최신의 지식으로 재구성하는 일이 어려운 경우에는 시맨틱 네트워크를 새롭게 구성하는 방법이 필수적으로 필요하다. 또한 시맨틱 네트워크의 기반이 되는 지식이 매우 빠르게 생성 및 변화하거나, 그 지식이 특정 분야에 국한되어 지식의 양이 비교적 적은 경우 역시 이러한 방식이 유용할 수 있다.

새롭게 시맨틱 네트워크를 구성하는 방식은 크게 두 가지로 나뉠 수 있다. 지식 베이스를 이용해 완전히 새로운 네트워크를 구성하는 방법이 가장 기본적인 방법일 것이다. 다른 한 가지 방법은 기존에 존재했던 복수의 시맨틱 네트워크를 재구성해서 새로운 네트워크를 만드는 방식이다. 이 방법은 새로운 네트워크를 구성하는데 필요한 요소를 기존의 네트워크에서 효율적으로 추출해 병합하는 과정이 매우 중요하다. 온톨로지의 경우 온톨로지 추출(Ontology Extraction)과 온톨로지 병합(Ontology Merging)을 통해 새로운 온톨로지를 구성하는 다양한 방법론들이 연구되고 있다.

특히, 시맨틱 웹(Semantic Web) 기술의 발전으로 인해 많은 정보들이 RDF 트리플의 형태로 데이터베이스에 저장되어 있는 경우가 많아지면서, 이를 기반으로 새롭게 온톨로지 등을 구축하는 방법론도 가능할 것으로 보인다. 지식 베이스가 네트워크의 형태로 변환 가능한 RDF로 되어있는 경우, 기존 텍스트나 테이블 형태로 되어 있는 지식 베이스에 비해 비교적 수월하게 네트워크로 변환할 수 있기 때문이다.

4.3 시맨틱 네트워크의 버져닝 및 업데이트 방법론을 포함하는 의미간의 유사도 연구

앞서 언급한 시맨틱 네트워크를 기반으로 하는 연구들의 특징은 시맨틱 네트워크에 대한 검증 절차를 거치지 않는다는 것이었다. 주로 사용되는 시맨틱 네트워크들은 특정 영역의 지식을 기반으로 만들어진 온톨로지(Ontology)나 워드넷(WordNet)과 같은 온톨로지 형태의 사전 등이 주를 이룬다. 따라서 이러한 시맨틱 네트워크의 변화 관리를 염두에 둔 연구가 필요하다고 할 수 있다.

물론, 이런 관리는 이미 온톨로지 버져닝(Onto-

logy Versioning)이나 온톨로지 관리(Ontology Management) 등의 연구에서 그 방안을 고안해 내고 있어 이러한 연구 영역을 의미간 유사도를 구하는 연구에까지 포함해야 하는지 회의적인 시각을 보일 수도 있다. 혹은 의미간 유사도를 연구하는 분야와 온톨로지 버져닝이나 온톨로지 관리분야에서 의미간의 유사도 구하는 연구와는 독립적인 영역이라는 시각도 존재할 수 있다.

하지만, 일반적인 온톨로지 버져닝이나 온톨로지 관리 등의 연구 분야에서는 주로 도메인이 정해져 있는 특정 분야의 온톨로지나 기업 운영에 필요로 하는 분야에 대한 특수한 지식을 기반으로 하고 있는 온톨로지를 대상으로 하는 경우가 대부분이다. 반면 키워드 추천이나 군집화, 연관어 추출 등의 경우에는 특정 분야가 아닌 일반적인 분야의 지식을 포괄하는 시맨틱 네트워크가 필요하다. 이러한 일반적인 지식은 사회의 변화에 따라 새롭게 생길 뿐 아니라 기존의 지식 및 구조 또한 쉽게 변한다. 더욱이 고전적인 시스템 통합 방법론에서 필요로 하는 의미간의 유사도 측정 또한 빠르게 진화하고 변화하는 비즈니스 환경을 고려할 때 결코 시맨틱 네트워크의 역동성과 의미간의 유사도를 구하는 연구를 분리해서 생각되어서는 안된다.

5. 결언

개념간의 유사도를 구하는 연구는 그 오래된 역사만큼이나 많은 연구가 있어 왔다. 그러나, 그간 연구되었던 성과들은 새로운 연구 방식이라기 보다는 과거 30여년 전의 연구 틀에서 미진한 진보를 이끌어온 노력의 일부라고 볼 수 있을 것이다. 아직까지도 Resnik이나 Rada의 시맨틱 네트워크 기반의 연구는 개념간의 유사도를 연구하는데 있어 기본이 되는 논리이며 이론적인 배경을 제공하고 있는 것은 사실이다. 하지만, 정보 기술의 발달과 정보 공유 체계의 확산은 과거에 비해 인간의 인지체계를 빠른 속도로 변화시키고 있는 것이 사실이다. 이러한 점에서 과거 연구들은 이와 같은 변화를 가정하고 인지하지 못했다고 평가할 수 있을 것이다. 이런 이유로 본 논문에서 과거의 개념간 유사도 연구들을 '정적인 개념간의 유사도 패러다임' 하에서 연구했다고 칭한 것이다.

본 연구에서 주장하고자 하는 '동적인 개념간의 유사도 패러다임'의 기본 전제는 서론에서 밝힌 바와 같이 세상의 변화와 그 변화에 대한 정보 및 지식을 인지하고 있는 인간의 인지구조를 보다 정확하게 즉시적으로 반영할 수 있는 개념간의 유사도 계산 방식이 필요하다는 것이다. 이러한 패러다임에 맞는 연구 방향으로 앞 장에서는 세 가지 가능한 연구 방향을 제안하였으며, 이 세 가지 연구 방향 이 외에도 보다 많은 연구들이 앞으로 진행될 것으로 믿는다.

의미간의 유사도 연구는 연구의 이론적인 기반이 매우 두텁다고 할 수는 없는 연구이지만, 그 활용 분야가 매우 광범위할 뿐만 아니라 기여점이 확실한 연구 분야이다. 정보통신 사회의 발전과 사회적 지식의 변화를 반영할 수 있는 연구 패러다임으로의 전환은 분명 관련 연구의 새로운 기회가 될 것이라고 확신한다.

참고문헌

Cho, I. D. and N. K. Kim, "Recommending core and connecting keywords of research area using social network and data mining techniques", *Journal of Intelligence and Information Systems*, Vol.17, No.1(2011), 127~138.

Couto, F. M., M. J. Silva, and P. M. Coutinho, "Measuring semantic similarity between Gene Ontology terms", *Data and Knowledge Engineering*, Vol.61, No.1(2007), 137~152.

Holland, D. and N. Quinn, (Eds.), *Cultural models in language and thought*, Cambridge : Cambridge University Press, 1987.

Kim, H. W., M. Sohn, and H. J. Lee, "Dyanamic decision making using social context based on Ontology", *Journal of Intelligence and Information Systems*, Vol.17, No.3(2011), 43~62.

Lin, D., "An information-theoretic definition of similarity", *in Proceedings of the 15th international conference on Machine Learning*, Vol.1, 296~304.

Löbner, Sebastian, 2002, *Understanding Semantics*. London : Arnold.

Ogden, C. K. and I. A. Richards, *The meaning of meaning : A study of the influence of language upon thought and of the science of symbolism*, 8th ed. Reprint New York : Harcourt Brace Jovanovich, 1923.

Park, J. S., N. W. Kim, M. J. Choi, Z. Jin, and Y. S. Choi, "Semantic search : A survey", *Journal of Intelligence and Information Systems*, Vol.17, No.14(2011), 19~36.

Resnik, P., "Using information content to evaluate semantic similarity in a taxonomy", *Proceedings of the 14th international joint conference on Artificial intelligence*, August Vol.20, No.25(1995), 448~453, Montreal, Quebec, Canada.

Rada, R. and E. Bicknell, "Ranking documents with a thesaurus", *Journal of the American Society for Information Science*, Vol.40, No. 5(1989), 304~310.

Rada, R., H. Mill, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets", *IEEE Transactions on Syst.* Man Cybern. Vol.19, No.1(1989), 17~30.

Rosch, E. H., "Natural categories", *Cognitive Psychology*, Vol.4(1973), 328~350.

Rosch, E. H., "Cognitive Reference Points", *Cognitive Psychology*, Vol.7(1975), 532~547.

Saeed, J. I., Semantics(Introducing Linguistics), Chichester : Wiley-Blackwell, 2003.

Abstract

The Need for Paradigm Shift in Semantic Similarity and Semantic Relatedness : From Cognitive Semantics Perspective

Youngseok Choi* · Jinsoo Park**

Semantic similarity/relatedness measure between two concepts plays an important role in research on system integration and database integration. Moreover, current research on keyword recommendation or tag clustering strongly depends on this kind of semantic measure. For this reason, many researchers in various fields including computer science and computational linguistics have tried to improve methods to calculating semantic similarity/relatedness measure.

The study of similarity between concepts is meant to discover how a computational process can model the action of a human to determine the relationship between two concepts. Most research on calculating semantic similarity usually uses ready-made reference knowledge such as semantic network and dictionary to measure concept similarity. The topological method is used to calculate relatedness or similarity between concepts based on various forms of a semantic network including a hierarchical taxonomy. This approach assumes that the semantic network reflects the human knowledge well. The nodes in a network represent concepts, and ways to measure the conceptual similarity between two nodes are also regarded as ways to determine the conceptual similarity of two words (i.e., two nodes in a network). Topological method can be categorized as node-based or edge-based, which are also called the information content approach and the conceptual distance approach, respectively. The node-based approach is used to calculate similarity between concepts based on how much information the two concepts share in terms of a semantic network or taxonomy while edge-based approach estimates the distance between the nodes that correspond to the concepts being compared. Both of two approaches have assumed that the semantic network is static. That means topological approach has not considered the change of semantic relation between concepts in semantic network.

However, as information communication technologies make advantage in sharing knowledge

* College of Business Administration, Seoul National University
** Corresponding Author: Jinsoo Park
Graduate School of Business, Seoul National University 1 Gwanak-ro, Gwanak-gu, Seoul 151-916, Korea
Tel: +82-2-880-9385, Fax: +82-2-877-0513, E-mail: jinsoo@snu.ac.kr

among people, semantic relation between concepts in semantic network may change. To explain the change in semantic relation, we adopt the cognitive semantics. The basic assumption of cognitive semantics is that humans judge the semantic relation based on their cognition and understanding of concepts. This cognition and understanding is called 'World Knowledge.' World knowledge can be categorized as personal knowledge and cultural knowledge. Personal knowledge means the knowledge from personal experience. Everyone can have different Personal knowledge of same concept. Cultural knowledge is the knowledge shared by people who are living in the same culture or using the same language. People in the same culture have common understanding of specific concepts. Cultural knowledge can be the starting point of discussion about the change of semantic relation. If the culture shared by people changes for some reasons, the human's cultural knowledge may also change. Today's society and culture are changing at a past face, and the change of cultural knowledge is not negligible issues in the research on semantic relationship between concepts.

In this paper, we propose the future directions of research on semantic similarity. In other words, we discuss that how the research on semantic similarity can reflect the change of semantic relation caused by the change of cultural knowledge. We suggest three direction of future research on semantic similarity. First, the research should include the versioning and update methodology for semantic network. Second, semantic network which is dynamically generated can be used for the calculation of semantic similarity between concepts. If the researcher can develop the methodology to extract the semantic network from given knowledge base in real time, this approach can solve many problems related to the change of semantic relation. Third, the statistical approach based on corpus analysis can be an alternative for the method using semantic network. We believe that these proposed research direction can be the milestone of the research on semantic relation.

Key Words : Semantic Relatedness, Semantic Similarity, Semantic Network

저 자 소 개

최영석

서울대학교 전기공학부를 졸업하였으며, 현재 서울대학교 경영대학에서 경영정보시스템 전공 박사과정에 재학 중이다. International Journal of Electronic Commerce (IJEC), 지능정보 연구, Information Systems Review 등 국내외 저널에 논문을 게재하고 있다. 주요 관심분야는 정보시스템 통합, 온톨로지, 시맨틱 웹, Network Analysis, 정보통신 산업 및 정책 등이 있다.

박진수

The University of Arizona에서 경영정보시스템을 전공하여 경영학 박사를 취득했으며, University of Minnesota의Carlson School of Management에서 조교수, 고려대학교 경영대학에서 조교수를 역임했다. 현재 서울대학교 경영전문 대학원/경영대학에 부교수로 재직 중이다. Journal of Database Management, International Journal of Principles and Applications in Information Science and Technology의 편집위원이며, 그의 논문은 MIS Quarterly, IEEE Transactions on Knowledge and Data Engineering (TKDE), IEEE Computer, ACM Transactions on Information Systems (TOIS), Data and Knowledge Engineering, Journal of Database Management, Expert Systems With Applications, Information Systems Frontiers, Communications of the AIS, Journal of Global Information Technology Management (JGITM), International Journal of Electronic Business, Asian Case Research Journal, Asia Pacific Journal of Information Systems, Information Systems Review, 지능정보연구 외 다수의 저널에 게재되었다. 주요 연구분야는 정보시스템 통합, 지식 경영, 온톨로지, 시멘틱웹 기반 혁신 기술이다.