2011년 2월
박사학위 논문

지능적인 웹 검색을 위한 의미적 문서 태깅 방법 연구

조선대학교 대학원

컴퓨터공학과

황 명 권

2011년 2월 박사학위논문 지능적인 웹 검색을 위한 의미적 문서 태깅 방법 연구

황 명 권

2011년 2월

박사학위 논문

지능적인 웹 검색을 위한 의미적 문서 태깅 방법 연구

조선대학교 대학원

컴퓨터공학과

황 명 권

지능적인 웹 검색을 위한
의미적 문서 태깅 방법 연구

Semantic Document Tagging Methods for Intelligent Web Retrieval

2011년 2월 25일

조선대학교 대학원

컴퓨터공학과

황 명 권

지능적인 웹 검색을 위한
의미적 문서 태깅 방법 연구

지도교수 김 판 구

이 논문을 공학박사학위신청 논문으로 제출함.

2010년 10월

조선대학교 대학원

컴퓨터공학과

황 명 권

황명권의 박사학위논문을 인준함

위원장　조선대학교 교수　이 성 주　(인)

위　원　조선대학교 교수　모 상 만　(인)

위　원　영남대학교 교수　정 재 은　(인)

위　원　대진대학교 교수　김 정 민　(인)

위　원　조선대학교 교수　김 판 구　(인)

2010년　12월

조선대학교 대학원

목 차

그 림 목 차

표 목 차

ABSTRACT

Semantic Document Tagging Methods for Intelligent Web Retrieval

Hwang, Myung Gwon
Advisor: Prof. Kim, Pan-Koo, Ph. D.
Department of Computer Engineering
Graduate School of Chosun University

Nowadays, the fast advance of digital technologies and the current Web environment have been accelerating the field of information retrieval and processing. The Internet space using the Web is not strange any more to most people and they can obtain any information desired from the Web. These changes have spawned a great deal of research aiming at enhancing service and convenience. Thus, many computer science researchers are committed to finding more useful and efficient methods to provide appropriate results to meet users' needs. Among those, the methods of this dissertation have been studied for semantic document tagging to realize Semantic Web as an ultimate purpose.

Semantic Web is a very important technique aiming at processing and understanding the information spread on the Web and subsequently providing semantic and exact retrieval results. To realize Semantic Web, this research concentrates on tagging methods of text documents. The amount of the texts

is increasing according to trend of Web 2.0 and it is the most frequently utilized communication medium to express and share information between people. Therefore, the text retrieval is important and this research proposes tagging methods of Web documents to provide standardized, systematic and semantic retrieval.

The previous works on Web document tagging generally choose core words from a document itself. However, the core words are not standardized taggers so, in retrieving, users should make an effort to grasp the tagger words first. To improve the point, this research contains methods to utilize titles (Wiki concept) of Wikipedia documents and to find the best Wiki concept which describes the Web documents (target documents). In addition to these methods, the research tries to classify target documents into Wikipedia category (Wiki category) for semantic document interconnections.

In order to use Wiki categories and concepts for classifying and tagging target documents, the research extracts context information from Wiki concepts, Wiki categories and target documents and finds the nearest Wiki categories and concepts of target documents through similarity measure. Experimenting diverse cases, it was confirmed that this research can provide semantic classification and tagging methods and that the context information of documents has much potentiality to be applied to various works for Semantic Web. By the way, it is worth noting that some future works, which can give semantics to proper nouns and technical terms, need to be done.

Ⅰ. 서 론

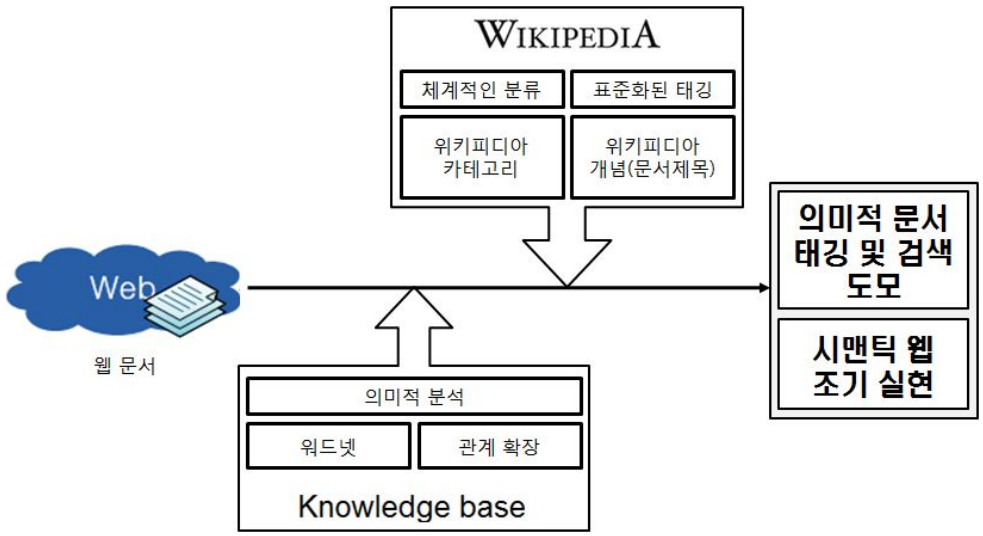
A. 연구 배경

1989년 영국의 컴퓨터 과학자 팀 버너스 리(Tim Berners-Lee)가 제안한 웹(World Wide Web)은 급격하고 꾸준한 발전을 거쳐 인간의 지식 저장과 여러 문화의 보고 및 정보 공유를 위한 공간으로 활용되고 있다[1]. 특히 2000년대에 웹의 지식화에 의해 그 부피는 급격하게 늘고 있다. 2003년 발표된 HMI(How Much Information)에 따르면 2000년 최소 2,132,238TB (이는 21만개의 미국 의회 도서관을 채울 수 있는 도서 분량과 같음)가 웹에 저장되어 있고, 매년 74.5%씩 증가하는 것으로 나타났다. 이에, 웹에 저장된 정보 검색에 대한 방법들은 매우 중요한 연구로 간주되고 있으며, 시맨틱 웹(Semantic Web)의 발전과 함께 의미적 검색(Semantic Retrieval)이 중요한 이슈로 떠오르고 있다. 이렇듯 현재는 다양한 정보를 체계적으로 정리하는 웹의 지식화 단계를 넘어 웹에 저장된 정보, 특히 웹 문서들을 의미적이고 지능적으로 검색하여 제시할 수 있는 시맨틱 웹(Semantic Web)에 집중되고 있다[2].

기존의 전통적인 방법인 키워드 기반 검색 방식으로는 검색자가 요구하는 깊고 방대한 지식을 의미적으로 찾아주기 어렵다. 이를 해결할 수 있는 시맨틱 웹은 크게 두 부류의 연구로 구성된다. 첫째는 사람이 갖는 지식체계(개념을 중심으로)와 비슷한 지식베이스(Knowledge Base)를 구축하는 것이고, 둘째는 그 지식베이스를 이용하여 웹에 있는 정보들을 처리하는 데 있어서 사람이 읽고 분석하는 것과 유사한 수준의 지능을 형성하는 것이다[16]. 지식베이스와 관련된 연구로는 온톨로지(Ontology) 확장[11, 12, 13], 지식 통합[15], 지식베이스 관계 확장[16] 등이 있으며, 지능적(의미적)으로 정보를 처리하기 위한 연구로는 의미적 문서 인덱싱(Indexing)[3, 4], 의미적 메타데이터 (Metadata) 생성[5, 6, 7], 자연어 처리 기반 의미적 웹 콘텐츠(Contents) 생성[8], 문서 주제 선정[9], 질의 확장[10], 의미적 유사도 기반 정보 검색[14] 등이 수행되었다. 이러한 연구들은 웹에 저장된 정보들의 의미적 분석 및 이해를 위한 방법들을 제안함으로써 최종적으로는 웹 자체적으로 시맨틱 검색을 제공하는데 기여하였다. 하지만 이들이 제시하는 방법이 문서 자체의 의미성, 문서와 문서, 질의어와 문서, 질의어와 질의어 사이의 단순한 의미성만을

고려하기 때문에 표준성, 체계성, 의미성을 함께 반영한 문서 검색에는 여전히 한계점이 존재한다.

시맨틱 웹을 위해 워드넷(WordNet)을 지식베이스로 다양하게 활용하고 있다. 하지만 워드넷의 한정된 개념들로는 사람들의 질의어, 문서 내의 모든 단어들을 처리하기 어렵다. 워드넷은 사람들 사이에 일반적으로 사용되는 단일 명사 또는 복합 명사만을 정의하고 있기 때문이다. 이에, 특정 도메인(Domain) 내의 전문 용어(Technical Terms), 고유명사(Proper Nouns) 등을 확장하기 위한 방법[11, 12]들이 연구되었다. 또한, 웹 지식화의 대표적인 문서 집합인 위키피디아(Wikipedia)의 문서 제목을 워드넷과 연결하는 연구[13]가 있었다. 그러나 이러한 연구들의 결과를 활용하기에는 신뢰성 측면에서 검증받을 수 없다는 한계점이 존재한다. 이에 본 연구는 문서를 체계적으로 분류하고 의미적인 주제로 태깅(tagging)을 통해 표준화된 검색을 제공하는 것에 주안점을 두었다.



[그림 1] 연구의 전체 개요

[그림 1]은 본 연구의 전체 개요를 보이고 있다. 웹에 존재하는 무수한 문서들을 체계적 의미적으로 분석하고, 표준화된 검색의 제공을 위한 본 연구는 위키피디아 내의 카테고리와 위키피디아의 문서 제목을 웹 문서의 분류와 태깅의 표준화된 정보로 활용한다. 그리고 각 문서의 핵심이 되는 문맥 정보(Context Information)를 확장된 워드넷을 이용한 의미적

방법으로 추출한다. 이러한 정보들을 이용하여 웹에 존재하는 문서들을 위키피디아 카테고리로 분류하고 위키피디아 문서 제목으로 태깅하는 방법을 제안하고 실험함으로써 그 결과를 평가한다. 본 연구는 문서에 포함된 어휘들을 기반한 의미적 검색, 위키피디아 카테고리를 이용한 문서 분류, 위키피디아 문서 제목을 이용한 표준화된 태깅 및 문서들 사이의 의미적 네트워크를 제공함으로써 시맨틱 웹 실현을 도모하고자 한다.

B. 연구 목적

본 연구는 웹에 존재하는 많은 문서들을 의미적으로 태깅할 수 있는 방법을 제안하며, 본 연구를 통해 추구하는 궁극적인 목적은 다음과 같다.

- 위키피디아 문서 제목으로의 웹 문서 태깅을 통해 표준화된 검색 질의어 제공: 사용자는 자신이 원하는 검색 키워드 기반 검색뿐만 아니라 검색엔진으로부터 그 키워드와 의미적으로 유사한 표준화된 검색 질의어를 추천받을 수 있다.

- 웹 문서의 문맥 정보 추출을 통한 의미적 키워드 인덱싱: 사용자가 입력한 검색 키워드와 의미적으로 가까운 웹 문서 검색이 가능하며, 그 키워드의 의미별로 검색 결과를 분류하여 제공할 수 있다.

- 문서의 내용기반 의미적 검색: 본 연구에서 추출하는 각 문서의 문맥 정보는 문서들 사이의 의미적 유사도를 측정할 수 있는 기반 데이터로 활용된다. 이는 사용자가 어떤 문서를 보유하고 있을 때 그 문서 자체를 질의어로 입력할 수 있으며, 그 문서와 의미적으로 유사한 순서대로 검색 결과를 제공할 수 있는 기반을 마련한다.

위와 같은 목적을 달성하기 위해 집단 지성(Collective Intelligence)의 집약체인 위키피디아 문서 집합을 이용한다. 먼저, 선행 연구를 통해 확장된 워드넷(Enriched WordNet)[16]을 지식베이스로 이용하고 확률적 가중치 및 의미적 가중치를 동시에 고려한 방법을 적용함으로써 위키피디아의 문서에서 핵심 문맥 정보를 추출한다. 그리고 핵심 문맥 정보를 바탕으로 위키피디아 내에 형성된 카테고리로 웹에 존재하는 문서들을 분류하며 위키피디아 문서 제목으로 태깅한다. 이러한 과정으로 진행되는 본 연구를 통해 기여할 수 있는 내용은 다음과 같다.

- 시맨틱 웹의 궁극적인 목적은 인간의 두뇌와 같은 지식을 보유하고 이를 의미적으로 처리하여 인간과 의미적 차이(Semantic Gap)없는 의사소통이다. 본 연구는 현 단계에서의 의사소통 즉, 검색에서 그러한 의미적 차이를 줄일 수 있는 방법을 제시하는 핵심연구이다.

- 시맨틱 웹 시대의 사용자들은 자신이 원하는 정보에 대해 더욱 정확하고 의미적으로 적합한 결과를 요구한다. 본 연구를 통해 사용자의 다양한 질의어 범위를 커버하고, 그에 부합하는 결과를 의미적으로 가까운 순서대로 제공할 수 있다.

- 인간의 지식을 기존에 형성된 지식베이스들로는 커버하기가 어렵다. 또한 지식은 시대의 변화, 시간의 흐름, 기술의 개발에 따라 다양해지기 때문에 그에 알맞은 지식베이스가 필요하다. 이에, 본 연구에서는 전 세계의 전문가들에 의해 꾸준히 지식을 누적하고 있는 위키피디아 문서 집합을 이용한다. 위키피디아는 기존의 지식 개념뿐만 아니라 현재와 미래에 중요할 수 있는 개념들을 정의함으로써 지속적으로 증가하고 있기 때문에 이를 새로운 형태의 지식베이스로 활용하기 위한 의미적 처리 방법을 제시한다.

- 웹 문서 검색은 단순히 사용자가 입력하는 질의어 기반의 검색뿐만 아니라, 사용자가 보유하고 있는 문서를 이용한 검색이 필요하다. 사용자의 질의어는 몇 개의 단어만으로 구성되어 있어 의미성을 판단하기 어렵고, 그 질의어에 의해 검색되는 문서 또한 의미적 분류 없이 방대하게 제공될 수 있다. 본 연구의 결과는 사용자가 보유하고 있는 문서와 웹에 있는 문서들 사이의 의미성을 고려함으로써 유사한 문서를 제공할 수 있으며, 사용자들의 전문 지식에 대한 욕구를 충족시켜줄 수 있다.

- 웹 문서들의 표준화된 분류 및 태깅 방법이 필요하다. 일반적으로 기존 연구들은 확률적 또는 의미적으로 핵심이 되는 키워드들을 추출하여 문서를 대표할 수 있는 어휘로 형성하고 있다. 그러나 이는 분류 및 태깅에 있어서 개념 활용에 대한 표준성이 부족하다. 이에, 본 연구에서는 위키피디아의 카테고리 구조와 문서 제목을 이용하여 분류 및 태깅하기 때문에 충분한 표준성을 제공할 수 있다. 이는 특히, 웹상에 존재하는 모든 문서들을 위키피디아에 연결할 수 있다는 강점을 가지고 있으며, 이러한 문서들의 의미적 네트워크는 궁극적으로 사용자에게 풍부한 정보 검색을 제공할 수 있다.

C. 연구 방법 및 내용

본 연구의 주요 내용은 웹에 존재하는 문서들을 표준화된 카테고리로 분류 및 태깅하는 것이다. 이를 위해, 개념 네트워크(Concept Network)를 확장한 워드넷을 이용하고, 확률적인 방법과 의미적인 방법을 함께 고려하여 문서의 문맥 정보를 추출한다. 또한 동일한 방법으로 웹에 존재하는 문서의 문맥정보를 추출한다. 문서 태깅의 정확도 증가 및 문서들의 의미적 네트워크 형성을 위해 표준화된 위키피디아 카테고리로 웹 문서를 분류하고, 각 문서의 내용을 위키피디아 문서 제목으로 태깅한다. 이러한 결과를 다양한 실험을 통해 본 연구의 유효성과 효율성을 제시한다.

본 논문의 구성은 다음과 같다.

본 장인 서론에 이어 2 장에서는 본 연구의 이론적 배경인 기존 연구들의 현재 수준과 개념들을 자세히 살펴본다. 그리고 연구 진행에 필요한 관련 연구들을 제시하여 3장부터 전개되는 연구 내용의 이해를 돕는다.

3장에서는 위키피디아의 각 문서에서 핵심이 되는 문맥 정보를 추출하는 방법에 대해 기술한다. 문맥 정보의 추출을 위한 세부 절차에 대해 살펴보고, 문서에서 명사 유형 추출을 위한 전처리 과정, 확률적인 가중치와 의미적 가중치를 함께 고려한 문맥 가중치 측정 방법에 대해 상세하게 다룬다.

4장에서는 3장에서 추출한 문맥 정보를 통해 웹 문서의 의미적 태깅 방법을 제시한다. 의미적 태깅에 필요한 전체 절차에 대해 살펴보고, 위키피디아 카테고리 구성 방법, 카테고리를 대표할 수 있는 문맥 정보 형성, 카테고리로의 웹 문서 분류 및 위키피디아 문서 제목을 이용한 태깅 방법을 기술한다.

4장의 결과물인 위키피디아 카테고리로의 문서 분류 및 위키피디아 문서 제목을 이용한 태깅 방법에 대해 5장에서는 다양한 실험을 시도한다. 추출된 각 문맥 정보에 대한 정확도 평가, 문서 분류의 정확도 평가, 문서 태깅에 대한 정확도 평가에 대해 기술한다. 이러한 각 평가 결과들을 기존의 타 연구결과들과 비교하여 본 연구의 유효성과 효율성을 제시한다.

마지막으로 6장은 결론을 맺고 향후 연구의 방향을 제시함으로써 본 연구의 필요성과 가능성을 도출한다.

II. 관련 연구

1990년대 말 정보화 시대를 거쳐 새로운 세기가 시작되면서 웹에는 무수히 많은 데이터들이 누적되고 있다. 또한 다양한 분야에 해당하는 지식들을 체계적으로 정리함으로써 웹의 지식화는 이미 완성단계에 있다. 현재는 이러한 웹 지식화 단계를 넘어 웹 자체에서 지능적인 서비스를 제공할 수 있는 시맨틱 웹으로의 도약이 진행되고 있다.

본 장에서는 지능적인 서비스를 제공하기 위해 필요한 요소들과 그것을 활용한 연구들을 살펴보고, 연구 진행에 필요한 저자의 선행 연구를 포함, 기존 연구들의 배경과 이론을 설명한다. 제시되는 내용들을 통해 3장부터 전개되는 연구의 동기를 이해하고 시맨틱 웹의 필요성과 중요성을 살펴본다.

A. 지식베이스 기반 연구

월드 와이드 웹(World Wide Web)의 창시자인 팀 버너스 리(Tim Berners-Lee)가 제안한 시맨틱 웹(Semantic Web)은 기계가 정보의 의미를 이해하여 인간과 의사소통을 원활하게 하는 것을 목표로 하고 있다[17]. 기계가 정보를 이해하기 위해서는 세상에 존재하는 개념들에 대한 지식베이스를 구축해야 한다. 이러한 지식베이스로 가장 일반적으로 활용되는 것이 워드넷(WordNet)이다. 본 단원에서는 워드넷의 특징에 대해 알아보고, 그를 활용한 다양한 연구들에 대해서 살펴본다.

1. 워드넷(WordNet)

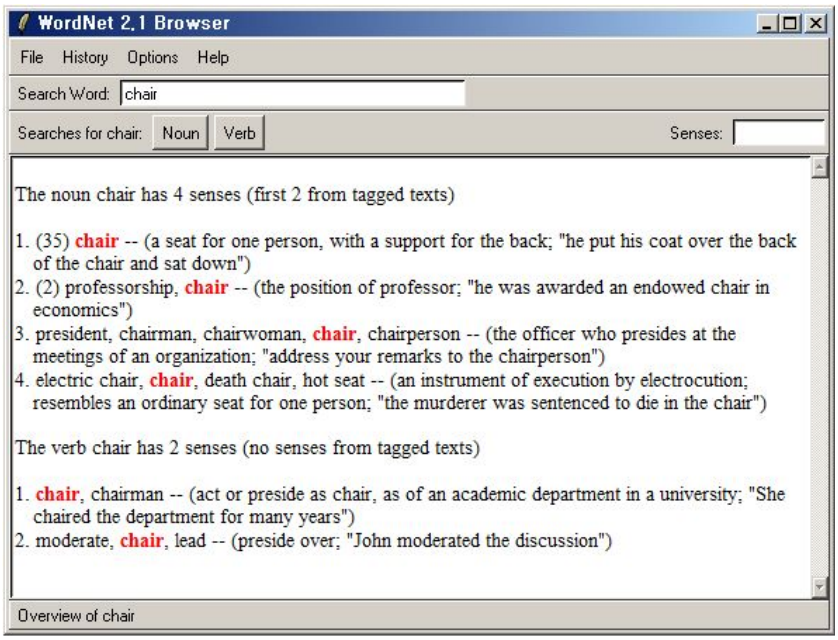
워드넷은 미국의 프린스턴 대학(Princeton University)에서 인간의 언어 심리학을 토대로 1985년부터 개발하기 시작한 영어 어휘 사전(Lexical Dictionary)이다. 워드넷은 영어 개념을 명사, 동사, 형용사, 부사로 분류하여 각 개념의 정의(Glossary)와 개념들 사이의 관계를

유형별로 정의하고 있으며 현재는 3.0 버전까지 개발되어 제공되고 있다. <표 1>은 연구에 가장 많이 사용되는 워드넷 2.1에서 정의하는 개념들의 통계를 보이고 있다.

<표 1> 워드넷이 정의하는 단어와 개념에 대한 통계

구 분	단어(개)	개념(개)
명사 (Nouns)	117,097	81,426
동사 (Verbs)	11,488	13,650
형용사 (Adjectives)	22,141	18,877
부사 (Adverbs)	4,601	3,644
합계	**155,327**	**117,597**

<표 1>과 같이 개념들을 정의하고 있는 워드넷은 각 개념의 고유한 의미를 번호로 분류하여 정의하고 있는데, 이를 신셋(Synset, Synonym Set)이라 부른다. 신셋이 동일한 것은 단어가 서로 다르더라도 같은 의미를 표현한다. 또한, 단어는 여러 가지 의미(Sense)를 표현할 수 있는데(다의어) 의미에 따라 각기 서로 다른 신셋을 갖는다. [그림 2]는 워드넷 브라우저(Browser)에서 단어 'chair'를 검색한 결과를 보이고 있다. [그림 2]에서 'chair'에 대한 명사 부분의 의미는 총 4개로 표현된다. 각 정의의 앞부분에 1,2,3,4와 같은 숫자는 센스 번호를 의미한다. 즉, '의자'에 관한 것은 'chair'의 의미 중에서 센스 1번이며 이를 'chair#1'로 표현한다. 이와 동일하게 '의장'에 대한 것은 'chair#3'으로 표현할 수 있다. 또한 'chair#3'은 'president#4', 'chairman#1', 'chairwoman#1', 그리고 'chairperson#1'과 동일한 의미를 갖는데, 이들은 서로 같은 신셋으로 표현된다. 즉, 이와 같이 워드넷은 기본적으로 특정 단어가 갖는 여러 가지 의미를 분류(다의어)하여 정의하고 있으며(의미가 결정된 것은 개념이라 불림), 같은 의미를 갖는 개념들(이음동의어)을 함께 보여주고 있다.



[그림 2] 워드넷 2.1 브라우저에서 'chair'를 검색한 결과

워드넷은 개념들의 정의뿐만 아니라 개념들 사이의 의미적 네트워크(Semantic Network)를 포함하고 있다. 이는 사람이 정보를 저장하는 형태와 동일한 구조를 따른다. '타이어'라는 단어(또는 그림)를 머릿속에 떠올렸다면 '자동차'가 그 다음으로 자연스럽게 연상이 될 것이다. 이는 인간이 어떤 정보를 처음 접할 때, 기존에 이미 알고 있는 다른 지식과 암묵적인 연결을 형성함으로써 기억하게 됨을 의미한다. 이러한 방식과 유사하게 워드넷은 개념들 사이에 존재하는 의미적 관계를 형성하고 있어 다양한 의미적 정보 처리에 활용되고 있다. <표 2>는 워드넷이 포함하는 명사 개념들 사이의 관계, 그를 표현하는 기호(Symbol), 각각의 정의 및 예제를 보이고 있다.

<표 2> 워드넷이 포함하는 명사 개념들 사이의 관계, 기호, 정의 및 예제

관계	기호 (S)	정의 $S(C_A, C_B)$	예제 기호(개념 관계쌍)
상/하위	~	C_A는 C_B를 하위 개념으로 가짐	~(country#1, South Korea#1)
	@	C_A는 C_B를 상위 개념으로 가짐	@(South Korea#1, country#1)
인스턴스	~i	C_A는 C_B를 인스턴스로 가짐	~i(lawyer#1, Abraham Lincoln#1)
	@i	C_A는 C_B의 인스턴스임	@i(Abraham Lincoln#1, laywer#1)
도메인	;c	C_A는 C_B를 도메인으로 가짐	;c(error#4, baseball#1)
	-c	C_A는 C_B의 도메인임	-c(baseball#1, error#4)
도메인과 지역	;r	C_A는 C_B를 도메인으로 가지며, C_B는 지역임	;r(multiple-voting#1, United Sates#1)
	-r	C_A는 지역이며, C_A는 C_B의 도메인임	-r(Janan#2, tee ceremony#1)
구성(실체)	%s	C_A는 C_B로 구성됨	%s(snowball#4, snow#2)
	#s	C_A는 C_B를 구성함	#s(snow#2, snowball#4)
사용형태	;u	C_A는 C_B의 형태로 사용됨	;u(sunglasses#1, plural form#1)
	-u	C_A는 C_B의 사용형태임	-u(plural form#1, sunglasses#1)
부분-전체	#p	C_A는 C_B의 일부임	#p(canyonside#1, canyon#1)
	%p	C_A는 C_B를 일부로 가짐	%p(canyon#1, canyonside#1)
반의	!	C_A와 C_B는 서로 반대 의미임	!(father#1, mother#1)

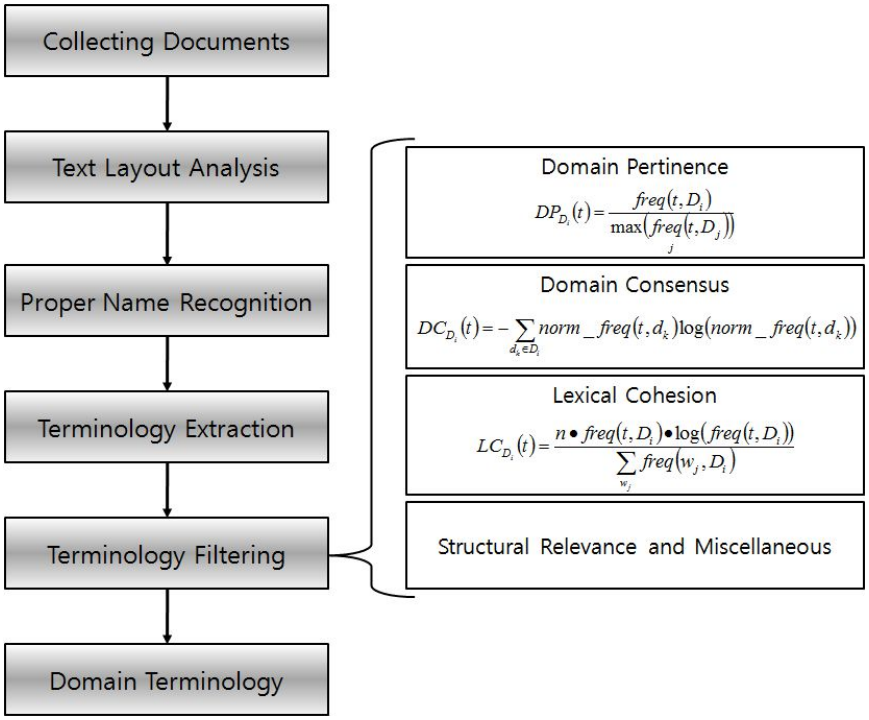
<표 2>와 같은 관계 구조를 이용하여 총 203,760개의 개념 관계쌍을 정의하고 있다. 이러한 워드넷의 개념 정의와 관계 형성은 다양한 자연어 처리 분야에서 의미성을 부여할 수 있는 기반 데이터로써 충분한 가치를 갖는다. 본 연구에서도 위키피디아 및 웹 문서에서 문맥 정보를 추출할 때 의미적 가중치를 측정하기 위해 워드넷을 활용한다.

2. 지식베이스 개념 확장

워드넷이 일상에 사용되는 단어의 의미들을 대부분 정의하고 있지만, 특정 도메인에 해당하는 전문 용어들은 포함하지 않는다. 또한, 용어들은 기술 개발, 새로운 트렌드(Trend)의 출현, 학문의 진화 등에 따라 새롭게 지속적으로 생성된다. 이러한 용어들을 사람의 수작업에 의해 새롭게 정의하는 것은 워드넷을 새롭게 제작하는 것 또는 그 이상의

시간, 비용, 노동 그리고 의견대립까지 요구할 수 있다. 이에, 워드넷에 정의되지 않은 용어들을 자동으로 판단하고 그 용어에 알맞은 상위 개념을 찾는 연구[12, 13, 22]가 있다.

특정 도메인에 속한 용어들을 추출하고, 그 용어와 워드넷 내의 개념 연결 및 새롭게 정의된 용어들 사이의 의미적 네트워크까지 파악하는 방법[12, 22]이 제안되었으며, 도메인 용어 추출 과정은 [그림 3]과 같다.



[그림 3] 도메인 용어 추출 과정

[그림 3]과 같이, 해당 도메인에 해당하는 문서들을 분류하여 수집하고 그 문서들에서 제목, 본문, 강조된 부분을 분석함으로써 고유명사와 명사구들을 파악한다. 파악된 명사구 부분에서 해당 도메인에 알맞은 것들만 추출하기 위한 필터링 과정을 거치는데, 이를 위해 도메인 적절성(Domain Pertinence, 추출된 명사구가 해당 도메인에 출현하는 가중치가 얼마나 되는가?), 도메인 일치성(Domain Consense, 추출된 명사구가 도메인 내의 문서들에서 얼마나 고루 출현하는가?), 어휘 밀집도(Lexical Cohesion, 명사구를 구성하는 단어들 사이의 밀집도가 얼마나 되는가?), 및 강조된 부분, 대/소제목 등에 포함된 것들의

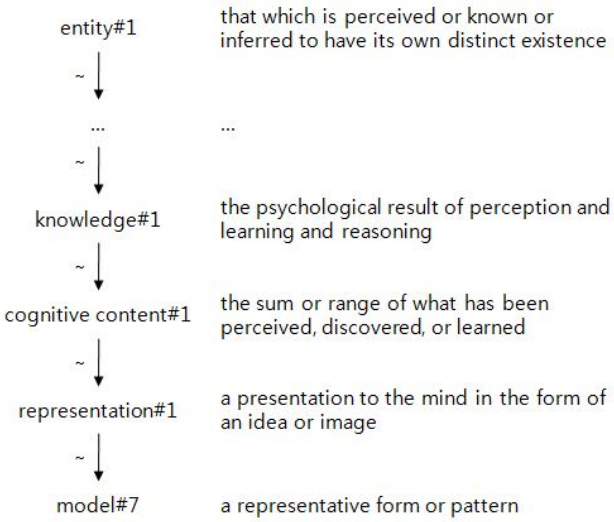
가중치를 측정한다. 최종적으로 측정된 각 가중치의 값을 더하여 해당 도메인을 파악하고, 그 도메인 내에 속한 워드넷 개념 하나를 상위어로 결정하는 부분이다.

위의 내용에 추가하여 추출된 명사구들 사이의 계층구조까지 파악하는 연구[12]가 수행되었다. 추출된 명사구는 일반적으로 수식어구와 주요부로 구성되는데, 주요부에 따라서 상위 개념의 결정이 도메인 용어 추출 연구[22]에서 성립되었다. 하지만, 이들 명사구들 사이에서 구체적인 계층구조를 형성하기에는 어려움이 존재하였으며 이를 극복하기 위해 워드넷을 기반으로 수식어구 사이의 관계를 파악함으로써 상/하위 관계를 판단하는 방법[12]이 연구되었다. <표 3>은 추출된 도메인 용어들을 이용하여 계층구조를 형성하는 과정을 보이고 있다.

<표 3> 도메인 용어 추출 및 계층구조 파악

추출된 도메인 용어	[22]에 의한 결과 (각 도메인 용어의 상위어만 파악)	[12]에 의한 결과 (각 도메인 용어들 사이의 계층구조까지 파악)
knowledge integration representation integration schema integration model integration information integration data integration program integration application integration service integration specification integration ontology integration	integration knowledge integration representation integration schema integration model integration information integration data integration program integration application integration service integration specification integration ontology integration	integration knowledge integration representation integration model integration schema integration ontology integration information integration data integration program integration application integration service integration specification integration ontology integration

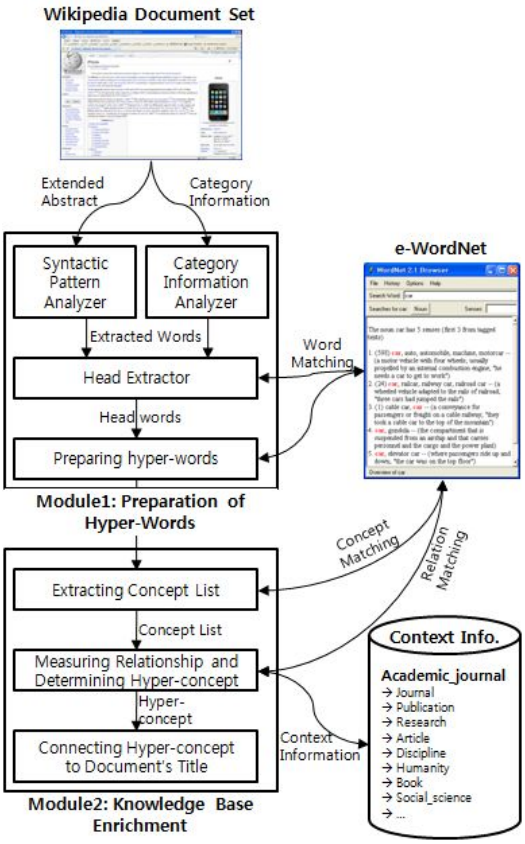
<표 3>에서 세 도메인 용어 'knowledge integration'과 'representation integration' 및 'model integration' 사이에 상/하위 관계가 형성되어 있음을 확인할 수 있다. 이들은 각 도메인 용어의 수식어 사이의 관계를 이용한 것인데, 워드넷을 통해 'model', 'representation', 'knowledge' 사이에 [그림 4]와 같은 관계를 파악함으로써 그러한 계층구조를 형성할 수 있게 된다.



[그림 4] 워드넷 내의 'model', 'representation', 'knowledge' 사이의 관계 구조

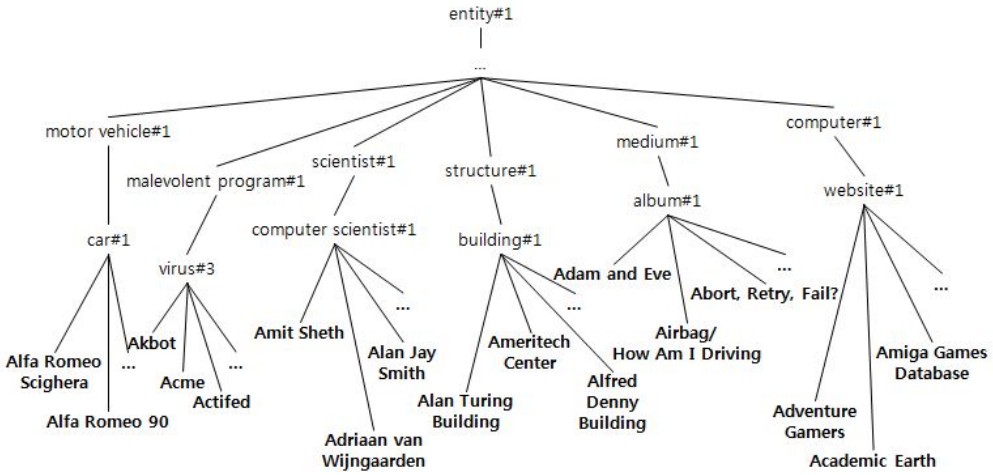
즉, 수식어구들 사이의 계층구조가 도메인 용어의 계층구조를 결정할 수 있다는 전제하에 지식베이스를 확장할 수 있었다.

또한 위키피디아가 포함하는 문서집합에서 각 문서의 제목을 개념화하여 워드넷에 추가하는 연구[13]가 있었다. 위키피디아는 역사, 사건, 철학, 문화, 예술, 과학, 시스템, 사람, 기술, 동물 등 한정되지 않은 주제들에 대한 상세한 설명을 포함하고 있으며, 각 분야의 전문가들에 의해 새로운 주제가 제기 또는 각 주제에 대한 내용이 지속적으로 채워지고 있다. 이에, 위키피디아는 시대의 흐름에 따른 새로운 개념들까지 포함한다는 측면에서 새로운 형태의 지식베이스로 각광받고 있다. 또한 각 문서의 제목은 알고리즘 제목(예. Euclidean algorithm, Monte Carlo method, 등), 사람 이름(예. Barack Obama, Amit Sheth, 등), 제품 이름(예. BMW Z4, Intel 80486DX2, 등), 노래 제목(예. I'm yours, Shape of My heart, 등), 그룹 이름(예. Backstreet Boys, Westlife, 등) 등의 유/무형의 객체 이름으로 구성되어 있다. 이러한 제목들은 하나의 명사구 또는 특정 개체의 전체 이름으로 구성되어 있어 워드넷 내의 특정 개념의 하위에 추가 될 수 있다. 이에 위키피디아의 문서 제목을 추출하여 워드넷의 특정 개념에 추가하는 연구[13]가 수행되었다. 이에 대한 전체 과정은 [그림 5]에서 보이고 있다.



[그림 5] 위키피디아 문서 제목의 개념화 과정

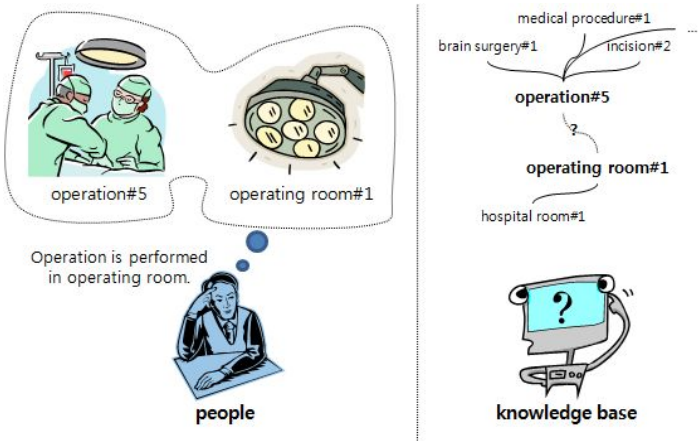
[그림 5]와 같이, 먼저 위키피디아의 문서 제목, 초록 및 카테고리 정보(category information)을 추출한다. 이 연구는 초록 문서에서 첫 번째 문장의 86%가 “위키피디아 제목 + be 동사 + 관사 + 정의 명사구”의 패턴을 따름을 파악하고, 이러한 형태에서 ‘정의 명사구’가 위키피디아 제목의 상위 용어 후보가 될 수 있다는 데에 근거하고 있다. 그 패턴을 따르지 않더라도 카테고리 정보에서 위키피디아 제목에 대한 상위 용어 후보를 추출하였다. 그리고 상위 용어 후보에서 정확한 워드넷 개념을 초록 문서에 포함된 어휘들 사이의 의미적 관계성을 측정하여 선정하였다. 이와 같은 방법으로 93.9(%)와 82.1(%)의 재현율과 정확도를 얻을 수 있었으며, [그림 6]은 그 결과에 의해 확장된 워드넷의 일부를 보이고 있다.



[그림 6] 위키피디아 문서 제목의 개념화에 의한 워드넷 확장 결과

3. 지식베이스 관계 확장

지식베이스로써 활용도가 높은 워드넷은 개념 확장에 대한 연구뿐만 아니라 관계를 확장하기 위한 연구도 다수 진행되었다. 워드넷은 다양한 개념들과 그들 사이의 관계를 정의하고 있지만, 사람들에 의해 실생활에 사용하는 개념 관계 쌍들을 충분히 커버하지 못하는 한계점이 지적되었다[21]. [그림 7]을 예로 보면, 수술(operation#5)은 수술실(operating room#1)에서 수행되기 때문에, 사람들은 암묵적으로 둘 사이의 관계를 생각해볼 수 있다. 하지만, 워드넷은 두 개념 사이의 관계를 멀게 정의하고 있기 때문에, 워드넷을 기반으로 한 의미적 정보 처리에 관한 연구들 또한 그들 사이의 관계를 파악하지 못하는 한계점을 이어받게 된다. 이러한 워드넷의 개념 관계쌍 부재의 한계를 해결하기 위해 개념 관계쌍을 확장하는 연구[16, 19]가 수행되었다. 두 연구는 워드넷의 개념 관계쌍을 다양한 근거에 의해 확장하였다는 점과, 그 결과를 WSD(Word Sense Disambiguation)에 적용하였다는 것에 공통점을 갖는다.



[그림 7] 사람과 지식베이스 사이의 의미적 차이(Semantic Gap)

먼저 다양한 지식을 통합하여 초대형 어휘 사전을 형성한 연구[19]에 대해 살펴보면, 워드넷 2.0, 도메인 레이블(Domain Labels)[23], 셈코(SemCor) 문서 집합[24], 엘디씨-디에스오(LDC-DSO) 집합[25], 워드넷 개념 정의(WordNet Glossaries와 WordNet Usage Examples), 및 사전(Oxford Collocations[26], Longman Language Activator[27], 그리고 Lexical FreeNet)을 이용하였으며, 통합된 어휘 사전을 기반한 SSI(Structural Semantic Interconnections) WSD 알고리즘을 제안하였다. SSI 알고리즘은 주어진 문맥(단어 집합)에서 관계성이 높은 것을 선정하는 것으로 방법은 <표 4>와 같다. 이러한 SSI 알고리즘은 지식베이스를 기반으로 한 연구 중에서 가장 높은 성능을 보이는 것으로 확인되었다.

<표 4> SSI에 의한 WSD 결정 과정
(실제 SSI가 사용한 지식베이스가 없어 확장된 워드넷을 이용한 결과임, 개념들 사이에 형성되는 관계들 중에서 일부만을 표기한 것임)

과정	예제	
주어진 단어 집합	house, apartment, room, wall, floor, window, guest	
단의어 파악	**apartment#1**	
결정된 개념과 의미가 결정되지 않은 단어들 사이의 관계성 파악	**apartment#1**->#p apartment_building->@ building->#p **room#1**	**room#1**
	apartment#1->#p apartment_building->@ building->#p **wall#1**	**wall#1**
	room#1->#p building#1->~ **house#1** **wall#1**->#p building#1->~ **house#1**	**house#1**
	room#1->#p **floor#1** **wall#1**->#p room#1->~ **floor#1**	**floor#1**
	room#1->#p building->#p **window#1** **wall#1**->#p building->#p **window#1**	**window#1**
	window#1->#p building->~ hotel->;gr **guest#3** **room#1**->-gr rooms->~ hotel->;gr **guest#3**	**guest#3**
결과	apartment#1, room#1, wall#1, house#1, floor#1, window#1, guest#3	

SSI 알고리즘이 여러 가지 사전을 통합한 초대형 지식베이스를 이용하고 있지만, 통합하는 과정에서 신뢰성이 결여되는 한계점이 지적되었다. 이에 확실한 근거를 바탕으로 하는 지식베이스 확장을 주장하며, 워드넷 자체의 특성을 이용한 개념 관계 쌍 확장에 대한 연구[16]가 시도되었다. 이 방법은 본 연구를 위한 선행연구로써 II-D 단원에서 상세하게 기술한다.

4. 개념들 사이의 의미적 유사도(Semantic Similarity) 측정 방법

개념들 사이에 존재하는 의미적 유사도를 측정하는 방법은 크게 에지 계산 방법(Edge Counting Methods), 정보량 방법(Information Content Methods), 특징 기반 방법(Feature based Methods)으로 나뉜다[14]. 에지 계산 방법은 두 개념 사이에 형성되는 관계 경로의

개수를 고려한 것으로 다양한 연구[20, 21, 28, 29, 30, 31, 32]가 이를 바탕으로 하고 있다. 정보량 측정 방법은 두 개념의 최소 공통 상위 개념을 찾아, 그 개념의 정보량을 유사도로 계산하는 방식으로 이 또한 많은 연구[33, 34, 35, 36, 37]에서 활용되고 있다. 특징 기반 방법은 두 개념이 갖고 있는 특징(예. 정의 구문 등)에서 공통점을 찾아 유사성을 측정하는 방식이다[38]. 이와 같이 정보의 의미를 파악하기 위해 개념들 사이의 유사도 측정에 관한 연구가 다양하게 진행되었으며, 본 논문에서는 사전(Dictionary)에 정의되지 않은 용어(고유명사)들의 개념화를 위한 연구[20, 21]에 대한 방법을 조금 더 깊게 다룬다.

기존의 의미적 유사도 측정에 관한 연구가 이미 어휘 사전(지식베이스)에 정의된 개념들 사이의 관계성을 측정하는 방식이라면, 사전에 정의되지 않은 어휘(고유명사)와 이미 사전에 정의된 개념 사이의 관계성을 측정하는 방식이 [20]과 [21]의 연구이다. 즉, 웹에 게시된 문서에 존재하는 고유명사 또는 신조어(Unknown Word)들을 자동으로 지식베이스에 확장하기 위해 관계성이 깊은 개념들을 파악하기 위해 제안되었다. 이 방법들은 한 문장에 함께 출현(Co-occurrence)하는 단어들은 서로 관계성이 높다는 근거를 바탕으로 진행되었는데, 고유명사를 그 중심에 두고 있다. 특정 고유명사를 포함하는 문장들을 수집하고, 문장들에 포함된 단어들이 출현빈도가 높고 그들 사이에 관계성이 높은 단어는 목적 고유명사와 관계성이 특히 높을 수 있다는 사실을 바탕으로 확률적 가중치와 의미적 가중치를 동시에 고려하고 있다. 확률적 가중치는 베이지안 확률인 (식 1)을 이용하여 측정한다.

$$pw(uw_i, rt_j) = \frac{P(oc(uw_i)|oc(rt_j)) \times P(oc(rt_j))}{P(oc(uw_i))} \quad \text{(식 1)}$$

(식 1)에서 oc는 출현한 횟수, uw는 고유명사 또는 신조어, rt는 함께 출현한 어휘(관련어휘, Related Term)을 의미한다. 이 수식을 통해 고유명사와 관련 어휘 사이의 확률적 응집력(Probabilistic Cohesion)을 측정할 수 있다. 여기에 의미성을 반영하기 위해 의미적 가중치를 측정하는데 이를 위해 관련 어휘들의 개념 리스트(CL, Concept List)를 워드넷을 통해 얻는다.

$CL(rt_j)=\{c_{jk},\ 1 \leqq k \leqq n\}$

n: 관련 어휘 j가 갖는 개념의 수,
k: 관련어휘 j가 갖는 개념

관련 어휘가 갖는 개념들이 갖는 값들 중에서 최대값을 그 어휘의 의미적 가중치로 결정한다. 이는 지식베이스인 워드넷을 통해 형성될 수 있는 두 개념 사이의 관계 그래프에서 에지(edge)의 수에 반비례 관계가 있음을 이용하며 (식 2)는 이를 표현하고 있다.

$$rd(c_{jk}, c_{lm}) = \frac{1}{\min(d(c_{jk}, c_{lm}))}, j \neq l \quad \text{(식 2)}$$

rd(related degree): 두 개념 사이의 관련성

(식 2)에 각 어휘가 갖는 여러 개념들의 관련성이 측정되는데, 그 값 중에서 최대값을 그 어휘의 의미적 가중치로 결정하며, 이를 위해 (식 3)과 (식 4)를 이용한다.

$$cw(c_{jk}) = \sum_{l=1}^{n} \max(rd(c_{jk}, c_{lm})), j \neq l \quad \text{(식 3)}$$

cw(concept weight): 특정 개념의 의미적 가중치,
n: 특정 고유명사와 함께 출현한 관련 어휘들의 개수

$$sw(rt_j) = \arg_{c_{jk} \in CL(rt_j)} \max(cw(c_{jk})) \quad \text{(식 4)}$$

sw(semantic weight): 관련 어휘의 의미적 가중치

이러한 방법으로 고유명사와 함께 출현한 어휘들의 확률적 가중치와 의미적 가중치를 측정할 수 있다. 최종적으로 이를 함께 반영하기 위해 다음과 같이 두 값의 곱 연산을 통해 최종 관련성을 측정한다. (식 5)는 확률적 가중치를 기본 값으로 두고, 의미적 가중치를 변수로 반영하기 위한 것이다. 즉, 확률적 가중치와 의미적 가중치가 함께 높을 때 관련성은

최대가 된다.

$$rd(uw_i, rt_j) = pw(rt_j) \times (sw(rt_j) + 1) \quad \text{(식 5)}$$

<표 5>는 이 방법에 활용된 문서(위키피디아에 정의된 'McDonald' 문서 초록을 이용)를 보이고 있으며, <표 6>은 의미적 관련성에 의해 'McDonald'와 관련된 상위 5개의 어휘를 보이고 있다.

<표 5> 위키피디아에 정의된 'McDonald' 문서

문서 내용	McDonald's (NYSE: MCD) is the world's largest chain of fast food restaurants, serving nearly 54 million customers daily. McDonald's primarily sells hamburgers, cheeseburgers, chicken products, French fries, breakfast items, soft drinks, milkshakes and desserts. More recently, it also offers salads, wraps and fruit. Many McDonald's restaurants have included a playground for children and advertising geared toward children, and some have been redesigned in a more 'natural' style, with a particular emphasis on comfort and the absence of hard plastic chairs and tables.
추출된 명사	NYSE, world, chain, food, restaurant, customer, hamburger, cheeseburger, chicken, product, fry, breakfast, item, drink, milkshake, dessert, salad, fruit, restaurant, playground, advertising, style, emphasis, plastic, chair, table

<표 6> 의미적 관련성 측정에 의한 'McDonald'와 관련된 상위 5개 어휘

고유명사	관련 어휘	확률적 가중치	의미적 가중치	의미적 관련성*
McDonald	restaurant	0.667	1.000	1.0000
	drink	0.334	2.200	0.8000
	breakfast	0.334	1.600	0.6500
	hamburger	0.334	1.350	0.5875
	food	0.334	1.350	0.5875

*최대 의미적 관련성 값으로 정규화한 결과임.

이 방법은 고유명사가 10회 이상 출현하고 의미적 관련성에 의해 상위 30%에 속한 관련 어휘들에 대한 평가에서 약 84.5%의 추출 적절성(relevance)를 얻었다. 이 방법은 본

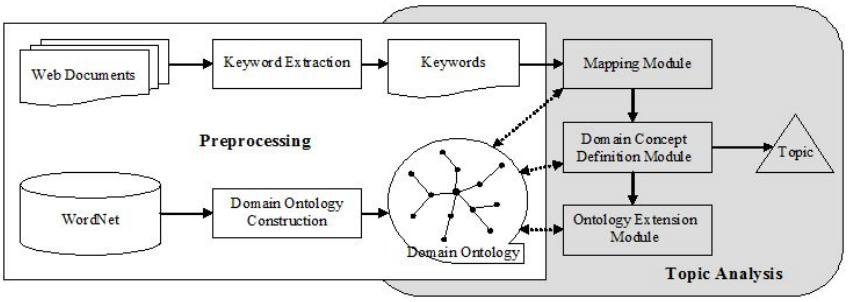
논문에서 위키피디아 문서의 의미적 문맥 정보 추출, 웹 문서의 핵심 키워드 추출 등에 응용되어 활용된다.

B. 문서 검색에 관한 연구

웹의 정보는 2000년부터 최소 약 2백만 TB(테라바이트)에서 매년 74.5%씩 증가하고 있다. 이에 웹에 저장된 정보의 검색은 매우 중요한 연구로 간주되고 있으며 시맨틱 웹의 발전과 함께 의미적 검색의 중요성이 증가하고 있다. 웹 문서 검색의 가장 대표적인 방식은 단어의 출현 빈도를 계산하는 TF-iDF(Term Frequency-inverse Document Frequency) 방식이 대표적이다. 하지만 본 단원에서는 검색하는 방법보다는 의미적 문서의 주제 선정, 유사도 측정 및 태깅 방식에 집중하여 다양한 기법들을 살펴본다.

1. 의미적 주제 선정 방법

문서의 주제를 선정하여 분류 및 검색에 적용하는 방법은 오랜 역사를 갖고 진행되었다. 이 연구들은 문서에 포함된 명사 어휘들을 바탕으로 그들을 모두 포함할 수 있는 대표어휘를 주제로 선정하는 방식이 일반적이다. 이 방법은 여러 문서의 학습을 통한 방식[44, 45]과 워드넷과 같은 지식베이스를 이용한 방식[43, 46]으로 나뉜다. 학습에 기반한 방식은 특정 주제와 관련된 문서들을 대량 수집하여 분석하고, 그 주제에 핵심이 되는 어휘들을 먼저 추출하여 웹 문서의 주제 선정 및 분류에 적용한다. 지식베이스를 이용한 방식은 문서에 출현하는 단어들의 의미성을 고려하여 그 문서의 내용을 대표할만한 개념을 지식베이스에서 찾는다. 일반적으로 문서에 포함된 내용은 하나 이상의 주제로 구성(문서를 대표할 수 있는 주제와 대표 주제의 특성을 설명하는 보조 주제)되는데 출현빈도만을 고려하는 것은 대표 주제를 선정하기 어렵기 때문에 지식베이스를 이용한다. 이에 지식베이스와의 매핑을 통해 핵심이 되는 어휘 그룹을 선정하고 그 그룹을 모두 포함할 수 있는 최소 상위 개념으로 주제를 선정한다는 점에서 공통점을 갖는다.



[그림 8] 문서의 의미적 주제 선정 방법[43, 46]

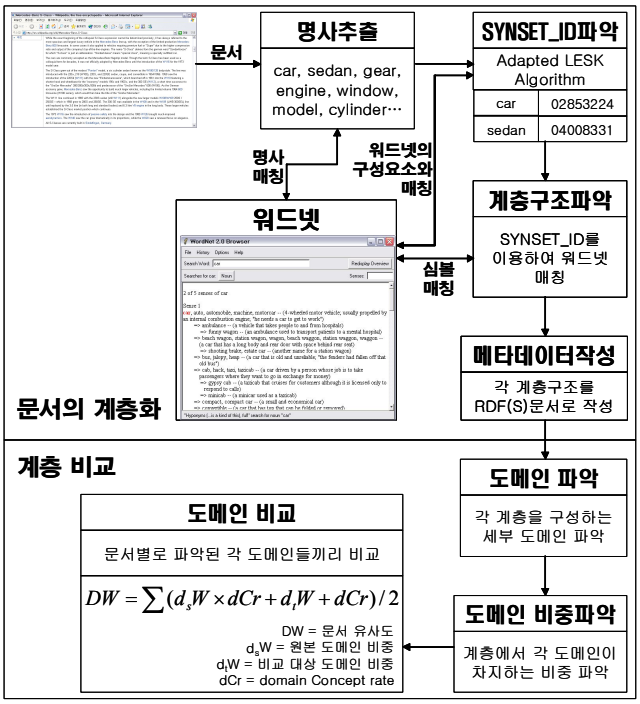
지식베이스를 이용한 방식[43, 46]에서 제안하는 방식을 살펴보면 공통적으로 [그림 8]과 같은 과정을 따른다. 웹 문서에서 명사어휘를 추출하고 출현 빈도(Term Frequency)를 이용하여 주요 어휘들을 먼저 선정한다. 주요 어휘들은 여러 도메인에 분산되어 있을 수 있다. 예를 들어, 정보기술을 적용한 의학에 관한 문서라 한다면 주요 어휘들은 '정보기술(IT)'과 '의학(BT)'로 나뉘어 추출될 수 있다. 이렇게 여러 도메인에 있는 주요 어휘들 중에서 핵심이 되는 도메인을 워드넷과의 매핑을 통해 파악한다는 점에서 두 연구는 유사하다. 하지만 문서의 주제를 선정하는 방식에서 [43]은 핵심 도메인에 속한 주요 어휘들을 모두 포함할 수 있는 워드넷 개념을 정보량(Information Content)에 기반한 유사도 측정(Similarity Measure)을 통해 선정하며, [46]은 주요 도메인 내에서 의미성이 가장 강한 하나의 워드넷 개념을 선정한다. 즉, [43]은 주요 도메인 내에 속한 어휘들을 모두 포함할 수 있는 상위 개념을 선정하는 반면, [46]은 하위 개념이라 할지라도 의미성이 강하다면 주제 개념으로 선정하여 문서 주제의 구체화를 도모할 수 있다. 이 방법들은 출현 빈도만을 이용한 방식과 주제 태그를 이용한 방식보다 높은 정확도를 보였지만, 워드넷에 포함하지 않은 어휘를 처리하지 못하는 한계점이 존재한다.

지식베이스를 기반으로 문서의 주제를 선정하는 것은 주제의 다양성 부재를 초래할 수 있다. 또한 새롭게 생성된 이슈(Issue)에 대해 다루지 못하는 한계점이 존재한다. 이러한 점을 극복하기 위해 본 연구에서는 위키피디아를 이용한다.

2. 문서 유사도 측정 방법

키워드만을 이용한 문서 검색은 사용자의 지식에 영향을 많이 받을 수 있다. 문서에 포함된 어휘들은 동일한 의미를 표현하지만 다른 모양(단어)로 구성될 수 있기 때문이다. 이에 문서들 사이의 의미적 유사성을 측정하는 연구[2, 47]가 제안되었다. 이들은 사용자가 관심있는 정보들의 핵심 키워드를 미리 파악할 필요 없이, 사용자가 보유하거나 읽고 있는 문서를 질의로 처리하여 유사한 문서들을 찾아 줄 수 있는 강점을 가진다.

문서 내용 계층화 방법[47]은 문서들 사이의 유사성을 측정하기 위해 명사 개념들을 이용하여 도메인을 파악하고, 각 도메인에 속한 개념들을 워드넷 계층구조로 형성한다. [그림 9]는 전체 과정을 보이고 있다.



[그림 9] 문서 계층화를 통한 유사도 측정 방법[47]

이 연구의 핵심은 문서가 표현하는 명사 개념들의 계층구조를 형성하는 것이다. 그리고 형성된 계층구조는 다른 문서의 계층구조와 비교되며, 그 개념들이 일치하는 것만 유사도에

반영된다. 이는 문서의 내용을 계층화하는데 있어서 의미성이 반영되지만 문서들의 유사도를 측정하는데 있어서 그러지 못하다는 한계점이 존재한다. 'computer'와 'notebook'을 각각 포함하는 문서들을 예로 들면, 두 단어의 모양은 다르지만 서로간의 관계성은 상당히 높을 수 있다. 하지만 이 방법으로는 이러한 관계성을 측정하지 못하기 때문에 둘 사이의 유사도를 0으로 판단해버린다. 이보다 발전된 의미적 문서 연결(SDI, Semantic Document Interconnections) 방법[2]이 수행되었다. SDI[2]에서는 문서에 출현하는 단어의 빈도와 단어들 사이에 존재하는 의미적 가중치를 반영하여 문서를 대표할 수 있는 핵심 키워드 개념 10개를 선정한다. 선정된 키워드들은 각각의 인덱싱 값(Indexing Value)을 갖는데, 문서들 사이의 유사성을 측정할 때 각 문서의 키워드들 사이에 존재하는 관계성을 이용한다. 문서들 사이의 유사성을 측정하는 수식은 (식 6)과 같다. 이 수식은 본 논문의 Ⅵ-C 단원에서 추출된 문맥 정보들 사이의 유사성을 측정할 때 한번 더 자세히 언급된다.

$$sim(d_k, d_l) = \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} \frac{I-value(s_{(k,i)}) + I-value(s_{(l,j)})}{2 \times \min dist(s_{(k,i)}, s_{(l,j)})}, k \neq l \qquad \text{(식 6)}$$

dist: 워드넷 내에서 두 개념 사이의 거리,
s(k,i): 문서 k에 속한 i번째 개념,
I-value: 인덱싱 값

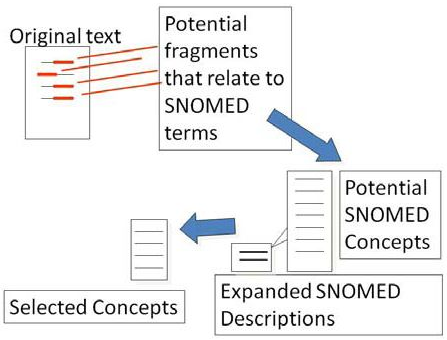
SDI[2]는 문서에서 핵심 키워드를 선정하거나 문서들 사이의 유사도를 측정할 때 모두 워드넷에 정의된 관계 구조를 기반으로 하기 때문에, 각 문서의 개념들이 동일하지 않더라도 의미성을 반영할 수 있는 강점을 갖는다.

3. 문서 태깅 방법

웹 2.0의 트렌드와 더불어 일반 웹 사용자가 블로그(Blog)나 개인 홈페이지를 통해 정보를 제공하는 정보 공급자로써의 역할을 활발히 수행하고 있다. 관심분야가 동일한 사람들의 커뮤니티(Community)가 형성되고 정보의 공유 및 재가공 또한 활발히 진행되고 있다. 기하급수적으로 늘어나는 문서들을 관리하고 검색의 용이성을 제공하기 위해 문서를 특정 키워드들로 표현하는 태깅에 대한 연구[48, 49, 50, 51, 52, 53, 54, 55, 56]가 활발히

진행되고 있다. 기존의 방식[50, 51, 52, 55]들은 퍼지 논리(Fuzzy Logic)[57]와 은닉 마르코프 모델(HMM, Hidden Markov Model)[58] 등을 활용한 기계 학습(Machine Learning)에 주로 의존하였다. 하지만, 기계가 정보를 이해하고 의미적으로 처리함으로써 사람과 의사소통을 원활하게 하는 것을 지향하는 시맨틱 웹(Semantic Web)의 발전과 더불어 워드넷, 위키피디아 및 온톨로지(Ontology) 등을 함께 사용하는 의미적 태깅에 대한 연구[48, 49, 53, 54, 56]가 다수 제안되었다.

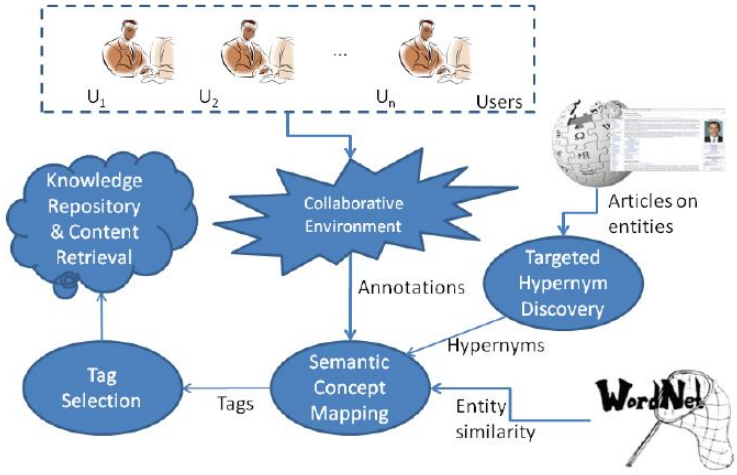
워드넷 또는 온톨로지를 이용하는 연구들은 태깅에 사용되는 어휘들을 수집하는 과정을 우선적으로 포함한다. 의학 문서를 태깅하기 위해 제안된 연구[49]는 의학 용어 사전인 SNOMED CT(Systematized Nomenclature of Medicin-Clinical Terms)를 확장하는 방법을 포함하고 있다. 또한, 개인 사용자들에 의해 생성된 블로그 문서들을 위한 의미적 태깅 방법[54]은 태깅 어휘들로 형성된 온톨로지를 구축하고 그것을 기반으로 하는 태깅 시스템(STSS, Semantic Tagging and Searching System)을 제작하였고, 이 시스템을 확장하여 응용한 연구[53]가 수행되었다. [그림 10]은 의학 문서 태깅 연구[49]에서 발췌한 SNOMED CT 사전을 확장하는 과정을 보이고 있으며, 일반적으로 태깅에 필요한 어휘 또는 개념 수집은 이와 유사하게 진행된다.



[그림 10] 대용량 문서로부터 온톨로지를 확장하는 일반적인 과정 [49]

또한 워드넷이 정의하는 개념들의 다양성 부재를 지적하고, 위키피디아를 이용하여 목적 어휘의 상위 개념을 파악하는 THD(Targeted Hypernym Discovery)[59]를 워드넷과 함께 이용하는 연구[56]가 수행되었다. 이는 워드넷을 통한 개념들 사이의 관계를 파악함으로써

태깅하는 과정에서 의미성을 부여하고, 태깅 어휘의 다양성을 위해 위키피디아로부터 개념들을 확장한다는 것에 본 연구와 유사하다 할 수 있다. [그림 11]은 이 연구에 대한 전반적인 과정을 보이고 있다.



[그림 11] 위키피디아와 워드넷을 함께 이용한 의미적 문서 태깅 방법 [56]

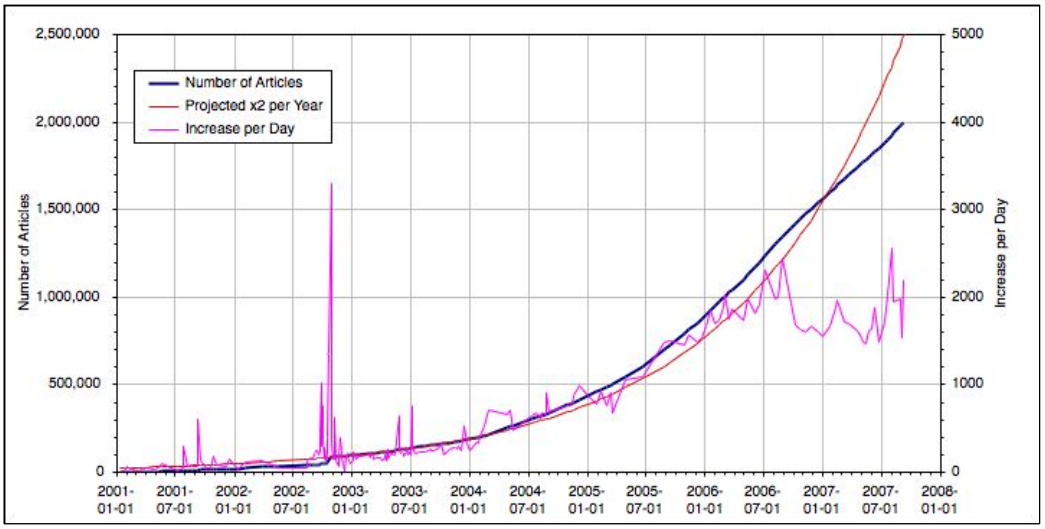
이러한 방법들은 60~70% 정도의 정확도를 달성하고 있지만, 지속적으로 생성되는 어휘, 주제 등에 대해서 태깅하지 못하는 한계점이 존재한다. 또한 이러한 어휘들을 처리하기 위해 온톨로지 재 확장 및 대용량 문서집합을 이용한 학습과정의 재 수행이 필요하며, 재학습을 통해 어휘들을 준비하더라도 태깅 어휘의 표준성이 결여되는 한계점이 존재한다. 이에, 본 논문에서는 태깅 어휘의 다양성과 표준성을 동시에 겸비할 수 있는 방법을 고안하며 웹 문서들을 의미적으로 태깅할 수 있는 방법을 제안한다.

C. 위키피디아 및 응용 연구

웹의 지식화를 위해 많은 정보들이 체계적으로 웹에 정리되고 있다. 그중 가장 대표적인 지식 창고는 위키피디아(Wikipedia)라 할 수 있다. 위키피디아는 특정 주제에 한정되지 않은 다양성, 지속적으로 증가하는 확장성, 그리고 집단 지성(Collaborative Intelligence)에 의한 정확성 등을 내포하고 있다. 이에 위키피디아에 정의된 문서들은 많은 연구들에 활용되고 있다. 본 단원에서는 이러한 위키피디아의 특징에 대해 기술하고 그것을 활용한 연구들에 대해 살펴본다.

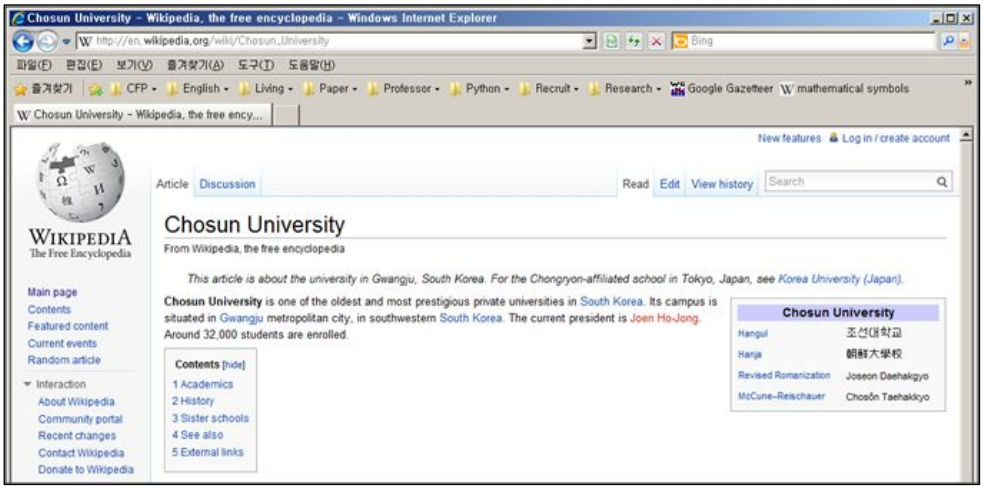
1. 위키피디아(Wikipedia)

위키피디아는 비영리 단체인 위키미디어 재단(Wikimedia Foundation)의 지원으로 사람들의 협업을 통해 제작되고 있는 웹 기반의 무료 전문 사전이다[39]. 이는 2001년부터 제작하기 시작하여 다양한 언어를 지원하며 현재(2010년 8월 19일 기준) 약 340만개의 영어 문서를 포함하고 있다. [그림 12]는 위키피디아 문서 규모의 성장 그래프를 보이고 있다.



[그림 12] 위키피디아의 규모 성장 그래프

또한 기존에 작성된 문서라 할지라도 해당 분야의 전문가에 의해 내용 수정 및 추가가 가능하며, 위키피디아에 정의되지 않은 어떤 주제를 등록할 수 있다. 이와 같이 전세계의 모든 사람에 의한 지식을 통합한 집단 지성의 집약체라 할 수 있다. 위키피디아는 하나의 주제에 대해 초록, 본문, 인포박스(Information box), 이미지(image), 참고문헌, 카테고리 정보(category information) 등으로 나누어 상세하게 기술하고 있다. [그림 13]은 위키피디아에서 'Chosun University'를 검색한 결과를 보이고 있다.



[그림 13] 위키피디아에서 'Chosun University' 검색 화면

이러한 위키피디아는 역사(history), 사건(events), 철학(philosophy), 문화(culture), 예술(arts), 과학(sciences), 시스템(systems), 인물(people), 기술(technology), 동물(animals), 식물(plants) 등 한정되지 않은 다양한 정보에 대해 기술하고 있어 기존의 검색 서비스의 통합검색과는 차별화된 전문 지식검색을 제공하고 있다. 특히, 위키피디아 문서에서 초록 부분은 그 주제의 핵심만을 다룬다는 특징이 있다. 이에, 본 연구에서는 위키피디아 문서 집합에서 초록 부분을 이용하여 주제(제목)에 따른 문맥 정보를 추출하는데 사용한다.

위키피디아의 부분별 정보는 DBPedia에서 제공하고 있는데, DBPedia는 위키피디아에서 추출된 정보를 구조적인 형태로 관리한다[40]. 특히, 위키피디아 문서를 정보박스, 제목, 초록, 이미지, 카테고리, 외부연결(External Links), 사람 데이터 등으로 구분하여 제공하며, 위키피디아의 내용이 역동적으로 증가, 수정 되는 것을 반영하기 위해 주기적으로 업데이트

및 버전별로 구분하여 제공한다. 특히 위키피디아 자체에 포함된 문서들의 연결을 온톨로지화 하여 제공 및 YAGO[42], Cyc[42] 등의 대표적인 온톨로지와도 연결을 제공하고 있어 위키피디아의 활용도를 극대화하는데 기여하고 있다.

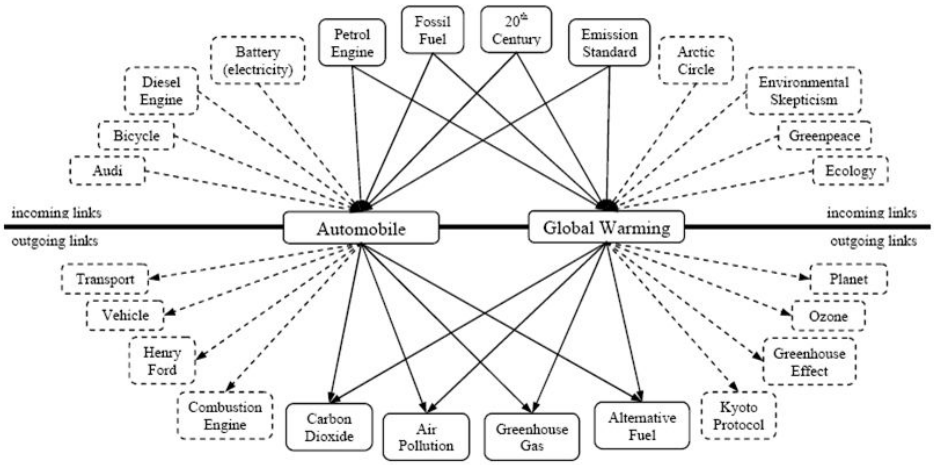
이러한 위키피디아는 다양성, 확장성, 정확성의 특징에 의해 웹에 있는 무수한 문서들을 관리하고 검색하는 기준이 될 수 있다. 위키피디아의 각 문서는 하나의 주제만을 다루고 있고 관련된 상위 주제로 카테고리화 되어 있다. 이에, 필요한 정보들을 주제별로 추출하여 웹 문서의 검색에 유용하게 활용된다.

2. 위키피디아 기반 연구

최근 위키피디아가 지식베이스로써 역할이 커지면서 이를 활용한 연구들이 급격히 늘고 있다. 위키피다아를 활용한 연구들의 주제로는 의미적 관계성(Semantic Relatedness)을 측정하는 방법[63, 66, 68, 71], 문서 주제 선정 방법[61], 문서 클러스터링 방법[62, 67, 70], 개체 명 분류(Named Entity Disambiguation) 방법[73], 문서 유사도 측정 방법[60, 78], 워드넷 및 온톨로지로 위키피디아에서 추출한 개념을 연결하는 방법[65, 72, 75], 문서 주석(Annotation) 처리 방법[64, 74], 문서 분류(Categorization) 방법[76, 79] 등이 대표적이며, 위키피디아 데이터 자체를 다루는 위키피디아 문서들의 의미성 풍부화 방법[77]과 위키피디아 내에서 잘못된 링크를 발견하는 방법[69] 등이 있다. 이들 중에서 본 연구와 깊이 관련된 의미적 관계성 측정, 문서 주석 그리고 문서 분류 방법에 대해 기술한다.

기존 연구에서 활발이 이용되는 워드넷 또는 로젯(Roget)과 같은 온톨로지 기반 방식이 포함하는 용어가 한정적이라는 한계점과 문서 집합인 코퍼스(Corpus)를 이용한 LSA(Latent Semantic Analysis) 방식이 비구조적이며 정확도가 낮다는 이유로 위키피디아를 활용한 의미적 개념 관계성 측정에 대한 연구가 수행되었다. 위키피디아 문서의 약 94%는 91,502개의 카테고리(2006년 1월 기준)에 포함되어 있으며[66], 카테고리들 사이에는 명명되지 않은 의미적 연결이 형성되어 있는데, 이러한 카테고리를 개념으로 간주하고, 카테고리의 관계를 기반으로 개념들 사이의 관계성을 파악하는 연구[66, 71]가 있었다. 하지만 이러한 위키피디아의 카테고리 구조 또한 관계성이 부족하고 개체명과 같은 구체적인

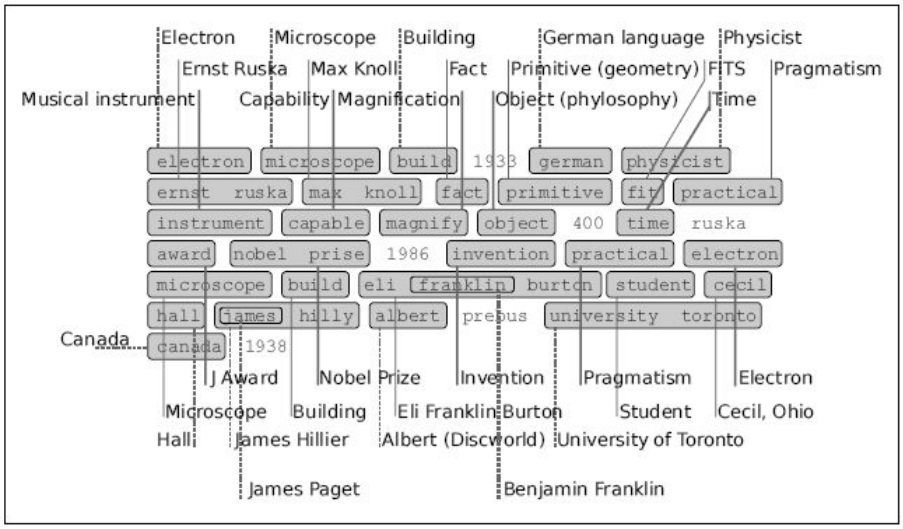
이름을 판단하지 못하는 한계점을 지적받았다. 이에 위키피디아 문서 제목을 개념으로 간주하여 그 한계점을 극복하려는 연구[68]가 있었다. 이는 위키피디아의 각 문서에 포함된 어휘들을 TF-iDF 방식으로 분석하여 개념(문서 제목)의 벡터 값으로 이용하고 있다. 그리고 어떠한 어휘들이 주어질 때 위키피디아에서 그 어휘들과 일치하는 개념들을 찾고 각 개념이 갖는 벡터를 비교함으로써 유사성을 측정한다. 이는 기존의 지식베이스 방식이나 카테고리 방식과 비교했을 때 정확도를 개선하였다. 이 정확도를 더욱 높이기 위해 위키피디아 문서들 사이에 존재하는 링크(Link)를 이용하는 연구[63] 또한 제안되었다. 이러한 연구들은 위키피디아 내에 존재하는 의미적 링크를 최대한 활용하여 개념들 사이의 유사도 및 문서들 사이의 유사도까지 활용되고 있다. [그림 14]는 위키피디아의 의미적 링크를 활용한 연구[63]에서 수행된 결과를 이용하여 'Automobile'과 'Global Warming' 사이의 관계성을 보이고 있다.



[그림 14] 위키피디아 문서들의 링크 기반 'Automobile'과 'Global Warming' 사이의 의미적 관계성[63]

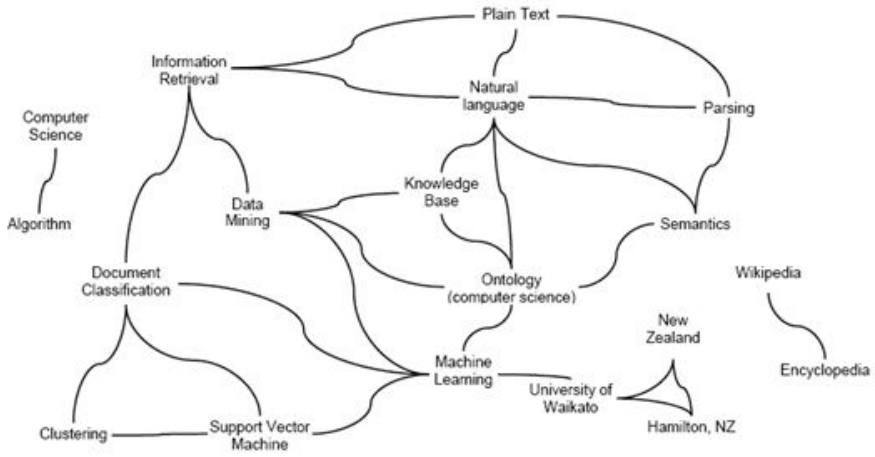
웹 문서에 포함된 여러 어휘들에 의미성을 부여하기 위해 위키피디아의 문서 제목(개념)으로 주석을 달기 위한 연구[64, 74]들이 수행되었다. 이들은 주석을 처리할 문서 내용에서 위키피디아의 제목과 일치하는 구문(타겟 구문)들을 파악하여 그 구문에 대해 위키피디아의 제목으로 주석을 처리함으로써 의미성을 부여한다. 이를 위해 문서의 여러 타겟 구문들로 확보된 위키피디아 문서들 사이의 연결 구조(Link Structure)를 파악하여 가장 유사성이 높은 개념들을 주석으로 선택한다. [그림 15]는 이 연구에 의해 생성된

문서의 주석을 도식화하고 있다.



[그림 15] 웹 문서의 내용을 위키피디아 제목(개념)으로 주석 처리[64]
(위키피디아 문서 중 'electron microscope'를 이용)

또한 이렇게 처리된 주석(개념)들의 링크 구조를 기반으로 문서의 내용을 의미적 네트워크로 형성하여 표현함으로써 문서의 주제를 부각할 수 있었다. [그림 16]은 주석 처리를 통해 형성된 의미적 네트워크를 보이고 있다.



[그림 16] 문서 내용에 포함된 구문들의 의미적 네트워크 형성[74]
([74]에 기술된 내용에서 추출된 네트워크)

기존의 문서 분류 방식들이 갖는 한계점 중 가장 큰것은 제한된 학습군이라 할 수 있다. 또한 학습군에 포함된 문서들이 그 주제와 관련성이 적거나 여러 노이즈 단어들을 포함할 수 있는 문제점들이 지적되었다. 이러한 문서 분류의 한계점을 극복하기 위해 위키피디아의 주제 다양성과 각 주제의 내용 일관성이라는 특징을 이용하는 연구[76, 79]가 있었다. 이들은 위키피디아를 기반으로 개념들 또는 문서들 사이의 유사도를 측정하는 방식과 비슷하게 문서를 분류함에 있어서도 위키피디아 문서 제목 또는 카테고리를 분류 카테고리로 간주하고 각 문서(또는 카테고리)에 포함된 어휘들을 분류 근거 데이터로 활용하고 있다. 이 방법들은 각 카테고리의 특징들을 추출하기 위해 TF-iDF, LSA(Latent Semantic Analysis), KPCA(Kernel Principal Component Analysis), 및 KCCA(Kernel Canonical Correlation Analysis) 방법으로 학습하였다. 위키피디아의 카테고리 또는 문서를 활용하고 각 문서에 포함된 어휘들을 활용하는 점에서 본 연구와 유사하지만 문서에 포함된 어휘들을 이용하는 방법에서는 확연한 차이가 있다.

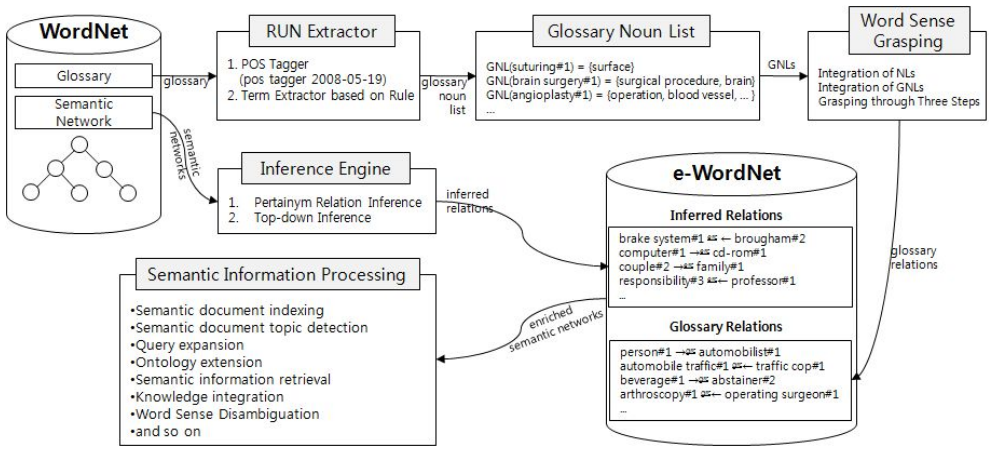
본 연구에서는 본 단원에서 기술한 위키피디아의 카테고리를 이용한 문서 분류 방법을 제안함으로써 웹 문서들 사이의 의미적 네트워크를 형성하고자 하며, 위키피디아 문서 제목(개념)으로 타겟 문서를 태깅함으로써 더욱 의미적이고 정확하며 표준화된 문서 검색 방법을 제공하고자 한다.

D. 선행연구

개념들 사이의 관계성을 측정하기 위해 가장 많이 사용되는 지식베이스는 워드넷이다. 그러나 워드넷은 실세계의 모든 개념 관계 쌍을 포함하지 못한다는 한계점이 지적되었고, 이를 극복하기 위해 본 연구의 선행 연구로써 지식베이스 확장 방법[16]이 제안되었다. 웹에 게시된 문서들을 위키피디아 제목(개념)으로 태깅하는 방법을 제안하는 본 연구는 확장된 워드넷[16]을 개념들의 의미성 파악에 활용한다. 본 단원에서는 워드넷을 확장하는 방법, 결과, 그리고 성능에 대해 상세하게 기술한다.

1. 워드넷 확장 방법

지식베이스는 의미적 정보처리의 기본 데이터로 활용되기 때문에 100% 정확성을 추구해야 한다. 이러한 특징을 반영하기 위해 개념의 정의 구문(Glossary)과 개념들 사이의 관계 자체에 포함된 공리(Axioms)를 이용하는 방법을 통해 워드넷의 개념 관계쌍을 확장하였으며 전체 과정은 [그림 17]과 같다.



[그림 17] 워드넷 확장 방법[16]

워드넷에 정의된 개념을 설명하는 정의 구문은 그 개념과 관계 깊은 명사들을 포함하고

있다. 이에 정의 구문에서 명사들을 추출하고 그 명사에 정확한 의미 번호(Sense Number)를 부여함으로써 그 주인 개념과 관계를 형성하였다. 또한 명사 개념들 사이에 존재하는 관계는 대부분 전이속성(Transitive Property)을 갖고 있다. 이는 관계 $\stackrel{P}{_}$가 전이속성을 갖고 있을 때, 개념 A, B, C사이에 A $-\!>\stackrel{P}{\_}$ B, B $-\!>\stackrel{P}{\_}$ C로 형성되어 있다면, A와 C사이에도 A $-\!>\stackrel{P}{\_}$ C가 형성될 수 있음을 의미한다. 이를 근거로 전이속성을 갖는 관계들을 모두 파악하여 개념 관계쌍을 확장하였다.

2. 워드넷 확장 결과 및 성능

워드넷 확장 방법은 위와 같이 2가지(정의구문과 관계공리)로 구성되는데, 확장된 정도에 따라 베이직 워드넷(Basic WordNet, 워드넷 2.1 자체), 라이트 워드넷(Light WordNet, 워드넷 2.1을 기본적으로 포함하고 정의 구문을 이용하여 확장한 것), 그리고 헤비 워드넷(Heavy WordNet, 라이트 워드넷을 기본적으로 포함하고 워드넷의 관계 공리를 이용하여 확장한 것)으로 구분하였다. <표 7>은 본 과정에 의해 확장된 개념 관계 쌍의 결과를 보이고 있다.

<표 7> 지식베이스 확장을 위한 연구[16]에 의한
개념 관계 쌍 확장 결과

구분	개념 관계 쌍의 수 (개)
워드넷 2.1 (베이직 워드넷)	203,760
정의 구문을 이용한 확장	114,400
공리에 의한 확장	1,430,467
라이트 워드넷	318,160
헤비 워드넷	1,748,627

이와 같이 확장된 워드넷을 사람들이 일반적으로 관계가 있을것으로 생각하는 개념 관계쌍에 대한 포함 정도(Coverage)에 대한 실험과 Senseval-3 문서 집합을 WSD-SemNet(Word Sense Disambiguation-Semantic Network) 알고리즘에 적용한 정확도 평가를 시도하였다. <표 8>과 <표 9>는 그 결과를 보이고 있다.

<표 8> 확장된 워드넷의 실생활 개념 관계 쌍 포함 정도

지식베이스	Coverage (%)
워드넷 2.1	74.25
라이트 워드넷	82.11
헤비 워드넷	89.13
SSI	85.45

<표 9> 확장된 워드넷 기반의 WSD-SemNet 평가(Senseval-3 이용)

지식베이스	정확도(%)	재현율(%)	F1 측정(%)
워드넷 2.1	70.3	74.3	72.2
라이트 워드넷	75.7	78.9	77.2
헤비 워드넷	71.3	85.2	77.7

헤비 워드넷이 실생활 개념 관계 쌍을 포함하는 경우가 가장 높았지만, WSD 평가 결과에서 라이트 워드넷의 정확도가 가장 높은 것으로 확인되었다. 이를 통해 풍부한 개념 관계쌍을 갖는 지식베이스가 오히려 주어진 문맥에 어울리는 의미를 선정할 때 역효과가 있음이 확인되었다. 자연어 처리 분야에서는 높은 재현율보다 높은 정확도를 중요하게 평가한다. 이에 본 논문에서는 개념들 사이의 의미적 관계성을 측정할 때 확장된 워드넷의 라이트 버전(이하 확장된 워드넷은 라이트 버전을 의미 함)을 이용한다.

Ⅲ. 위키피디아 문맥 정보 추출

본 논문은 웹 문서를 위키피디아의 카테고리로의 분류 함으로써 최종적으로 위키피디아 문서 제목으로 태깅하는 방법을 다루고 있다. 본 논문의 핵심을 기술하고 있는 본 장은 위키피디아의 문서 제목(개념)과 관련 깊은 어휘들을 추출하는 방법을 소개한다. 이는 선행연구[87]의 방법을 개선하고 있으며, 본 장에서는 전체적인 절차와 본 논문에서 사용한 용어들을 간략히 소개하고, 전체를 구성하는 사전 처리 방법, 확률을 이용한 키워드 가중치 측정 방법, 확장된 워드넷을 이용한 의미적 가중치 측정 방법, 키워드 가중치 및 의미적 가중치를 동시에 고려한 문맥 가중치 측정 방법, 그리고 문맥 가중치에 의해 확보된 위키피디아 문맥 정보에 대해 상세히 기술한다. 카테고리로 분류 및 태깅 방법은 다음 장에서 기술한다.

A. 위키피디아 문맥 정보 추출을 위한 절차 및 용어 설명

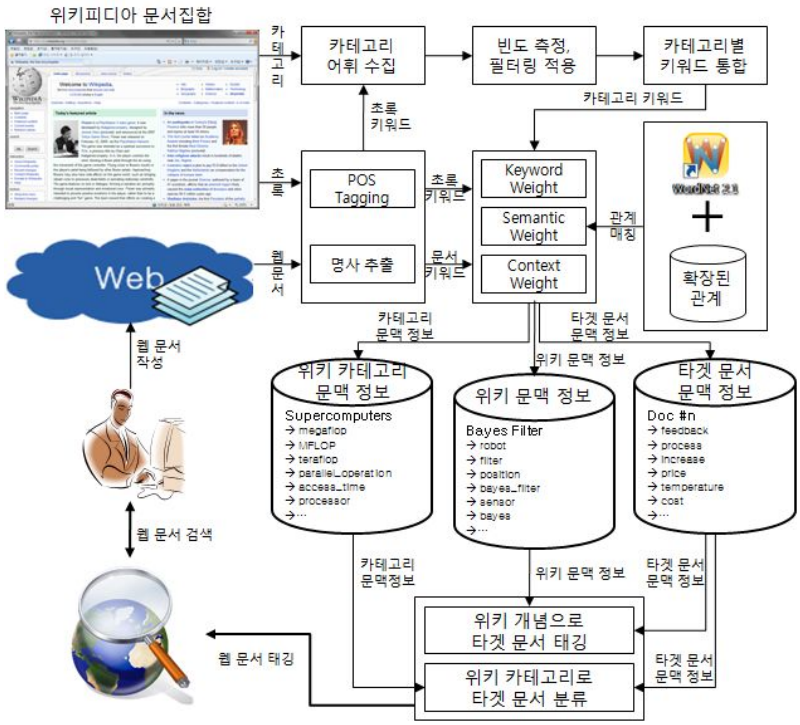
위키피디아의 문맥 정보는 문서의 초록(이하 위키 문서, DBPedia에서 공급하는 3.4버전에 포함된 초록 문서 집합을 이용)에 포함된 명사 어휘들을 의미한다. 이러한 문맥 정보들은 특정 개념(위키피디아 문서 제목을 의미하며 이하 위키 개념으로 사용됨)과 관련된 핵심 어휘들로 웹 문서(이하 타겟 문서)를 태깅할 때 타겟 문서에서 추출된 어휘(이하 타겟 문서 문맥정보)들과 유사성을 판별하는 근거 데이터로 활용된다. 위키 문서에서 문맥 정보를 추출하기 위해 먼저 품사 태깅(POS Tagging, Part-of-Speech Tagging)과정을 거쳐 단일 명사와 복합 명사를 모두 추출하는 전처리 과정을 거친다. 전처리를 통해 추출된 문맥 정보들 사이의 키워드 가중치, 관계성을 기반한 의미적 가중치, 그리고 두 가중치를 함께 반영한 문맥 가중치를 측정하는 과정을 통해 위키 문맥 정보를 형성한다. 여기서 의미적 가중치를 측정할 때 선행연구에서 기술한 확장된 워드넷을 활용한다. 이 과정을 통해 타겟 문서 태깅을 위한 준비 과정을 마친다.

타겟 문서들 사이에 존재할 수 있는 의미적 네트워크 형성을 위해 위키피디아 문서들을

분류하고 있는 카테고리 정보 추출 과정을 포함한다. 그리고 카테고리(DBPedia 3.4버전에 포함된 카테고리 정보를 이용)에 포함된 위키 문서의 문맥 정보를 통합(카테고리 별로 통합된 문맥 정보를 카테고리 문맥 정보라 칭함)하고 문맥 가중치를 재 측정하여 카테고리와 깊이 관련된 어휘들을 순위화 한다. 이에 대한 자세한 설명은 4장에서 다룬다.

또한 타겟 문서를 태깅하기 위해 타겟 문서의 문맥 정보를 추출하는 과정을 포함한다. 이는 위키 문맥 정보를 형성하는 과정과 동일하다.

위의 전 과정을 [그림 18]에서 도식화하고 있다. 본 장에서는 위키 문맥 정보를 추출하기 위한 과정을 예제와 함께 구체적으로 기술하며, 그에 대한 평가는 5장에서 포함한다.



[그림 18] 위키피디아 문맥 정보 추출 및 태깅의 전체 절차

B. 전처리 (Pre-processing)

문서를 기술하는 단어들의 여러 품사 중에서 명사가 문서 주제의 특징을 가장 잘 반영한다는 이유로 여러 연구[9, 13, 16, 21, 46, 47, 64]에서 명사를 추출하여 이용하고 있다. 본 연구에서도 이러한 명사를 추출하기 위한 과정을 포함하며, 이를 위해 스탠포드 대학의 자연어 처리 연구실(The Stanford Natural Language Processing Group)에서 제작하여 배포한 POSTagger 2008-09-28 버전[80, 81]을 이용하였다. 이는 입력받은 문장을 구성하는 각 단어에 적절한 품사 태그를 달아준다. 다음은 위키피디아의 'computer' 문서를 이용한 예를 보이고 있다.

- 입력문장: A computer is a machine that manipulates data according to a set of instructions. Although mechanical examples of computers have existed through much of recorded human history, the first electronic computers were developed in the mid-20th century. (이하 생략)
- 태깅 결과: A/DT computer/NN is/VBZ a/DT machine/NN that/IN manipulates/VBZ data/NNS according/VBG to/TO a/DT set/NN of/IN instructions./NNS Although/IN mechanical/JJ examples/NNS of/IN computers/NNS have/VBP existed/VBN through/IN much/JJ of/IN recorded/VBN human/JJ history,/NN the/DT first/JJ electronic/JJ computers/NNS were/VBD developed/VBN in/IN the/DT mid-20th/JJ century./NN

명사는 하나로 구성된 단일 명사뿐만 아니라 두 단어 이상으로 구성된 복합 명사가 존재한다. 복합 명사는 그 의미가 구체적이고 중의성이 적어 의미적 텍스트 처리에 중요한 단서로 활용된다. 이에, 워드넷이 정의하는 명사 사전과 간단한 규칙을 반영하여 명사 유형(단일 명사, 복합 명사, 고유 명사, 복합 고유명사)을 추출하는 RUN(**Ru**le based **N**oun Extractor) 명사 유형 추출기를 제작하였다. 먼저 태깅된 결과에서 명사(NN, NNS, NNP, NNPS)가 존재하면 명사 앞과 뒤의 연속하는 단어들을 고려하여 가능한 명사 유형을 모두 추출한다. 명사 유형을 추출하기 위한 규칙들은 다음과 같다.

- 명사 유형 추출 규칙: 연속한 명사, 형용사 명사(예. (연속한)형용사 (연속한)명사),

조사와 함께 쓰인 명사 유형 (예. (연속한 또는 단일 형용사) (연속한)명사 조사 (연속한 또는 단일 형용사) (연속한) 명사)

명사에 의미성을 부여하기 위해 추출된 명사 유형들의 원형 파악(Stemming)을 하고 워드넷과 매칭하여 존재하지 않는 것을 제거하는 과정을 거친다. 단, 추출된 명사가 고유명사(NNP 또는 NNPS)일 경우에는 워드넷과 일치하는 명사가 존재하지 않더라도 제거하지 않는다. 다음은 RUN에 의해 추출된 가능한 명사 유형들을 보이며, 진하게 표시된 것은 복합 명사로 가능한 형태를 의미한다.

- 추출된 명사 유형: computer, machine, data, **set of instruction**, set, instruction, **mechanical example**, example, computer, **human history**, history, **first electronic computer**, **electronic computer**, computer, **mid-20th century**, century, ...

추출된 명사 유형에서 복합 명사로 최종 판정이 된다면 그 복합 명사를 구성하는 단어들은 제거한다. 예를 들어, 'first electronic computer', 'electronic computer', 'computer'는 하나의 명사구에서 파생된 것이다. 이를 워드넷과 단어의 수가 많은 것부터 매칭을 한다. 본 예에서 'electronic computer'가 워드넷에 존재하므로 'first electronic computer'와 'computer'는 제거한다. 다음은 워드넷과 매칭을 통해 최종 선정된 명사 유형을 보이고 있다.

- 워드넷과 매칭을 통해 선정된 명사 유형: computer, machine, data, set, instruction, example, computer, history, electronic computer, century, ...

본 과정에서 추출된 명사 유형들은 문맥 정보로 간주되며, 단어가 위키 문서의 키워드로써 확률적 가중치 및 주제로써 의미적 가중치를 측정하는 대상이 된다.

C. 키워드 가중치(Keyword Weight) 측정

키워드 가중치는 추출된 명사들이 해당 위키 개념에 확률적으로 가중치를 얼마나 갖는지를 측정한다. 만약 특정 단어가 출현 빈도가 높다면 위키 개념과 높은 관계성을 가질 수 있다. 이에 가중치 측정을 위해 TF(Term Frequency)를 적용한다. 여기서 TF는 단순히 출현한 횟수를 위키 문서가 포함하는 문맥 정보의 수로 나눈 수치를 의미한다. <표 10>은 앞에서 사용된 위키 문서 'computer'에서 추출된 문맥 정보들의 TF값을 보이고 있다.

<표 10> 위키 문서 'computer'에서 추출된 문맥 정보의 TF 값

문맥 정보	TF	문맥 정보	TF	문맥 정보	TF	문맥 정보	TF
computer	0.167	capacity	0.016	set	0.016	versatility	0.016
instruction	0.033	simple	0.016	program	0.016	player	0.016
time	0.033	circuit	0.016	power	0.016	history	0.016
mobile phone	0.016	room	0.016	principle	0.016	million	0.016
device	0.016	battery	0.016	form	0.016	fraction	0.016
mathematical statement	0.016	personal computer	0.016	information age	0.016	electronic computer	0.016
supercomputer	0.016	ability	0.016	size	0.016	billion	0.016
fighter aircraft	0.016	list	0.016	space	0.016	wristwatch	0.016
machine	0.016	data	0.016	toy	0.016	personal	0.016
calculator	0.016	robot	0.016	capability	0.016	minimum	0.016
storage	0.016	watch	0.016	task	0.016		
icon	0.016	modern	0.016	people	0.016		
MP3	0.016	thesis	0.016	century	0.016		

위와 같이 TF에 의해 측정된 문맥 정보는 노이즈를 포함할 수 있으며, 주제와 상관없이 다른 문서에서도 자주 출현하는 어휘들이 있을 수 있다. 또한 TF 값은 단순히 출현 빈도만을 고려하기 때문에 수치가 동일한 결과를 많이 만든다. 예를 들어, <표 10>에서 'electronic computer'는 'century' 또는 'minimum' 보다 주제의 의미에 가까울 수 있다. 하지만 각각의 TF값은 동일하다. 이는 출현 빈도만을 고려한 TF는 해당 주제와 가까운 정도의 차별화를

주기 어려움을 의미한다.

이러한 한계점을 개선하기 위해 TF와 가장 많이 사용되는 것은 iDF(inverse Document Frequency)이다. 본 연구에서는 이 수식을 변경하여 위키 문서 집합 전체에 포함된 단어의 수를 고려하였다. 즉, 위키 문서 집합 전체에 다수 존재한다면 그 단어는 위키 개념과 확률적으로 관련된 정도가 작아진다는 의미이다. 이를 iTF(inverse Term Frequency)라 칭하였으며 (식 7)과 같다.

$$iTF(t) = \log\left(\frac{|W|}{fr_W(t)}\right) \quad \text{(식 7)}$$

W: 본 연구에 활용된 위키 문서의 수(본 연구에서는 2,944,417),
fr_W: 위키 문서 전체에 출현하는 명사 t의 빈도

iTF는 각 위키 문서가 포함하는 노이즈 판별 및 위키 개념과 관계가 적은 단어를 구별해 줄 수 있다. 앞에서 측정된 TF에 iTF 값을 서로 곱하여 (식 8)과 같이 KW(Keyword Weight)를 측정한다.

$$KW = TF(t) \times iTF(t) \quad \text{(식 8)}$$

<표 11>은 <표 10>의 문맥 정보들을 KW 값에 의해 정렬한 결과를 보이고 있다.

<표 11> 위키 문서 'computer'의 문맥 정보를 KW로 정렬한 결과

문맥 정보	KW	문맥 정보	KW	문맥 정보	KW	문맥 정보	KW
computer	0.362	billion	0.056	million	0.045	size	0.035
mathematical statement	0.090	Personal	0.054	Modern	0.044	set	0.034
instruction	0.090	robot	0.054	task	0.044	list	0.031
Information Age	0.080	fraction	0.053	circuit	0.042	power	0.031
electronic computer	0.076	watch	0.053	principle	0.041	program	0.029
wristwatch	0.076	toy	0.051	machine	0.041	form	0.028
MP3	0.064	icon	0.050	room	0.040	player	0.027
supercomputer	0.063	thesis	0.050	device	0.040	history	0.026
calculator	0.063	Simple	0.050	ability	0.038	century	0.025
versatility	0.061	battery	0.048	capacity	0.038	people	0.024
fighter aircraft	0.059	minimum	0.047	time	0.037		
mobile phone	0.058	capability	0.046	data	0.035		
personal computer	0.058	storage	0.046	space	0.035		

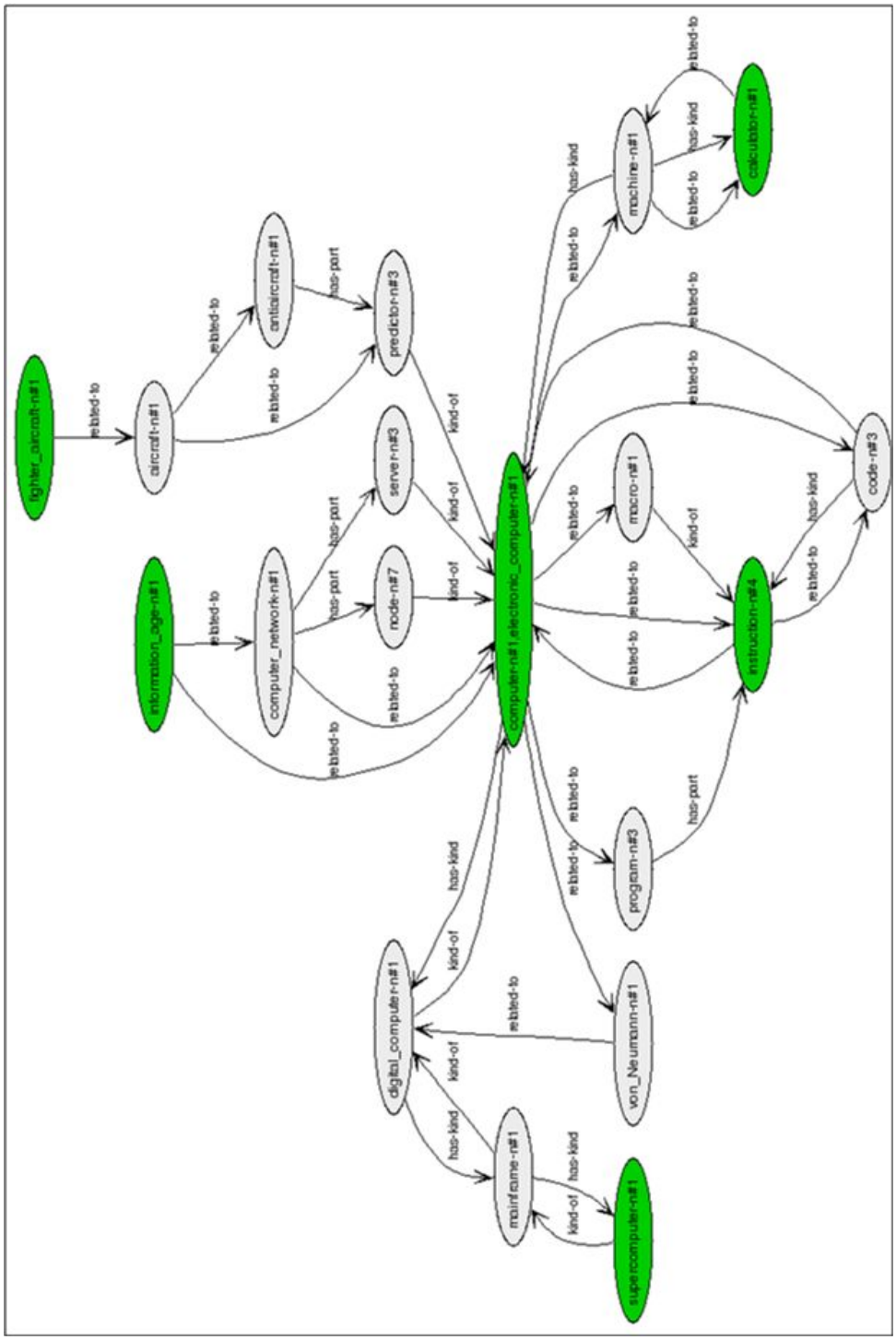
<표 11>에서 보는 바와 같이 iTF를 적용한 것만으로도 위키 개념과 가까운 순서대로 어느 수준 정렬할 수 있는 것처럼 보인다. 이렇게 측정된 키워드 가중치에 의미성을 부여하기 위해 다음 단원에서 확장된 워드넷 기반의 의미적 가중치 측정 방법에 대해 기술한다.

D. 의미적 가중치(Semantic Weight) 측정

의미적 가중치는 추출된 문맥 정보가 얼마나 그 주제와 의미적으로 가까운지를 측정하는 것이다. 만약 어떤 문서가 하나의 주제에 대해 기술하고 있다면, 그 문서에는 그 주제를 설명하기 위해 의미적으로 가까운 단어들을 많이 포함할 것이다. 즉, 주제와 가까운 단어들은 서로 관계를 형성할 수 있는데, 그 관계성이 높은 단어는 그 문서를 의미적으로 대표할 수 있을 가능성이 있다는 근거를 바탕으로 하고 있다. [그림 19]는 <표 11>에 KW값에 따른 상위 10개 단어('MP3'는 워드넷에 정의되어 있지 않으므로 제외함)에 대해 SSI(Structural Semantic Interconnections) 알고리즘[19]을 기반으로 추출한 의미적 관계 그래프를 보이고 있다. 그래프에서 'computer#1 / electronic computer#1'를 중심으로 다른 단어들과의 관계성이 형성되는 것을 확인할 수 있는데, 이 때 'computer#1 / electronic computer#1'가 의미적으로 주제에 가까울 수 있음을 내포한다. 이에 여러 연구[2, 14, 16, 20, 21, 30, 31, 32, 33, 34, 36, 37, 38, 60, 78]에서 다양한 의미적 가중치 측정 방법을 제안하고, 각 방법을 의미적 정보 처리에 활용하고 있다.

본 연구에서는 확장된 워드넷[16]을 바탕으로 단어들 사이의 의미적 관계성을 추출하고, 형성된 관계성에서 노드(node)기반 가중치 측정 방법[2, 16, 20, 21]을 이용한다. 각 단어는 하나 이상의 의미를 포함하고 있다. 이에 먼저 각 단어가 갖는 개념을 워드넷과 매칭하여 추출한다. 본 연구의 이해를 위해 <표 11>에 있는 10개의 단어(상위 10개의 단어들은 대부분 단의어(monosemy)임. 수식의 설명을 위해 다의어(polysemy) 9개와 단의어 1개를 선정함)를 주어진 문맥으로 간주하고 예로 든다. 10개 단어(computer, capacity, instruction, program, circuit, principle, battery, list, data, storage)에 대해 워드넷에 정의된 개념 리스트(SL, sense(concept) list)들은 다음과 같다:

$SL_{computer}$ = {computer#1, computer#2},
$SL_{capacity}$ = {capacity#1, capacity#2, ... , capacity#9},
$SL_{instruction}$ = {instruction#1, instruction#2, instruction#3, instruction#4},
$SL_{program}$ = {program#1, program#2, ... , program#8},
$SL_{circuit}$ = {circuit#1, circuit#2, ... , circuit#6},
$SL_{principle}$ = {principle#1, principle#2, ... , principle#6},
$SL_{battery}$ = {battery#1, battery#2, ... , battery#7},



[그림 19] <표 11>의 상위 10개에 대한 의미적 관계성 (SSI 알고리즘[19] 이용)

SL_{list} = {list#1, list#2},

SL_{data} = {data#1},

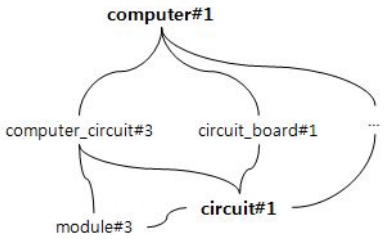
$SL_{storage}$ = {storage#1, storage#1, ... , storage#6}.

위와 같이 각 단어의 개념들을 추출하고, 각 개념이 다른 단어의 개념들과 갖는 관계성(relatedness)을 측정한다. 이때 확장된 워드넷(관계를 중심으로 확장되었으며 Ⅱ-D에 상세히 기술됨)을 기반으로 (식 9)를 적용한다. 이 수식에서 두 개념 사이의 관계는 거리 5까지 형성되는 것만 고려하며, 5를 넘는 관계는 없는 것으로 간주한다. 이는 기존에 수행된 지식베이스에 형성된 노드기반의 유사도 측정 연구[2, 13, 16, 19, 20, 21]들이 일반적으로 활용하는 방법이다.

$$relatedness(s_{ia}, s_{jb}) = \frac{1}{\arg_{s_{jb} \in SL_j} \min(dist(s_{ia}, s_{jb}))}, i \neq j \qquad \text{(식 9)}$$

s_{ia}: SL에 포함된 j번째 개념,
min: 최소값,
dist: 두 개념 사이의 거리

(식 9)는 두 개념 사이에 형성된 관계들에서 가장 가깝게 형성된 하나를 선택하여 역수를 취하는 것이다. 두 개념을 연결하는 노드가 많을수록 두 개념 사이의 거리는 멀어짐(반비례 관계)을 의미한다. 두 개념 'computer#1'과 'circuit#1'을 예로 들어 수식을 설명한다. 두 개념 사이의 관계성은 [그림 20]과 같이 형성될 수 있다.



[그림 20] 확장된 워드넷 기반의 'computer#1'과 'circuit#1' 사이의 관계도

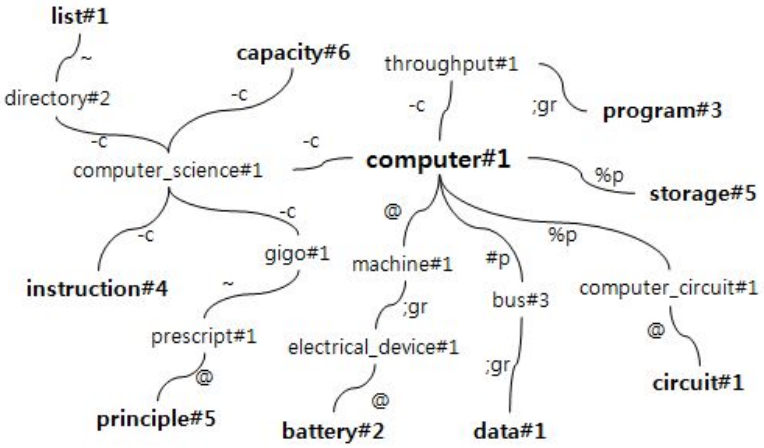
실제로 워드넷 내에는 두 개념 사이에 더욱 많은 관계를 포함하지만 (식 9)를 간략히 설명하기 위해 나머지는 생략하였다. [그림 20]과 같이 형성된 관계에서 추출할 수 있는 관계는 크게 'circuit#1 -> module#3 -> computer_circuit#3 -> computer#1'과 'circuit#1 -> computer_circuit#3 -> computer#1'라 할 수 있다. 여기서 두 개념을 연결하는 노드의 수가 가장 적은 것('circuit#1 -> computer_circuit#3 -> computer#1')을 선택하고, 포함된 노드 수의 역수를 취한다. 즉, 두 개념 사이의 관계성은 1/3(≒0.333)으로 계산될 수 있다. 이처럼 (식 9)는 특정한 두 개념 사이의 관계성만을 측정한다.

하나의 문맥에는 여러 단어들로 구성되기 때문에 주제와 가까운 개념은 다른 여러 개념들과도 관계가 많이 형성될 수 있다. (식 9)에서는 두 개념 사이의 관계성만을 측정했다면, (식 10)에서는 특정 개념이 갖는 다른 개념들과의 관계성을 계산한다. 이를 개념의 의미적 가중치(sw, semantic weight)라 칭한다.

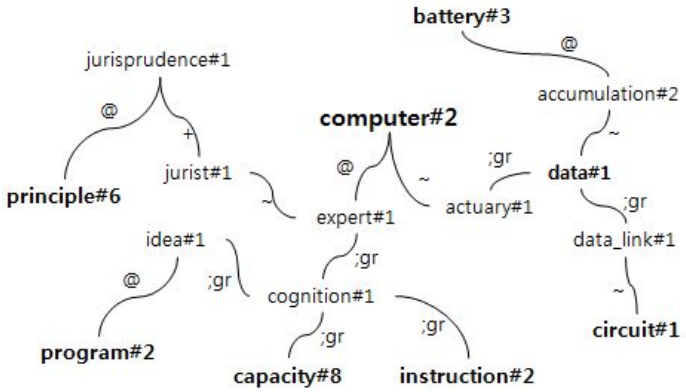
$$sw(s_{ia}) = \sum_{j=1}^{n} \arg_{s_{jb} \in SL_j} \max(relatedness(s_{ia}, s_{jb})), i \neq j \qquad \text{(식 10)}$$

max: 최대값

앞의 10개 단어에서 'computer'를 예로 들겠다. 워드넷에는 'computer#1'과 'computer#2' 두가지 개념으로 정의를 하고 있다. 각 개념이 다른 개념들과 갖는 관계성은 [그림 21]과 [그림 22]에서 보이고 있다.



[그림 21] 'computer#1'이 갖는 다른 개념들과의 관계성



[그림 22] 'computer#2'가 갖는 다른 개념들과의 관계성

[그림 21]과 [그림 22]처럼 형성된 각 개념들 사이의 관계성과 의미적 가중치를 (식 9)와 (식 10)에 의해 각각 <표 12>와 같이 계산된다.

<표 12> 개념의 관계성(relatedness)과 의미적 가중치(semantic weight)

computer#1		computer#2	
storage#5	0.5		
data#1	0.333	data#1	0.33333334
list#1	0.25	battery#3	0.2
battery#2	0.25	principle#6	0.2
principle#5	0.2	circuit#1	0.2
circuit#1	0.333	program#2	0.2
program#3	0.333	instruction#2	0.25
instruction#4	0.333	capacity#8	0.25
capacity#6	0.333		
sw(computer#1) = 2.866		sw(computer#2) = 1.633	

<표 12>와 같이 각 개념은 의미적 가중치를 갖게 된다. 이는 예에 사용된 10개 단어의 각 개념 모두에게 해당되며, 한 단어가 표현할 수 있는 개념들 중에서 최대 의미적 가중치를 갖는 하나를 선택하여 그 단어의 의미적 가중치로 선택한다. (식 11)은 본 절차를 표현하고 있다.

$$sw(t_i) = \arg_{s_{ia} \in SL_i} \max(sw_{ia}) \qquad \text{(식 11)}$$

(식 11)에 의해 단어 'computer'의 개념으로 'computer#1'이 선택되며, 의미적 가중치 또한 그 개념이 갖는 2.866으로 결정된다. 본 결과에서 보이는 바와 같이 의미적 가중치를 측정하는 부분에서 단어가 해당 문맥에 적절한 의미를 결정하는 WSD(Word Sense Disambiguation) 과정이 암묵적으로 포함되는데, 이는 이미 관계 구조 기반 WSD 연구[89]에서 그 성능을 인정받았으며, 이를 응용한 지식베이스 관계 확장 방법[16]에서 제안한 WSD-SemNet 알고리즘의 기반이 되기도 하였다. 이러한 과정에 의해 주어진 문맥에 포함된 10개 단어의 의미적 가중치와 개념은 <표 13>과 같이 결정될 수 있다.

<표 13> 주어진 문맥의 의미적 가중치와 대표 개념

단어	개념 번호	의미적 가중치	워드넷 정의 구문
computer	1	2.866	a machine for performing calculations automatically
capacity	6	2.367	(computer science) the amount of information (in bytes) that can be stored on a disk drive
instruction	4	2.733	(computer science) a line of code written as part of a computer program
program	3	2.733	(computer science) a sequence of instructions that a computer can interpret and execute
circuit	1	2.200	an electrical device that provides a path for electrical current to flow
principle	5	1.833	rule of personal conduct
battery	2	1.933	a device that produces electricity; may have several primary or secondary cells arranged in parallel or series
list	1	2.200	a database containing an ordered array of items (names or topics)
data	1	2.750	a collection of facts from which conclusions may be drawn; "statistical data"
storage	5	2.400	an electronic memory device; "a memory and the CPU form the central part of a computer ... "

<표 13>과 같이 하나의 주제를 기술하는 문서에서 추출된 명사들 사이에, 의미적 가중치는 특히 그 주제와 깊이 관련된 어휘들을 의미적으로 추출할 수 있다. 또한 각 단어가 그 문맥에서 의미하는 개념을 파악함으로써 단어 기반이 아닌 의미 기반의 문서 검색 및 분류가 가능하게 한다. 이와 같이 측정한 의미적 가중치는 다음 단원에서 문맥 가중치를 측정하기 위해 키워드 가중치와 함께 사용된다.

E. 문맥 가중치(Context Weight) 측정 및 문맥 정보 형성

앞의 과정에서 위키 개념의 문맥 정보에 대해 키워드 가중치와 의미적 가중치를 측정하였다. 문맥 정보에서 위키 개념과 가까운 정도를 계산하는 문맥 가중치(CW, context weight)는 두 값을 모두 반영하는 (식 12)를 이용한다.

$$cw(t_i) = kw(t_i) + kw(t_i) \times sw(t_i) \quad \text{(식 12)}$$

문맥 가중치를 측정하는 수식은 키워드 가중치(KW)를 그대로 보존하면서 의미적 가중치(SW) 값에 따라 그 가중치를 조절하기 위한 것으로, 워드넷에 정의되지 않은 어휘들의 가중치까지 측정할 수 있다는 점에서 다양한 연구[2, 20, 21]에서 활용되었다. 본 연구에서는 (식 13)을 이용하여 각 문서의 문맥 가중치를 최대값으로 정규화 하는 과정을 포함한다. 이는 웹 문서의 분류 및 위키 개념 태깅을 위해 단어의 수에 영향을 받지 않기 위함이다.

$$n-cw(t_i) = \frac{cw(t_i)}{\arg_{t_k \in WC_j} \max(cw(t_k))} \quad \text{(식 13)}$$

n-cw: 정규화된 문맥 가중치(normalized context weight),
WC_i: 위키 문서 i에 대한 문맥 정보(Wiki Context)

<표 14, 15, 16, 17, 18>은 본 과정을 통해 측정된 위키 개념(문서 제목) 'computer'[82], 'A* search algorithm'[83], 'Java'[84], 'Apple Inc.'[85], 'Amit Sheth'[86]에 대해 문맥 가중치에 따라 정렬된 상위 10개의 문맥 정보를 각각 보이고 있다. 각 표에서 WS는 워드넷 센스(WordNet Sense) 번호를 의미하고, KW, SW, CW, n-CW는 각각 키워드 가중치, 의미적 가중치, 문맥 가중치, 정규화된 문맥 가중치를 나타낸다.

<표 14> 위키 개념 'computer'에 대한 문맥 정보

위키 개념	문맥 정보	WS	KW	SW	CW	n-CW
computer	computer	1	0.362	10.600	4.196	1
	electronic computer	1	0.076	10.600	0.885	0.211
	instruction	4	0.090	6.800	0.703	0.167
	device	1	0.040	9.400	0.412	0.098
	mathematical statement	1	0.090	3.517	0.408	0.097
	fighter aircraft	1	0.059	5.733	0.400	0.095
	machine	1	0.041	8.800	0.398	0.095
	calculator	2	0.063	5.233	0.393	0.094
	storage	5	0.046	7.300	0.379	0.090
	personal computer	1	0.058	6.533	0.362	0.086

<표 15> 위키 개념 'A* search algorithm'에 대한 문맥 정보

위키 개념	문맥 정보	WS	KW	SW	CW	n-CW
A* search algorithm	node	7	0.421	4.050	2.127	1.000
	function	1	0.200	3.550	0.912	0.429
	distance	1	0.165	2.600	0.593	0.279
	algorithm	1	0.160	2.267	0.524	0.246
	graph	1	0.121	2.450	0.416	0.196
	heuristic	1	0.158	1.433	0.385	0.181
	computer science	1	0.054	5.850	0.371	0.174
	search	1	0.096	2.400	0.327	0.154
	cost	1	0.085	2.717	0.315	0.148
	edge	3	0.086	2.300	0.283	0.133

<표 16> 위키 개념 'Java'에 대한 문맥 정보

위키 개념	문맥 정보	WS	KW	SW	CW	n-CW
Java	Java	3	0.389	2.167	1.231	1.000
	technology	2	0.099	4.167	0.511	0.415
	computer architecture	2	0.105	3.567	0.478	0.388
	Sun	1	0.184	0.750	0.323	0.262
	programming language	1	0.076	3.183	0.320	0.260
	library	4	0.059	3.583	0.273	0.221
	component	3	0.059	3.633	0.271	0.220
	core	2	0.062	3.300	0.268	0.217
	syntax	1	0.085	2.100	0.264	0.214
	specification	1	0.074	2.450	0.256	0.208

<표 17> 위키 개념 'Apple Inc.'에 대한 문맥 정보

위키 개념	문맥 정보	WS	KW	SW	CW	n-CW
Apple Inc.	software	1	0.089	7.400	0.750	1.000
	consumer	1	0.105	5.883	0.721	0.960
	company	1	0.099	5.183	0.615	0.819
	software product	1	0.102	4.767	0.589	0.785
	hardware	3	0.075	6.333	0.552	0.735
	electronics	1	0.079	5.783	0.536	0.714
	computer	1	0.058	8.083	0.526	0.701
	audio	3	0.072	4.433	0.391	0.521
	retail store	1	0.048	6.017	0.340	0.453
	Apple	1	0.162	1.100	0.340	0.453

<표 18> 위키 개념 'Amit Sheth'에 대한 문맥 정보

위키 개념	문맥 정보	WS	KW	SW	CW	n-CW
Amit Sheth	Georgia	1	0.284	1.283	0.648	1.000
	Dayton	1	0.205	1.483	0.508	0.783
	Athens	2	0.162	1.283	0.370	0.571
	Lexis Nexis Ohio Eminent Scholar	0	0.369	0.000	0.369	0.569
	Amit Sheth	0	0.369	0.000	0.369	0.569
	Knoesis	0	0.369	0.000	0.369	0.569
	Advanced Data Management	0	0.369	0.000	0.369	0.569
	Ohio	1	0.136	1.583	0.351	0.542
	Information System	1	0.266	0.200	0.319	0.492
	computer scientist	1	0.210	0.400	0.295	0.454

이후부터의 'CW'는 정규화된 문맥 가중치(n-CW)로 활용됨.

<표 18>에서 WS가 '0'인 것은 워드넷에 정의되지 않은 단어를 의미한다. 위와 같이 추출된 문맥 정보들은 웹 문서의 분류와 위키 개념 태깅에 중요한 단서로 활용되며, 웹 문서에서 주요 키워드들을 추출하는 방법도 동일한 과정을 따른다. 다음 장에서는 본 장에서 추출한 위키 개념의 문맥 정보를 기반하여 웹 문서의 분류와 태깅 방법에 대해 기술한다.

Ⅳ. 의미적 문서 태깅 방법

Ⅲ장에서 생성한 위키 개념에 대한 문맥 정보를 이용하여 본 논문의 가장 중요한 부분인 웹 문서의 의미적 태깅 방법을 제안한다. 이를 위해 타겟 문서(웹 문서)에서 키워드(타겟 문서 문맥정보)들의 문맥 가중치 측정, 위키피디아의 카테고리 형성 및 카테고리에 따른 문맥 정보 형성에 대해 기술하고, 본 연구의 궁극적인 목적인 카테고리로의 문서 분류와 위키 개념으로 타겟 문서를 태깅하는 방법을 설명한다.

A. 타겟 문서의 문맥 정보 추출

웹 문서(타겟 문서)의 분류와 위키 개념 태깅을 위해 타겟 문서의 핵심이 되는 문맥 정보를 추출하고 가중치를 계산하는 과정이 필요하다. 이 과정은 Ⅲ장에서 기술한 위키 문서를 분석하는 과정과 동일하게 진행된다. 본 과정의 이해를 위해 지식베이스 관계 확장에 관한 논문[16]에서 초록, 서론 및 결론을 이용하여 예를 보이며, <표 19>는 그 논문에서 추출한 내용의 일부를 보이고 있다.

<표 19> 지식베이스 관계 확장 논문[16]에서 추출한 내용의 일부

The most fundamental step in semantic information processing (SIP) is to construct knowledge base (KB) at the human level; that is to the general understanding and conception of human knowledge. WordNet has been built to be the most systematic and as close to the human level and is being applied actively in various works. In one of our previous research, we found that a semantic gap exists between concept pairs of WordNet and those of real world. This paper contains a study on the enrichment method to build a KB. We describe the methods and the results for the automatic enrichment of the semantic relation network. A rule based method using WordNet's glossaries and an inference method using axioms for WordNet relations are applied for the enrichment and an enriched WordNet (E-WordNet) is built as the result. Our experimental results substantiate the usefulness of E-WordNet. An evaluation by comparison with the human level is attempted. (이하 생략)

위키 개념의 문맥 정보를 추출하는 과정과 같이 품사 태깅 과정을 거쳐 RUN 추출기를 통해 명사 유형을 추출한다. 추출된 명사 유형들을 워드넷과 매칭하고, 최종적으로 선정된 명사 유형들의 키워드 가중치를 측정한다. 키워드 가중치에서 iTF 계산은 위키피디아 분석을 통해 얻은 수치를 이용하였다. 이는 실험 군이 다르지만 대용량의 문서 집합을 이용하여 이미 각 단어의 iTF 값을 계산하였기 때문에 실험 군에 따른 오차는 적을 것으로 간주한다. <표 20>은 <표 19>에서 추출된 명사들을 키워드 가중치에 따라 내림차순으로 보이고 있다.

<표 20> <표 19>에서 추출된 명사들의 키워드 가중치

키워드	KW	키워드	KW	키워드	KW
WordNet	0.304	SSI	0.050	paper	0.026
KB	0.191	knowledge	0.049	type	0.026
E-WordNet	0.156	result	0.046	gap	0.025
WSD-SemNet	0.090	network	0.040	interconnection	0.023
SIP	0.089	glossary	0.037	indexing	0.023
relation	0.078	evaluation	0.036	web	0.023
enrichment	0.076	document	0.036	usefulness	0.022
WSD	0.076	inference	0.033	extension	0.022
method	0.074	axiom	0.032	performance	0.022
real_world	0.065	level	0.032	retrieval	0.022
knowledge base	0.063	disambiguation	0.029	query	0.021
concept	0.058	machine	0.029	relationship	0.019
research	0.055	information	0.028	detection	0.018

또한 추출된 명사들 사이의 의미적 가중치를 측정하여 키워드 가중치와 함께 문맥 가중치를 측정한다. <표 21>은 추출된 명사들의 의미적 가중치와 정규화된 문맥 가중치를 보이고 있다.

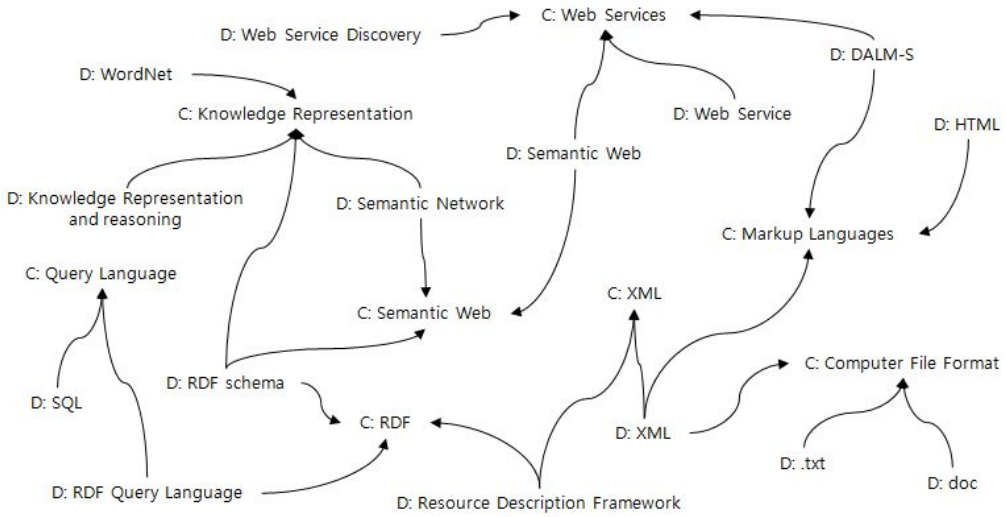
<표 21> <표 19>에서 추출된 명사들의 의미적 가중치와 문맥 가중치

문맥 정보	SW	CW	문맥 정보	SW	CW	문맥 정보	SW	CW
relation	20.07	1.00	document	12.35	0.26	manner	14.37	0.13
WordNet	4.78	0.96	machine	15.13	0.25	axiom	5.82	0.12
concept	18.67	0.62	result	8.30	0.23	usefulness	8.50	0.12
knowledge	21.48	0.60	level	12.17	0.23	point	12.93	0.12
knowledge base	12.83	0.47	type	14.62	0.22	paper	6.90	0.11
method	8.30	0.38	glossary	7.73	0.18	information processing	15.03	0.11
KB	2.48	0.36	interconnection	11.88	0.17	extension	8.02	0.11
research	10.38	0.34	relationship	13.87	0.16	indexing	7.43	0.11
real world	8.40	0.33	retrieval	12.20	0.16	conception	18.67	0.10
inference	14.55	0.28	web	11.52	0.15	general knowledge	13.25	0.10
network	11.53	0.28	gap	9.75	0.15	word	15.77	0.10
algorithm	5.65	0.28	order	17.00	0.15	pair	11.42	0.09
information	16.65	0.27	performance	10.38	0.14	people	18.22	0.09

위와 같은 방법으로 타겟 문서가 포함하는 문맥정보의 문맥 가중치를 측정함으로써 위키피디아 카테고리로 분류와 위키 개념 태깅에 중요한 단서로 활용한다. 이를 위한 과정은 다음 단원부터 설명한다.

B. 위키피디아 카테고리 형성 및 문맥 정보 형성

위키피디아 카테고리(이하 위키 카테고리)는 위키 문서들이 기술하는 주제(제목)보다 일반적인 개념으로 형성되며, 여러 위키 문서들을 분류하고 있다. 또한 각 위키 문서는 하나 이상의 위키 카테고리에 속해 있다. 예를 들어 위키 문서 'RDF schema'는 'RDF', 'Knowledge Representation', 'Semantic Web' 등의 위키 카테고리로 분류되고 있다. 이렇듯 위키 카테고리와 위키 문서 사이는 포함 관계를 형성하고 있으며 [그림 23]은 이에 대한 일부를 도식화 하고 있다.



[그림 23] 위키 문서(개념)들의 위키 카테고리 분류 (C: Category, D: Document)

이러한 위키 카테고리로 먼저 분류하는 것은 타겟 문서를 조금 더 일반적인 개념으로 분류함으로써 타겟 문서들 사이의 의미적 네트워크를 형성하고, 위키 개념 태깅의 정확도를 높이기 위함이다. 예를 들면, <표 21>의 내용을 포함하는 타겟 문서가 있을 때, 먼저 위키 개념 'WordNet', 'Semantic Network' 또는 'Knowledge Representation and reasoning'을 찾는 것보다 'Knowledge Representation' 이라는 카테고리를 파악하는 것이 용이하기 때문이다. 또한 'Knowledge Representation'에 속한 타겟 문서들과 'Semantic Web'에 속한 타겟 문서들 사이의 의미적 네트워크 형성도 용이하다.

타겟 문서를 적절한 위키 카테고리로 분류하기 위해 해당 카테고리와 깊이 관련된 문맥 정보를 형성하는 과정이 필요하다. 이미 앞에서 위키 개념에 대한 문맥 정보를 형성하였다. 이러한 위키 문맥 정보를 해당 카테고리에 통합하고, 통합된 문맥 정보의 키워드 가중치와 의미적 가중치를 측정함으로써 최종적으로 각 카테고리에 대한 문맥 가중치를 계산한다. 이때 하나의 위키 문맥 정보(예. 'Semantic Network')는 여러 카테고리(예. 'Semantic Web', 'Knowledge Representation' 등)에 중복으로 활용될 수 있다. <표 22>는 본 과정을 통해 습득된 위키 카테고리 'Knowledge Representation'에 대한 문맥 정보의 키워드 가중치, 의미적 가중치, 그리고 문맥 가중치를 보이고 있다.

<표 22> 위키 카테고리 'Knowledge Representation'에 대한 문맥 정보

키워드	KW	SW	CW	키워드	KW	SW	CW
synset	1.000	45.350	1.000	artificial intelligence	0.026	50.500	0.029
lexical database	1.000	27.133	0.607	informatics	0.020	64.050	0.028
information processing system	0.333	74.933	0.546	knowledge base	0.024	52.000	0.027
conceptualisation	0.500	40.417	0.447	fuzzy logic	0.071	16.533	0.027
polysemous word	1.000	13.583	0.315	cognitive science	0.016	74.983	0.027
universal quantifier	0.333	39.317	0.290	natural language processing	0.042	28.733	0.027
Reification	1.000	11.017	0.259	Peirce	0.043	26.600	0.026
first blush	1.000	8.967	0.215	library science	0.038	28.000	0.024
nomogram	0.500	13.967	0.161	hypermedia system	0.250	3.450	0.024
WordNet	0.200	25.700	0.115	knowledge	0.012	91.533	0.024
Alfred North Whitehead	0.125	31.283	0.087	metadata	0.025	40.917	0.023
logical implication	0.333	10.067	0.080	Web personalization	1.000	0.000	0.022
SUMO	0.160	21.967	0.079	Adaptive hypermedia	1.000	0.000	0.022
Object-oriented programming language	0.063	50.900	0.070	Teknowledge Corporation	1.000	0.000	0.022
coreference	0.333	8.117	0.066	SUO-KIF	1.000	0.000	0.022
metalanguage	0.167	12.117	0.047	Articulate Software	1.000	0.000	0.022
NLP	0.071	28.733	0.046	Suggested Upper Merged Ontology	1.000	0.000	0.022
utilisation	0.042	40.983	0.038	CDS/ISIS	1.000	0.000	0.022
data structure	0.033	45.700	0.034	OpenIsis	1.000	0.000	0.022
ontology	0.155	8.383	0.031	ISIS DLL	1.000	0.000	0.022
local area network	0.038	36.217	0.031	HTML Web	1.000	0.000	0.022
natural language	0.029	45.283	0.029	GenISIS	1.000	0.000	0.022

<표 22>와 같이 위키 카테고리는 대부분 특정 개념에 대해 일반화된 의미로 이루어져 있지만, 그 일반화의 차이가 아주 큰 것도 존재한다. 예를 들어, 위키 개념 'quick sort'를 포함하는 카테고리로는 'sorting algorithms', 'comparison sorts', 'Articles with example pseudocode', '1961 in science'를 포함하며, 컴퓨터 과학자인 'Amit Sheth'의 카테고리는 'living people', 'indian computer scientists', 'Ohio State University alumni', 'Write State University faculty'를 포함한다. 이러한 예에서 'quick sort'에 대한 카테고리로 'Articles with example pseudocode' 및 '1961 in science' 그리고 'Amit Sheth'에서 'living people'은 카테고리로써의 성격보다는 해당 문서의 특징이나 위키 개념의 일시적인 성격을 의미하기 때문에 적절지 않다. 또한 여러 위키 개념의 문맥 정보를 통합하여 위키 카테고리 문맥 정보가 형성되기 때문에 문맥 정보의 수가 아주 많아진다. 이에 간단한 조건을 두어 불필요한 것들을 필터링 하는 과정이 필요하며, 이러한 필터링에 대한 것은 실험 부분에서 구체적으로 기술하도록 한다.

C. 카테고리로 문서 분류 및 의미적 문서 태깅

앞의 과정에서 위키 개념의 문맥 정보, 위키 카테고리의 문맥 정보, 그리고 타겟 문서에서 문맥 정보를 추출하는 방법을 제안하였다. 본 단원에서는 이러한 정보들을 바탕으로 본 논문에서 핵심이 되는 타겟 문서의 위키 카테고리 분류와 의미적 문서 태깅 방법에 대해 기술한다. 이를 위해 확장된 워드넷을 다시 이용하고, 각 문맥 정보 사이에 포함된 개념들 사이의 관계성을 측정한다. 하지만, 각 문맥 정보는 워드넷이 정의하는 명사 유형(예. <표 22>에서 'synset', 'lexical database' 등)과 그렇지 않은 명사 유형(예. <표 22>에서 'SUO-KIF', 'Articulate Software' 등)을 포함한다. 이는 워드넷이 정의하는 명사 유형들끼리는 의미적 관계성을 측정할 수 있지만, 그렇지 않은 경우는 관계성을 측정할 수 없음을 의미한다. 이에, 문맥 정보들(위키 카테고리 문맥정보와 타겟 문서 문맥정보, 또는 위키 개념 문맥 정보와 타겟 문서 문맥정보) 사이의 유사성을 측정하기 위해 두가지 사항을 각각 반영할 수 있는 방법을 제안한다. 먼저 워드넷이 정의하는 명사 유형을 위한 수식은 (식 14)에서 보이고 있다.

$$sim(ci_k, ci_l) = \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} \frac{CW(s_{(k,i)}) + CW(s_{(l,j)})}{2 \times \min dist(s_{(k,i)}, s_{(l,j)})}, k \neq l \qquad \text{(식 14)}$$

ci_k: 타겟 문서의 문맥정보,
ci_l: 위키 개념 또는 위키 카테고리의 문맥정보(ci, context information),
n: 해당 문맥 정보에 속한 개념의 수,
CW: 정규화된 문맥 정보 가중치(Context Weight),
dist: 확장된 워드넷에서 두 개념 사이의 거리(distance),
s: 특정 문맥 정보에 속한 개념

(식 14)는 k 문맥 정보와 l 문맥 정보에 각각 포함되는 개념 i 와 j 사이에 관계가 형성될 때(두 개념 사이에 형성되는 거리가 5이하일 때), 두 개념의 문맥 가중치의 합에 대한 평균을 내는 것이다. 이때, 두 개념 사이의 관계는 거리에 반비례하는 특징을 반영하기 위해 최소 거리 값으로 나눈다. 이 수식은 의미적 문서 연결(SDI, Semantic Document Interconnections)에 대한 연구[2]에서 제안되어 그 성능을 이미 인정받은 바 있다.

또한 워드넷에 정의되지 않은 개념들은 단어기반 매칭을 이용한다. 이는 두 값이 갖는 문맥 가중치의 평균을 이용한 것으로 (식 15)와 같다.

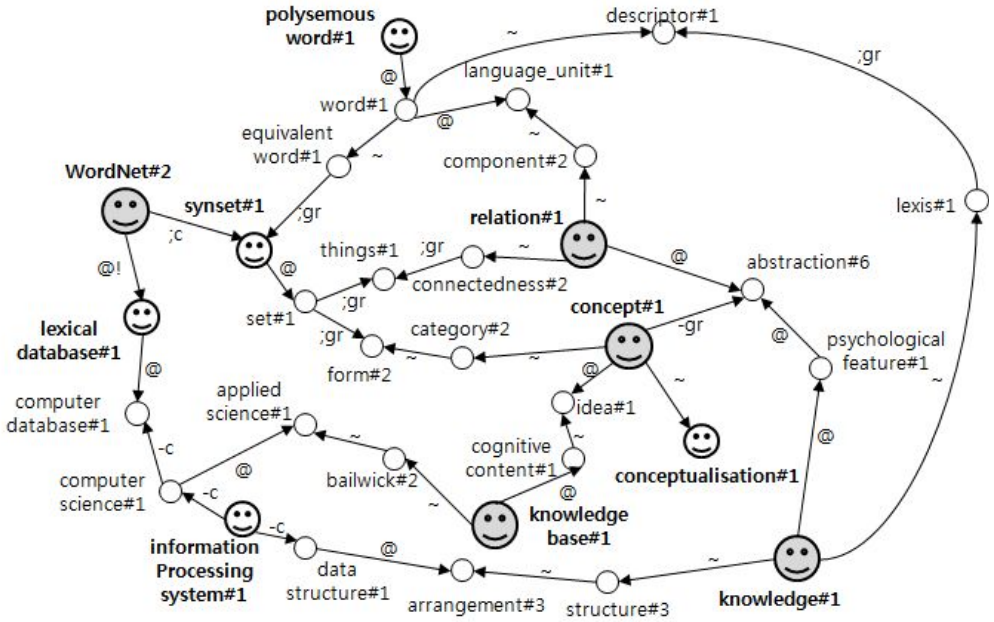
$$sim(ci_k, ci_l) = \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} \frac{CW(s_{(k,i)}) + CW(s_{(l,j)})}{2}, k \neq l \qquad \text{(식 15)}$$

예를 들어 설명하면, 두 카테고리 'knowledge representation'과 'query language'와 타겟 문서 'Doc 1'이 존재하고, 각각에서 추출된 문맥 정보는 <표 23>과 같이 각각 5개의 개념을 포함한다고 먼저 가정하겠다. 본 가정에서 워드넷에 정의되지 않은 개념은 수식이 직관적이므로 포함하지 않는다.

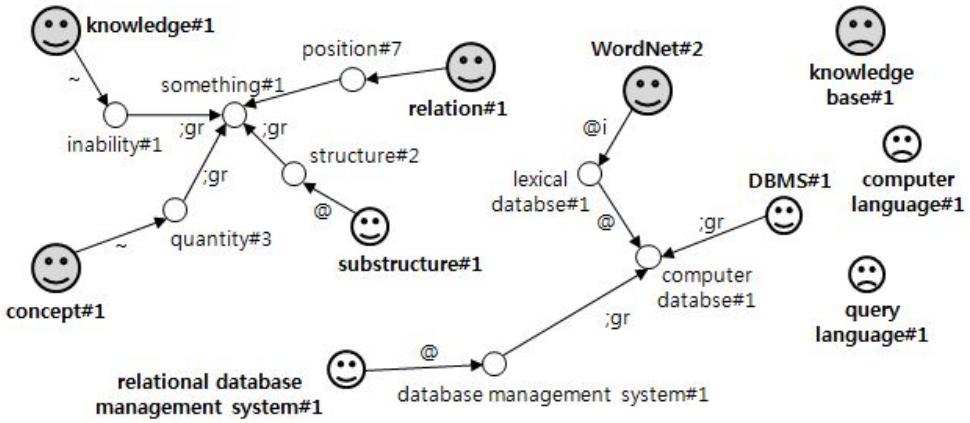
<표 23> 카테고리 분류를 위한 예시

Knowledge Representation			Query Language			Doc 1		
문맥 정보	#	CW	문맥 정보	#	CW	문맥 정보	#	CW
synset	1	1.00	query language	1	1.00	relation	1	1.00
lexical database	1	0.61	substructure	1	0.42	WordNet	2	0.96
information processing system	1	0.55	relational database management system	1	0.40	concept	1	0.62
conceptualisation	1	0.45	computer language	1	0.36	knowledge	1	0.60
polysemous word	1	0.32	DBMS	1	0.27	knowledge base	1	0.47

<표 23>과 같이 각 문맥 정보에서 'Doc 1'과 'Knowledge Representation'에 속한 개념들의 관계들은 [그림 24]와 같이 형성되고, 'Doc 1'과 'Query Language' 사이에는 [그림 25]와 같이 추출된다. 이러한 관계 구조는 확장된 워드넷을 이용하여 추출한 것이다.



[그림 24] 확장된 워드넷 기반 문맥 정보들의 관계 구조
('Knowledge Representation'과 'Doc 1'의 문맥정보)
*서로 다른 문맥 정보에 속한 개념들 사이의 가장 짧은 거리 관계를 위주로 표기한 것임.
실제는 더 많은 관계들이 형성됨.



[그림 25] 확장된 워드넷 기반 문맥 정보들의 관계 구조
('Query Language'와 'Doc 1'의 문맥정보)
*서로 다른 문맥 정보에 속한 개념들 사이의 가장 짧은 거리 관계를 위주로 표기한 것임.
실제는 더 많은 관계들이 형성됨.

[그림 24]와 [그림 25]와 같이 형성된 각 문맥 정보들 사이의 관계를 (식 14)를 이용하여 계산한다. 다음은 'Doc 1'을 중심으로 각 문맥 정보의 개념들과 관계성 계산과 그 관계성을 모두 합하여 최종적으로 문맥 정보들 사이의 유사도를 측정하는 과정을 보인다. 먼저 'Doc 1'과 'Knowledge Representation' 사이의 유사도 측정과정이다.

CI_relatedness(relation#1, synset#1) = (1.00+1.00)/(2*5) = 0.2
CI_relatedness(relation#1, lexical database#1) = 관계 없음.
CI_relatedness(relation#1, information processing system#1) = 관계 없음.
CI_relatedness(relation#1, conceptualisation#1) = (1.00+0.45)/(2*4) = 0.18125
CI_relatedness(relation#1, polysemous word#1) = (1.00+0.32)/(2*5) = 0.132

CI_relatedness(WordNet#2, synset#1) = (0.96+1.00)/(2*2) = 0.49
CI_relatedness(WordNet#2, lexical database#1) = (0.96+0.61)/(2*2) = 0.3925
CI_relatedness(WordNet#2, information processing system#1) = (0.96+0.55)/(2*5) = 0.151
CI_relatedness(WordNet#2, conceptualisation#1) = 관계 없음.
CI_relatedness(WordNet#2, polysemous word#1) = (0.96+0.32)/(2*5) = 0.128

CI_relatedness(concept#1, synset#1) = (0.62+1.00)/(2*5) = 0.162
CI_relatedness(concept#1, lexical database#1) = 관계 없음.
CI_relatedness(concept#1, information processing system#1) = (0.62+0.55)/(2*4) = 0.14625
CI_relatedness(concept#1, conceptualisation#1) = (0.62+0.45)/(2*2) = 0.2675
CI_relatedness(concept#1, polysemous word#1) = 관계 없음.

CI_relatedness(knowledge#1, synset#1) = 관계 없음.
CI_relatedness(knowledge#1, lexical database#1) = 관계 없음.
CI_relatedness(knowledge#1, information processing system#1) = (0.60+0.55)/(2*5) = 0.115
CI_relatedness(knowledge#1, conceptualisation#1) = (0.60+0.45)/(2*5) = 0.105
CI_relatedness(knowledge#1, polysemous word#1) = (0.60+0.32)/(2*5) = 0.092

CI_relatedness(knowledge base#1, synset#1) = 관계 없음.

CI_relatedness(knowledge base#1, lexical database#1) = 관계 없음.

CI_relatedness(knowledge base#1, information processing system#1) = (0.47+0.61)/(2*5) = 0.102

CI_relatedness(knowledge base#1, conceptualisation#1) = (0.47+0.45)/(2*5) = 0.092

CI_relatedness(knowledge base#1, polysemous word#1) = 관계 없음.

sim('Doc 1', 'Knowledge Representation') = 2.7565

다음은 'Doc 1'과 'Query Language' 사이의 유사도 측정과정이다. 아래의 과정에서는 실제로 관계가 있는 사례만을 작성한다.

CI_relatedness(relation#1, substructure#1) = (1.00+0.42)/(2*5) = 0.142

CI_relatedness(WordNet#2, relational database management system#1) = (0.96+0.40)/(2*5) = 0.136

CI_relatedness(WordNet#2, DBMS#1) = (0.96+0.27)/(2*4) = 0.15375

CI_relatedness(concept#1, substructure#1) = (0.62+0.42)/(2*5) = 0.104

CI_relatedness(knowledge#1, substructure#1) = (0.60+0.42)/(2*5) = 0.102

sim('Doc 1', 'Query Language') = 0.63775

(식 14)에 의해 <표 23>에 기술된 각 문맥 정보 사이의 유사도는 위와 같이 측정되며, 유사도 측정 결과에 따라 'Doc 1'의 내용은 'Query Language' 보다는 'Knowledge Representation' 카테고리로 분류된다.

또한 각 위키 카테고리는 다양한 위키 개념을 포함하고 있다. 본 연구의 최종 목적은 타겟 문서가 기술하는 의미적 주제들을 파악하여 그 주제들을 대표할 수 있는 위키 개념들로 태깅하는 것이다. 이에, 위와 같이 파악된 카테고리에 속한 위키 개념의 문맥 정보들과 유사도 측정 과정을 거친다. 본 과정 또한 (식 14)와 (식 15)를 이용하여 진행되며, 3장에서

준비한 각 위키 개념의 문맥 정보를 이용한다. 본 예제를 위해 위키 개념 'Knowledge Representation and reasoning', 'WordNet', 'Semantic Network', 'RDF schema'를 예로 이용한다. 이들은 [그림 23]에서 위키 카테고리 'Knowledge Representation'에 속한 위키 개념들이다. <표 24>는 각 위키 개념의 문맥정보 중에서 상위 5개를 이용한 태깅 유사도 측정 과정을 보이고 있다.

<표 24> 의미적 태깅에 대한 예시

Knowledge Representation and Reasoning			Semantic Network			WordNet		
문맥 정보	#	CW	문맥 정보	#	CW	문맥 정보	#	CW
inference	1	1.00	semantic relation	1	0.71	WordNet	2	1
reasoning	1	0.94	concept	1	0.52	synset	1	0.28
knowledge	1	0.73	graph	1	0.41	lexical database	1	0.22
logic	1	0.64	representation	1	0.34	database	1	0.19
representation	1	0.55	network	1	0.33	synonym	1	0.18

타겟 문서	위키 개념	유사도	
Doc 1	Knowledge Representation and Reasoning	relation#1 = 0.81525	WordNet#2 = 0
		knowledge#1 = 1.71437	concept#1 = 0.798
		knowledge base#1 = 0.63375	**3.96137**
	Semantic Network	relation#1 = 0.5842	WordNet#2 = 0
		knowledge#1 = 0.7845	concept#1 = 0.4516
		knowledge base#1 = 0.24458	**2.06483**
	WordNet	relation#1 = 0.065	WordNet#2 = 1.1667
		knowledge#1 = 0.025	concept#1 = 0.037
		knowledge base#1 = 0	**1.296667**

* 'Semantic Network'에서 가장 큰 CW가 1이 아닌 이유는 고유명사가 최대값을 가졌기 때문임.

<표 24>의 결과에 따라, 'Doc 1'의 내용은 'Knowledge Representation and

Reasoning'으로 태깅될 수 있다. 위의 과정은 본 연구의 이해를 위해 단순히 상위에 속한 5개의 문맥을 이용하고 있지만, 실제로 문맥의 수가 다양하기 때문에 그 결과는 달라질 수 있다. 또한 하나의 문서는 여러 개의 위키 개념으로 태깅될 수 있다. 예를 들어, 컴퓨터를 적용한 의료 방법(예. 영상처리 기법과 의료 영상)이라는 내용을 기술한다면 그것은 컴퓨터와 관련된 위키 개념, 의료와 관련된 위키 개념으로 태깅될 수 있으며, 그 카테고리 또한 복수개로 형성될 수 있음을 의미한다. 이러한 부분은 실험 부분에서 상세하게 기술한다.

V. 실험 및 결과 평가

본 장에서는 앞에서 제안한 타겟 문서의 위키 카테고리 분류 및 위키 개념 태깅 방법에 대한 평가를 수행한다. 먼저 의미적 분류 및 태깅을 위해 각 위키 개념, 위키 카테고리, 타겟 문서의 문맥 정보에 대해 문맥 가중치를 측정하였다. 또한 문맥 정보들 사이의 유사도를 측정함으로써 타겟 문서를 특정 위키 카테고리로 분류 및 위키 개념으로 태깅하였다. 이에, 본 장에서는 문맥 정보 추출에 대한 정확도 평가, 타겟 문서 분류 정확도 평가, 그리고 타겟 문서의 태깅 정확도 평가를 수행하고, 기존에 수행된 유사 연구들과 비교함으로써 본 연구에 대해 객관적 분석을 하도록 한다.

A. 문맥 정보 추출 정확도 평가

본 연구의 실험 및 평가를 위해 위키피디아 초록에 단어 'computer'를 포함하는 문서들을 선정하였다. DBPedia 3.4 버전에 포함된 위키피디아 초록 집합(총 2,944,417개 문서)에서 'computer'를 포함하는 문서의 수는 총 36,566개이며, 이를 모두 분석하여 각 위키 개념과 카테고리에 대한 문맥 정보 추출에 활용하였다. 위키피디아의 초록과 카테고리 및 타겟 문서에서 문맥 정보를 형성하는 것은 본 연구의 핵심이라 할 수 있으며, 이는 타겟 문서 분류 및 태깅 결과에 큰 영향을 미친다. 또한 타겟 문서 분류 및 태깅을 위해 웹에서 논문 1,000건을 수집하였다. 이에, 본 단원에서는 본 연구에 의해 형성된 각 문맥 정보의 추출 정확도에 대해 평가하고, 그 결과를 TF-iDF 방식, 의미적 주제 선정 방식(Semantic Topic Selection, STS) 방식[43], 그리고 본 연구의 선행 연구인 의미적 문맥 추출(Semantic Context Extraction, SCE) 방식[87]과 비교 평가를 수행한다. TF-iDF 방식은 키워드 기반 검색의 대표적인 연구이며, 본 연구가 키워드 기반 검색보다 어느 정도 성능향상을 가져올 수 있는지 비교하기 위함이다. 또한 STS와 SCE 방식은 도메인 온톨로지(Domain Ontology) 또는 워드넷을 지식베이스로 활용한 연구로써 각 문서의 핵심 키워드(문맥 정보)를 추출함에 있어서 의미성을 부여한다는 측면에서 유사한 연구라 할 수 있기 때문에 그들과의 비교는 타당하다.

1. 위키 개념에 따른 문맥 정보 추출 정확도

단어 'computer'를 포함하는 문서 36,566개 초록에서 추출된 총 단어의 수는 1,448,664개로 하나의 위키 개념에 대해 약 39.6개인 것을 확인할 수 있었다. 여기서 300개의 위키 개념(초록)에서 추출된 문맥 정보에 대한 평가를 시도하였으며, 이에 속한 단어의 수는 9,416(고유명사 7,082, 일반명사 2,334)개이다. 평가 방법으로는 2명의 평가자를 두어 위키 개념-문맥 정보 쌍 만(앞의 과정에서 계산된 모든 수치를 제외하였음)을 나누어 주고 관련성이 있다는 것에 1, 그렇지 않은 것에 0을 부여 하도록 하였다. 그리고 최종적으로 2명이 모두 1을 부여한 것만 관련성이 있는 것으로 간주하였다.

먼저 평가자가 판단한 문맥 정보의 정확도는 62.34%(=5,870÷9,416)에 도달하는 것을 확인하였다. 이 수치는 각 위키 초록에 포함된 모든 단어에 대해 관련성을 평가한 결과이기 때문에 TF-iDF, STS, SCE 방식에 의한 결과가 모두 일치하는 기본 선(Base Line)이라 할 수 있다. 이에 각 가중치에 의해 상위에 포함된 문맥 정보의 관련성을 판단하기 위해 각 위키 개념의 문맥 정보에서 상위 10, 20, 30, 40, 50, 70, 100(%)에 속한 단어들의 관련성에 대해 평가를 시도하였다. 각 방법에서 고유 명사와 일반 명사에 대해 확연히 다른 수치를 계산한다. 특히, 지식베이스를 활용한 STS, SCE, 그리고 본 방법은 고유명사에 의미적 가중치를 반영하지 못하기 때문에 일반명사와 그 값의 차이가 현저히 크다. 이에, 고유명사와 일반명사를 서로 분리하여 상위에 속한 비율에 따른 평가를 시도하였다. <표 25>는 추출된 위키 문맥 정보에서 위키 개념과 관련된 정도를 일반 명사(워드넷에 정의된 명사)에 대한 부분과 고유 명사(워드넷에 정의되지 않은 명사)에 대한 것을 구분하여 보이고 있다.

<표 25> 추출된 위키 문맥 정보의 관련 정확도

일반 명사					고유 명사			
비율 (%)	관련 정확도 (%)				비율 (%)	관련 정확도 (%)		
	TF-iDF	STS	SCE	TM		TF-iDF	SCE	TM
10 (708)	70.04 (496)	66.79 (473)	70.88 (502)	83.17 (589)	10 (233)	84.83 (198)	74.98 (175)	85.26 (199)
20 (1,416)	72.93 (1,033)	69.12 (979)	73.64 (1,043)	85.36 (1,209)	20 (467)	85.48 (399)	78.19 (365)	85.69 (400)
30 (2,125)	72.95 (1,550)	70.22 (1,492)	78.89 (1,676)	87.69 (1,863)	30 (700)	86.12 (603)	78.26 (548)	86.12 (603)
40 (2,833)	71.03 (2,012)	69.12 (1,958)	78.76 (2,231)	87.58 (2,481)	40 (934)	86.98 (812)	81.83 (764)	86.44 (807)
50 (3,541)	69.98 (2,478)	65.22 (2,451)	75.23 (2,664)	83.25 (2,948)	50 (1,167)	83.80 (978)	79.79 (931)	84.06 (981)
70 (4,957)	65.70 (3,257)	65.72 (3,258)	70.50 (3,495)	72.20 (3,579)	70 (1,634)	79.51 (1,299)	74.31 (1,214)	78.90 (1,289)
100 (7,082)	59.45 (4,210)	59.45 (4,210)	59.45 (4,210)	59.45 (4,210)	100 (2,334)	71.12 (1,660)	71.12 (1,660)	71.12 (1,660)

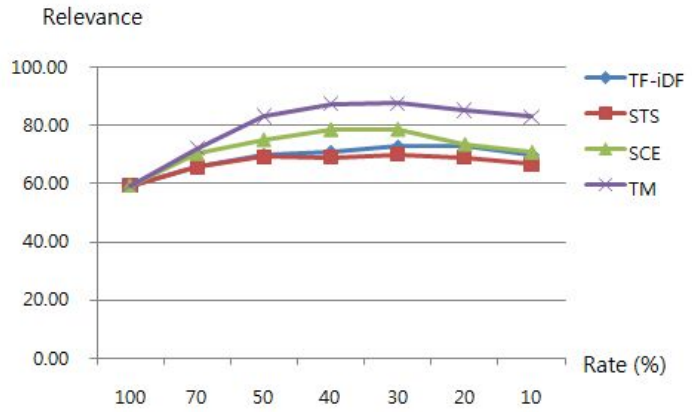
*TM: 본 연구의 방법

<표 25>에서 STS 방식은 고유명사 부분을 고려하지 않기 때문에 일반 명사 부분에 대한 평가만을 보였다. 추출된 문맥 정보를 일반과 고유 명사로 분류한 평가에서 각각 59.45(%)와 71.12(%)로 판단되었으며, 고유 명사에 대한 관련 정확도가 높은 것으로 확인되었다. 이는 고유 명사의 중의성(Ambiguity)이 일반 명사의 것보다 적고, 위키 개념을 설명하는 고유 명사들이 직접적으로 관련된 것들로 구성되기 때문인 것으로 확인되었다.

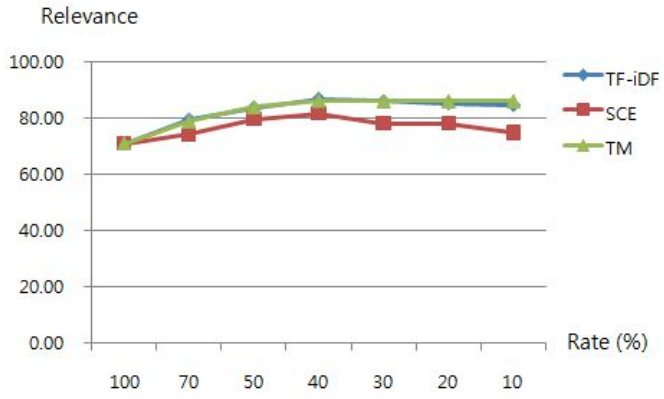
또한 각 방법에 의한 문맥 가중치에 따라 상위에 포함된 문맥 정보들의 관련 정확도를 계산하면, 일반 명사 부분에서 본 연구의 방법이 확연히 높은 것을 확인할 수 있다. 특히 최고 정점은 대부분 상위 30% 부분인 것으로 확인되었다. 일반 명사의 상위 30%에서 관련 정확도의 차이는 본 연구의 방법이 TF-iDF, STS, SCE 보다 각각 14.73, 17.46, 그리고 8.8 포인트가 높았으며, TF와 워드넷에서 직접적인 관계만을 고려하는 STS 방법이 가장 낮은 것으로 확인되었다. 본 연구의 선행연구였던 SCE 방법은 TF-iDF 방법보다 높았으나, 위키 초록 집합 전체에서 출현하는 단어의 빈도를 추가로 고려한 본 연구가 8.8 포인트를 더 높게 만들 수 있었다.

고유 명사에 대한 평가 부분에서는 본 방법과 TF-iDF 방법이 아주 유사한 결과를 보이고 있으며, 이는 본 연구에서 워드넷에 정의되지 않은 단어들에 대해서는 TF-iDF의 장점을 최대한 반영하고 있기 때문이다. 이 부분에서 역시 SCE 방법은 위키 초록 집합에서 출현하는 단어의 빈도를 고려하지 않기 때문에, 즉 TF만을 이용하기 때문에 가장 낮은 정확도를 기록하였다.

[그림 26]과 [그림 27]은 <표 25>의 내용을 그래프로 요약하고 있다.



[그림 26] 추출된 위키 문맥 정보에서 일반 명사의 개념 관련성
(TM: 본 연구의 방법)



[그림 27] 추출된 위키 문맥 정보에서 고유 명사의 개념 관련성

본 연구의 결과에 따라 불필요한 노이즈를 최소화하고 관련성이 높은 단어들을 최대화하기 위해 일반 명사와 고유 명사의 상위 30%만을 각 위키 개념의 문맥 정보로 활용하며, 이를 타겟 문서의 태깅에 활용한다. 이는 각 위키 개념에 대해 평균 11.89개의 단어로 구성된 문맥 정보이며, 일반 명사와 고유 명사로 분류하였을 때 10.87개와 1.02개로 각각 파악되었다. <표 26>은 본 연구에 의해 추출된 위키 개념 'Microsoft'에 대한 문맥 정보의 상위 30%를 보이고 있다.

<표 26> 위키 개념 'Microsoft'에 대한 문맥 정보 중에서 상위 30%

구분	순위	문맥 정보	KW	CW	판단
일반 명사	1	computer	0.0914	1.0000	1
	2	product	0.0797	0.7450	1
	3	company	0.0785	0.6225	1
	4	computing device	0.0492	0.5380	1
	5	software	0.0470	0.4396	1
	6	system	0.0506	0.4115	1
	7	market	0.0434	0.3759	1
	8	stock	0.0543	0.3741	1
	9	BASIC	0.0326	0.2943	1
	10	software product	0.0403	0.2542	1
	11	hardware	0.0297	0.2368	1
	12	desktop	0.0353	0.2363	1
	13	interpreter	0.0348	0.2291	0
	14	employee	0.0262	0.2266	1
	15	Windows	0.0311	0.2191	1
	16	mouse	0.0329	0.2169	1
	17	strategy	0.0268	0.2122	0
	18	technology	0.0208	0.2077	1
고유 명사	1	Microsoft	0.2325	0.2178	1
	2	XBox	0.0678	0.0635	1
	3	MSN TV	0.0621	0.0582	1
	4	MSN Internet	0.0621	0.0582	1

2. 위키 카테고리에 따른 문맥 정보 추출 정확도

본 연구에 활용된 36,566개의 위키 개념 문서에 기술된 카테고리는 중복을 허용하여 총 111,228개에 달하며, 이는 위키 개념 당 평균 3.06개이다. 중복된 것을 제거한 개별 위키 카테고리는 26,123개이며, 최소 1개부터 최대 4,371개의 위키 개념의 카테고리로 기술되고 있다. 본 연구에서는 원활한 문서 분류를 위해 카테고리로써 자질이 부족한 것을 제거하는 필터링 과정이 필요하였다. 그 기준은 다음과 같다.

- 상황에 따라 카테고리가 바뀔 수 있는 개념: 예) Living People, Articles lacking sources, year of birth missing living people 등
- 위키 개념에 카테고리로 기술된 횟수가 5회 미만인 것 (이는 'computer'를 포함하고 있긴 하지만 해당 도메인('computer')의 카테고리로 판단하기 어려움): 예) Tianhe District, Building and structures in Trondheim, People from Walworth County, People from St Helens 등

위의 기준에 의해 3,908개의 카테고리를 형성할 수 있었으며, 각 카테고리에 포함된 위키 개념의 문맥 정보를 통합하였다. 통합된 카테고리 문맥 정보를 V-A-1 과정에서 보인 방법과 동일한 방법으로 비교 평가하였다. 평가를 위해 30개 카테고리를 임의로 선정하였으며, 30개의 카테고리에 포함된 문맥 정보의 수는 11,483개였다. <표 27>은 위키피디아 카테고리에 속한 문맥 정보 추출에 대한 평가 결과를 보이고 있다. 이 역시 고유 명사 부분과 일반 명사 부분을 분류하여 평가하였으며, 위키 개념 문맥 정보 평가에서 비교 대상이었던 TF-iDF, STS, SCE를 이용하였다. 본 과정에서 분류된 일반 명사는 7,911개, 고유 명사는 4,075개 였다. 또한 위키 카테고리 전체에 대해 계산하면 1,197,086개 문맥 정보 단어 중에 일반 명사 814,019개, 고유 명사 383,067개로 고유 명사가 약 32%를 차지하는 것을 파악할 수 있었다. 이는 위키 문맥 정보에서 약 10%를 차지했던 것보다 3배에 이르는 수준인데, 하나의 카테고리 문맥 정보로 통합될 때, 각 위키 문서에 포함된 고유 명사들은 서로 다른 반면, 일반 명사들은 동일한 것이 많이 형성되었기 때문이다.

<표 27> 카테고리 문맥 정보의 관련 정확도

일반 명사					고유 명사			
비율 (%)	관련 정확도 (%)				비율 (%)	관련 정확도 (%)		
	TF-iDF	STS	SCE	TM		TF-iDF	SCE	TM
10 (791)	82.29 (651)	74.45 (589)	78.50 (621)	91.01 (720)	10 (408)	85.40 (348)	76.56 (312)	85.40 (348)
20 (1,582)	83.24 (1,317)	75.34 (1,192)	78.44 (1,241)	91.45 (1,447)	20 (815)	86.13 (702)	78.04 (636)	86.13 (702)
30 (2,373)	84.44 (2,004)	76.81 (1,823)	81.07 (1,924)	91.27 (2,166)	30 (1,223)	90.06 (1,101)	81.72 (999)	90.22 (1,103)
40 (3,164)	83.90 (2,655)	75.09 (2,376)	78.88 (2,496)	88.61 (2,804)	40 (1,630)	86.87 (1,416)	80.80 (1,317)	87.06 (1,419)
50 (3,956)	80.29 (3,176)	73.29 (2,899)	77.94 (3,083)	84.54 (3,344)	50 (2,038)	83.29 (1,697)	77.35 (1,576)	83.48 (1,701)
70 (5,538)	74.56 (4,129)	71.17 (3,941)	73.62 (4,077)	77.14 (4,272)	70 (2,853)	79.02 (2,254)	77.41 (2,208)	78.84 (2,249)
100 (7,911)	64.44 (5,098)	64.44 (5,098)	64.44 (5,098)	64.44 (5,098)	100 (4,075)	73.52 (2,996)	73.52 (2,996)	73.52 (2,996)

<표 27>의 결과에서 카테고리 문맥 정보의 관련 정확도가 전반적으로 위키 문맥 정보보다 높은 것을 확인할 수 있는데, 이는 문맥 정보가 많으면 많을수록 주요 단어의 확률적 가중치와 의미적 가중치가 동시에 증가하기 때문인 것으로 판단되었다. 또한 고유 명사 관련 정확도 부분에서 본 연구의 방식은 여전히 TF-iDF와 거의 흡사하다. 하지만, 일반 명사 부분에서는 확연한 차이를 보이는데, 이는 주요 단어에 대한 기본적인 추출은 확률에 의존하는 반면 결정적인 영향은 의미적 가중치가 미치기 때문인 것으로 해석될 수 있다.

카테고리 문맥 정보에서도 상위 30%(카테고리 평균 약 91.90개, 일반 명사 62.49 및 고유 명사 29.40개)를 이용하여 타겟 문서 분류를 위한 근거 데이터로 활용한다. <표 28>은 위키 카테고리 'UNIX'에 포함된 문맥 정보 중에서 본 연구의 방법에 의한 문맥 정보 가중치를 기준으로 높은 것부터 정렬한 일부를 보이고 있다.

<표 28> 위키 카테고리 'UNIX'에 대한 문맥 정보의 일부

구분	순위	문맥 정보	순위	문맥 정보
일반 명사	1	data communication	11	system call
	2	library routine	12	Tarantella
	3	MIPS	13	data structure
	4	descriptor	14	initialization
	5	UNIX System	15	video display
	6	Unix	16	device driver
	7	peripheral device	17	computing system
	8	RISC	18	directory
	9	system administrator	19	file system
	10	typewriter	20	kernel
고유 명사	1	MIPS OS	4	USENIX
	2	VUE	5	MIPS Magnum
	3	mtXinu	6	MIPS Computer Systems

3. 타겟 문서의 문맥 정보 추출 정확도

본 연구에서 제안하는 타겟 문서(웹 문서) 태깅 방법의 실험을 위해 웹(IEEE Xplore, J.UCS 사이트, Google 등)에서 'computer'와 관련된 논문을 수집하였으며, 2005년부터 현재까지의 논문 1,000건을 활용하였다. 먼저 이들 문서에서 추출되는 문맥 정보의 정확도에 대한 실험을 진행하였으며, 이를 위해 각 논문에서 초록, 서론, 그리고 결론 부분만을 추출하였다. 이는 본문 및 실험의 내용에 다양한 수식, 표, 그림 등이 포함되어 아주 많은 노이즈를 포함하기 때문이다. 타겟 문서의 위키 카테고리로의 분류와 위키 개념 태깅에는 수집된 타겟 문서 집합을 모두 이용하였지만, 문맥 추출 정확도를 위해서는 50개의 문서를 이용하였다. 50개의 문서에 포함된 문맥 정보의 단어 수는 4,310개이며, 앞의 방법과 동일한 방법을 통해 관련 정확도를 계산하여 TF-iDF 및 SCE 방법과 비교하였다. <표 29>는 그 결과를 보이고 있다. 50개의 타겟 문서 분포하는 일반 명사와 고유 명사는 각각 3,750개와 560개(약 12.99%)로 구성되었다.

<표 29> 타겟 문서에서 추출된 문맥 정보 관련 정확도

일반 명사				고유 명사			
비율 (%)	관련 정확도 (%)			비율 (%)	관련 정확도 (%)		
	TF-iDF	SCE	TM		TF-iDF	SCE	TM
10 (375)	73.87 (277)	70.67 (265)	79.47 (298)	10 (56)	75.00 (42)	67.86 (38)	75.00 (42)
20 (750)	78.80 (591)	77.73 (583)	83.33 (625)	20 (112)	78.57 (88)	70.54 (79)	78.57 (88)
30 (1,125)	77.51 (872)	77.24 (869)	83.20 (936)	30 (168)	79.17 (133)	70.24 (118)	78.57 (132)
40 (1,500)	74.20 (1,113)	76.07 (1,141)	79.87 (1,198)	40 (224)	75.45 (169)	69.64 (156)	75.45 (169)
50 (1,875)	68.21 (1,279)	69.87 (1,310)	70.77 (1,327)	50 (280)	71.79 (201)	67.50 (189)	72.14 (202)
70 (2,625)	61.45 (1,613)	64.57 (1,695)	65.79 (1,727)	70 (392)	65.56 (257)	59.18 (232)	65.56 (257)
100 (3,750)	47.65 (1,787)	47.65 (1,787)	47.65 (1,787)	100 (560)	55.71 (312)	55.71 (312)	55.71 (312)

<표 29>에서 관련 정확도가 위키 개념 및 카테고리의 것보다 낮음을 볼 수 있다. 본 연구에 이용된 위키 개념의 문맥 정보는 위키 문서의 초록 부분에서 추출한 것인데, 초록은 위키 개념을 설명하기 위해 주요한 단어들을 많이 포함하기 때문이다. 반면, 타겟 문서로 활용된 논문은 그 주제와 관련되지 않은 다양한 단어들(예. paper, study, feature, result, experiment 등)을 포함하여 노이즈로 작용함을 확인할 수 있었다. 하지만, 여기에서도 다른 문맥 정보들과 마찬가지로 상위 30%일때 대체적으로 정확도가 정점에 도달하기 때문에, 노이즈를 다수 포함하지만 타겟 문서의 문맥정보로 활용한다. <표 30>은 본 과정에서 추출된 타겟 문서 중 온톨로지 매핑에 대한 논문[88]에서 추출한 상위 30% 문맥 정보를 보이고 있다.

<표 30> 타겟 문서 중
온톨로지 매핑에 대한 논문[88]에서 추출한 상위 30%에 속한 문맥 정보

구분	순위	문맥 정보	순위	문맥 정보
일반 명사	1	mapping	19	study
	2	information system	20	problem
	3	knowledge	21	amount
	4	system	22	result
	5	ontology	23	sharing
	6	information	24	query
	7	environment	25	similarity
	8	interoperability	26	process
	9	composition	27	database
	10	network	28	paper
	11	domain	29	discovery
	12	user	30	work
	13	resource	31	number
	14	measurement	32	centrality
	15	platform	33	schema
	16	transformation	34	telecommunication
	17	concept	35	algorithm
	18	peer	고유명사	P2P

4. 비교평가 결과

TF-iDF, STS, SCE, 그리고 본 연구의 방법을 이용하여 문맥 정보를 추출한 결과는 본 연구에 의한 방법이 가장 높은 것으로 확인되었다. 이에 대한 결과를 분석하면, TF-iDF 방법에서는 전체 문서 집합을 이용하여 필터링을 하기 때문에 노이즈 정보의 가중치를 낮게 만드는 장점이 있지만, 단지 확률적인 정보만을 이용하기 때문에 정작 주요한 개념에 대해서는 가중치를 부여하는데 한계점이 있음이 확인되었다. 예를 들어, 'Caxton College'(스페인 발렌시아의 사립학교)에 대한 위키 개념에서 'school'과 'Valencia' 모두 관련된 어휘로 추출될 수 있지만 순위를 결정하는 측면에서는 'school'이 더욱 관련 깊다고 할 수 있다. 하지만 TF-iDF에서는 'Valencia'를 오히려 더 높게 판단하는 한계점이 존재하였으며, 본 연구의 방법에서는 이를 개선할 수 있었다. STS 방법은 본 비교 평가에서

가장 저조한 성능을 보였는데, 대표적인 원인으로는 필터링(Filtering, iDF)의 부재와 의미성 부족인 것으로 확인되었다. 단순히 해당 문서에서 다수 출현하는 어휘에 대해 확률 가중치를 높게 부여한 것이 첫 번째 한계점이다. 또한 지식베이스로 활용되는 워드넷에서 직접적으로 형성되는 관계(본 연구를 기준으로 거리 2까지만 STS에 활용)만을 이용하기 때문에 의미성 부여에 한계점이 있음이 추가로 확인되었다. 실제 전문가 또는 일반인들이 문서를 작성할 때, 문서 주제 개념과 직접적으로 관련된 어휘보다는 간접적으로 관련된 어휘를 많이 사용하는 것이 일반적이기 때문에 STS 방법으로는 그 관계성을 측정할 수 없는 것이 대부분이었다. 예를 들면, 'Caxton College'를 기술하기 위해 'director', 'classroom', 'computer' 등을 활용하며, 이들은 'college'와 직접적인 관계가 워드넷 내에 형성되어 있지 않다. 본 연구의 선행 연구였던 SCE 방법은 본 연구와 유사하게 풍부한 관계성을 활용하지만 필터링을 반영하지 않기 때문에 본 연구보다 저조한 성능을 보였다. 이와 같이, 문맥 정보를 추출함에 있어서 필터링을 겸한 확률적 가중치와 풍부한 관계성을 반영한 의미적 가중치는 그 정확도를 높일 수 있음이 확인되었다.

B. 위키 카테고리로 분류 및 위키 개념 태깅 정확도 평가

타겟 문서의 분류를 위해 'computer'와 관련된 위키 개념 36,566개, 위키 카테고리 3,908개, 그리고 논문 1,000개를 분석하였다. 그리고 실제 문서 분류와 태깅을 위한 근거 데이터로 사용될 각 위키 개념, 각 카테고리, 그리고 각 논문에 대한 문맥 정보를 추출하였다. 추출된 문맥 정보에서 일반 명사와 고유 명사를 분류하여 상위 30%에 속한 단어들을 선정하였으며, 본 단원에서는 그 문맥 정보들을 이용하여 타겟 문서를 위키 카테고리로 분류 및 위키 개념을 이용한 태깅에 대해 정확도를 평가한다.

본 연구에 의한 결과를 국내외 저명한 저널에 이미 소개된 문서 유사도를 측정하는 방식들인 CS(Cosine Similarity, Vector Space Model에서 질의어와 문서 사이의 유사도 측정에 자주 사용되는 방식), DCM(Document Clustering Method)[47], SDI(Semantic Document Interconnections)[2]와 비교함으로써 객관적인 평가를 한다. CS 방식은 키워드 기반의 대표적 문서 유사도 측정 방식이며, 이와 비교를 통하여 본 연구가 키워드 방식보다 어느정도 성능개선이 있는지 비교하기 위함이다. 또한 DCM과 SDI는 유사도를 측정하는 방법은 다르지만 워드넷을 지식베이스로 활용하는 것으로써 본 연구와 아주 유사하며, 동일한 지식베이스를 활용하였을 때의 성능 개선 효과를 보기 위한 것이다.

평가를 시도하기 위해 각 방식에 따라 문맥 정보를 구성하였다. 또한, 타겟 문서로 사용된 각 논문의 분류 및 태깅 정확도를 계산하기 위해 Gold Standard를 형성하였다. 그 방법으로는 2명의 평가자를 두어, 각 평가자에게 타겟 문서 1,000개에 대해 주제가 될 수 있는 위키 개념 3개를 위키피디아 웹 자체에서 선정하도록 요구하였으며, 두 평가자가 모두 일치하는 위키 개념만을 각 타겟 문서의 태거로 간주하였다. 두 평가자의 일치도는 <표 31>과 같다.

<표 31> 평가자에 의한 위키 개념 태깅 일치도

	모두 일치	2개 일치	1개 일치	불일치
비율 (%)	9.8	27.4	41.7	21.1
문서 수	98	274	417	211

<표 31>의 결과에서 두명의 평가자가 선정한 위키 개념 중에서 일치하는 것을 본 연구의 태깅 정확도 평가를 위한 Gold Standard로 이용한다. 그리고 타겟 문서의 주제로 선정된 위키 개념의 카테고리들을 이용하여 분류 정확도 평가에 이용하였다. 평가자들의 판단이 불일치 한 타겟 문서들을 제외한 총 789개를 이용하였으며, 각 방법이 타겟 문서의 위키 카테고리를 얼마나 잘 찾아내는지에 대한 실험을 먼저 진행하였다. 타겟 문서는 최소 2개(할당된 위키 개념이 1개인 문서)부터 최대 13개(할당된 위키 개념이 3개인 문서)까지의 위키 카테고리를 갖는 것으로 확인되었다. <표 32>는 그 통계를 보이고 있다. 5개 이상부터는 원활한 평가를 위해 하나의 그룹으로 형성하였다.

<표 32> 위키 카테고리 수에 따른 타겟 문서의 개수

	5개 이상	4개	3개	2개
문서 수	173	288	197	131

타겟 문서에 할당된 위키 개념이 많을수록 카테고리 형성이 많을 것으로 예상하였으나, 하나의 카테고리 내에 존재하는 두 개 이상의 위키 개념이 평가자에 의해 태깅된 경우가 다수 존재하였다. 위키 카테고리 분류 정확도 측정을 위해 타겟 문서들을 본 연구에서 제안하는 방법을 따라 유사도가 높은 순으로 5개의 카테고리를 선정하고, 평가자에 의해 선정된 카테고리의 일치 여부를 계산하였다. 그리고 비교 평가를 위해 기존에 수행된 연구들을 적용하여 동일하게 5개의 카테고리를 선정하였다. <표 33>은 평가 결과를 보이고 있다. 카테고리 수는 본 연구에 의해 선정된 것을 의미하며, 5일 경우에는 5개 이상의 카테고리를 갖는 타겟 문서, 4, 3, 2는 상위에 선정된 4, 3, 2개의 위키 카테고리를 의미하며, 각각 4, 3, 2개의 카테고리를 갖는 타겟 문서에 대해 정확도를 계산한 것이다. 만약 타겟 문서가 4개의 카테고리를 갖고 있고 본 연구에 의해 4개의 카테고리가 추출된 것 중에 3개가 일치한다면 0.75 값을 정확도에 반영한다. 또한 <표 33>에서 카테고리 수가 1인 경우는 유사도를 가장 높게 선정한 하나의 카테고리가 타겟문서의 카테고리에 존재한다면 정확한 것으로 판단하였다.

<표 33> 위키 문서 분류 정확도

방법	카테고리 수				
	5	4	3	2	1
CS	약 71.7%	약 67.2%	약 56.3%	약 51.5%	약 84.3%
DCM	약 72.4%	약 70.7%	약 62.6%	약 55.0%	약 85.1%
SDI	약 85.9%	약 73.5%	약 67.2%	약 63.7%	약 90.8%
제안된 방법	약 97.6%	약 87.4%	약 82.1%	약 79.4%	약 97.8%

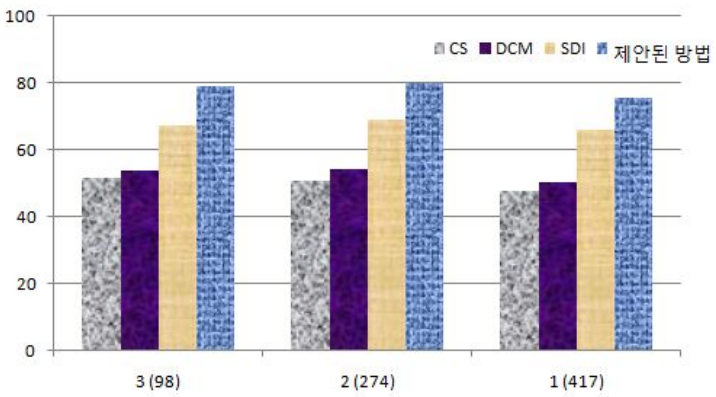
<표 33>과 같이 본 연구에 의한 방법이 기존의 연구보다 카테고리를 선정함에 있어 우수한 결과를 보였다. CS와 DCM의 방법은 일치하는 단어만을 이용하여 유사성을 계산하기 때문에 카테고리를 분류하는 정확도가 낮으며, 본 연구의 기반이 된 SDI 방식은 문맥 정보로 활용하는 단어의 수가 적고, 고유 명사의 경우 거의 반영이 되지 않았기 때문에 CS와 DCM 방법보다는 우월하지만 본 연구보다는 낮은 정확도를 보였다. 반면 풍부한 문맥 정보를 활용하고, 일반 명사와 고유 명사를 구분하여 유사도 측정에 적용하는 본 방법은 가장 우수한 결과를 보일 수 있었다.

타겟 문서의 위키 개념 태깅에 대한 정확도 평가를 위해 유사도가 높은 순서로 3, 2, 1개의 위키 개념을 분류하여 선정하였다. 분류 기준은 평가자에 의해 위키 개념을 3개 갖게 된 98의 타겟문서에 대해서는 본 연구에서도 3개의 개념을 추출하였고, 2개와 1개의 경우도 그 수에 맞춰 각각 추출하였다. 그리고 그 일치 여부를 이용하여 정확도를 산정하였다. 이 부분에서도 카테고리 선정 부분과 같이, 3개의 위키 개념 중 2개가 일치한다면 0.67을 정확도에 반영하였다. <표 34>는 그 결과를 보이고 있다.

<표 34> 위키 개념 태깅 정확도

방법	위키 개념 수		
	3 (98)	2 (274)	1 (417)
CS	약 51.4%	약 50.7%	약 47.7%
DCM	약 53.7%	약 54.2%	약 50.4%
SDI	약 67.3%	약 69.0%	약 65.7%
제안된 방법	약 78.9%	약 79.7%	약 75.5%

태깅 정확도가 카테고리 분류 정확도보다 낮은 것을 확인할 수 있는데, 그 이유로 첫째는 평가자에 의해 태깅된 위키 개념의 문맥 정보에 'computer'를 포함하지 않은 경우이다. 본 연구에서 사용한 위키 개념은 초록에 단어 'computer'를 포함하는 것인데, 컴퓨터 분야의 위키 문서 일지라도 그 단어를 포함하지 않은 경우가 다수 존재하였다. 예를 들어 'Knowledge representation and reasoning' 문서는 컴퓨터 과학(Computer Science)의 인공지능(Artificial Intelligence)에 깊이 관련된 것이지만 그 초록에는 'computer'를 포함하지 않고 있다. 예로 <표 30>에서 기술한 타겟 문서인 온톨로지 매핑에 관한 논문[88]의 경우는 평가자에 의해 위키 개념 'ontology engineering'과 'ontology merging'이 태거로 선정되었으나, 전자의 경우는 제대로 추출이 되었지만 후자의 경우는 본 실험에 반영되지 않았다. 또한 'semantic web'의 경우도 마찬가지였는데, 평가자에 의해 태깅된 개념 중에 이들이 일부 포함되어 있었던 것이 정확도를 낮게 만드는데 원인으로 작용하였다. 두 번째 이유로는 위키 개념의 문맥 정보 길이의 차이가 작용하였다. 위키 개념에 대한 문맥 정보의 평균 수는 일반 명사와 고유 명사를 포함하여 11.89개이지만, 3개 또는 4개로 구성된 위키 개념이 일부 존재하였다. 이들은 관련성이 존재하더라도 다른 것보다 낮게 측정되기 때문에 선정되지 못하는 결과를 만들었다. [그림 28]은 <표 34>의 내용을 그래프로 표현한 것이다.



[그림 28] 위키 개념 태깅 정확도 비교 평가

비록 이러한 이유로 카테고리 분류보다는 낮은 정확도를 얻었지만, 다른 연구들과 비교했을 때 여전히 본 방법이 가장 높은 정확도를 얻을 수 있음을 확인하였다. 이는 본 연구에서 추출한 문맥 정보를 추출하는 방법이 의미적이며, 문맥 정보를 선택하는 기준과 그들 사이의 유사도를 측정하는 방식이 효과적임을 내포할 수 있다.

Ⅵ. 결론

본 논문은 웹 문서의 검색에 표준성, 체계성, 의미성을 함께 제공할 수 있는 시맨틱 웹을 위해서 위키피디아의 문서 제목(위키 개념)을 이용하여 웹 문서를 태깅하는 방법에 대해 제안하였고 이를 위해 각 문서에서 핵심이 될 수 있는 문맥 정보를 의미적으로 선정하는 방법, 그 문맥 정보를 이용하여 유사도를 측정하는 방법을 제안하였다.

위키 개념, 위키 카테고리 그리고 웹 문서(타겟 문서)를 대표하는 문맥 정보를 형성하기 위해 출현 빈도를 고려한 키워드 가중치와 확장된 워드넷 기반의 의미성을 고려한 의미적 가중치를 측정하였으며, 이 두 가중치를 함께 반영함으로써 각 단어에 대한 문맥 가중치를 측정하였다. 또한 추출된 각 문맥 정보를 이용하여 타겟 문서의 위키 카테고리로의 분류와 위키 개념 태깅을 위해 그들 사이의 유사도를 측정하였다.

이와 같이 본 논문은 위키 개념으로 웹에 존재하는 문서들을 태깅하는 방법을 제안했고, 여러 가지 실험을 통하여 제안한 방법들의 타당성을 입증하였다. 문맥 정보 추출에 대한 관련 정확도 실험(상위 선정 30% 기준), 위키 카테고리로의 문서 분류 정확도 실험, 그리고 본 연구의 최종 목적인 위키 개념 태깅에서 기존에 수행된 연구들 보다 각각 최소 7%에서 최대 29%까지 향상 시킬 수 있었다.

본 논문에서 제안한 방법을 이용하여 위키 전체 개념에 대해 문맥 정보를 추출한다면, 웹에 존재하는 여러 문서에 대해 표준화된 태깅이 가능하며, 각 문서가 표현하는 복합적인 주제에 대해서도 태깅을 할 수 있다. 또한 이러한 위키 개념 태거와 위키 카테고리로의 분류를 통해 웹에 존재하는 임의의 문서들 사이의 의미적 연결을 도모할 수 있다. 이는 사용자가 입력한 표준화된 질의어에 대한 의미적 검색뿐만 아니라, 검색된 문서와 의미적으로 가까운 문서까지 제공할 수 있는 밑거름이 될 수 있음을 의미한다.

본 연구의 결과인 위키 태거들은 사용자에게 표준화된 검색 질의어를 추천할 수 있다. 또한 각 문맥 정보 내의 문맥 가중치를 이용하여 사용자의 임의 키워드 기반 의미적 검색이 가능할 뿐 아니라 키워드의 의미별로 검색 결과를 분류하여 제공할 수 있다. 이러한 문맥 정보는 문서들 사이의 의미적 유사도를 측정할 수 있는 기반데이터로 활용되며, 나아가서

사용자가 보유하고 있는 문서를 질의어로 입력함으로써 의미적으로 유사한 웹 문서들을 검색해줄 수 있다.

특히 본 연구에서 추출한 위키 문맥 정보는 본 연구뿐만 아니라 지식베이스 확장(knowledge base enrichment), 온톨로지 병합(ontology merging), 의미적 문서 연결(Semantic Document Interconnections), 질의 확장(query expansion) 및 변환(transformation) 등에 중요한 기반 데이터로 활용될 수 있다. 또한 문맥 가중치를 겸비한 문맥 정보는 문서에서 의미적으로 중요한 순서대로 키워드 인덱싱을 가능하게 하며, 이는 사용자가 입력한 질의어를 기반으로 한 의미적 검색을 제공할 수 있다.

하지만 지식베이스에 정의되지 않은 고유 명사와 전문 용어들에 대한 의미적 처리의 부재가 아쉬움으로 남았다. 이러한 용어들은 중의성이 적고 특정 문서에서 큰 의미를 가질 수 있기 때문에 정보 검색에서 주요한 단서로 작용할 수 있다. 또한 이들은 새로운 사회 현상, 트렌드, 기술이나 제품 생산 등에 의해 지속적으로 생성되고 있다. 만약 본 연구가 이러한 용어들의 의미적 처리까지 가능하게 된다면 그 성능은 더욱 우수해질 것으로 기대된다. 이러한 용어들을 자동으로 판단 및 추출하여 지식베이스에 추가하고, 그 지식베이스를 본 연구에 적용하는 것은 앞으로 해결해야할 과제로 남아있다.

참 고 문 헌

[1] Wardrip-Fruin, N. and Montfort, N., "The New Media Reader," Section 54. The MIT Press. ISBN 0-262-23227-8, 2003.

[2] Hwang, M.G., Choi, D.G., Choi, J.H., Kim, H.I., and Kim, P.K., "Similarity Measure for Semantic Document Interconnections," Information-An International Interdisciplinary Journal, Vol. 13, No. 2, pp. 253-267, 2010.

[3] Hemayati, R., Meng, W., and Yu, C., "Semantic-Based Grouping of Search Engine Results Using WordNet," Advanced in Data and Web Management, LNCS 4505, pp. 678-686, 2007.

[4] Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D., "Semantic annotation, indexing, and retrieval," Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 2, No. 1, pp. 49-79, Dec. 2004.

[5] Jovanovic, J., Gasevic, D., and Devedzic, V., "Ontology-Based Automatic Annotation of Learning Content," International Journal on Semantic Web and Information Systems, Vol. 2, No. 2, pp. 91-119, April, 2006.

[6] Handschuh, S., Staab, S., and Ciravegna, F., "S-CREAM - Semi-automatic CREAtion of Metadata, Knowledge Engineering and Knowledge Management," Ontologies and the Semantic Web, LNCS 2473, pp. 165-184. 2002.

[7] Alani, H., Kim, S., Millard, D., Weal, M., Hall, W., Lewis, P., and Shadbolt, N., "Automatic Ontology-based Knowledge Extraction from Web Documents," IEEE Intelligent Systems, Vol. 18, No. 1, pp. 14-21, Jan. 2003.

[8] Java, A., Nirenburg, S., McShane, M., Finin, T., English, J., and Joshi, A., "Using a Natural Language Understanding System to Generate Semantic Web Content," International Journal on Semantic Web and Information Systems, Vol. 3, No. 4, pp. 50-74, 2007.

[9] Kong, H.J., Hwang, M.G., and Kim, P.K., "The Method for the Unknown Word Classification," In Proceeding of The 2006 Pacific Rim Knowledge Acquisition Workshop, LNCS4303, pp. 207-215, August, 2006.

[10] Liu, S., Liu, F., Yu, C., and Meng, W., "An effective approach to document retrieval via utilizing WordNet and recognizing phrases," In Proceeding of SIGIR 2004, pp. 266-272, 2004.

[11] Navigli, R., and Velardi, P., "Ontology Enrichment Through Automatic Semantic Annotation of OnLine Glossaries," EKAW 2006: Managing Knowledge in a World of Networks, LNCS 4248, pp. 125-140, 2006.

[12] Velardi, P., Cucchiarelli, A., and Petit, M., "A Taxonomy Learning Method and Its Application to Characterize a Scientific Web Community," IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 2, pp. 180-191, Feb., 2007.

[13] Hwang, M.G., Choi, D.G., and Kim, P.K., "A Method for Knowledge Base Enrichment using Wikipedia Document Information," Information-An International Interdisciplinary Journal, Vol. 13, No. 5, 2010.

[14] Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E., and Milios, E., "Information Retrieval by Semantic Similarity," International Journal on Semantic Web and Information Systems, Vol. 2, No. 3, pp. 55-73, 2006.

[15] Missikoff, M., Velardi, P., and Fabriani, P., "Text Mining Techniques to Automatically Enrich a Domain Ontology," Applied Intelligence, Vol. 18, No. 3, pp. 323-340, 2003.

[16] Hwang, M.G., Choi, C., and Kim, P.K., "Automatic Enrichment of Semantic Relation Network and its Application to Word Sense Disambiguation," IEEE Transactions on Knowledge and Data Engineering, 03 Sept. 2010. IEEE computer Society Digital Library. IEEE Computer Society.

[17] Tim, B., Hendler, J., and Lassila, O., "The Semantic Web," Scientific American Magazine, May, 2001.

[18] Fellbaum, C., "WordNet: An Electronic Lexical Database," MIT Press.

[19] Navigli, R. and Velardi, P., "Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 7, pp. 1075-1086, July 2005.

[20] Hwang, M.G., Youn, B.S., Chung, I.Y., and Kim, P.K., "Semantic Measurement of Related Degree between Unknown Word and Related Word for Automatic Extension of Lexical

Dictionary," in Proceedings of Fifth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 5, pp.484-488, 2008.

[21] Hwang, M.G. and Kim, P.K. "A New Similarity Measure for Automatic Construction of the Unknown Word Lexical Dictionary," International Journal on Semantic Web & Information Systems, vol. 5, no. 1, pp. 48-64, 2009.

[22] Scalno, F., and Velardi, P., "TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities," In proceedings of Conference TIA-2007, Sophia Antipolis, October 2007.

[23] Magnini, B. and Cavaglia, G., "Integrating Subject Field Codes into WordNet," In Proceedings of Second International Conference Language Resources and Evaluation (LREC2000), pp. 1413-1418, 2000.

[24] Miller, G.A., Chodorow, M., Landes, S., Leacock, C., and Thomas, R.G., "Using a Semantic Concordance for Sense Identification," In proceedings of ARPA Human Language Technology Workshop, pp. 240-243, 1994.

[25] Ng, H.T. and Lee, H.B., "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach," In proceedings of 34th Annual Meeting Association for Computational Linguistics, 1996.

[26] Crowther, J., Dignen, S., and Lea, D., "Oxford Collocations Dictionary for Students of English," Oxford University Press, 2002.

[27] Longman, K., "Longman Language Activator," Pearson Education, 2003.

[28] Rada, R., Mili, H., Bicknell, E., and Blettner, M., "Development and Application of a Metric on Semantic Nets," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 19, No. 1, pp. 17-30, 1989.

[29] Wu, Z., and Palmer, M., "Verb Semantics and Lexical Selection," In: Annual Meeting of the Associations for Computational Linguistics (ACL'94), pp. 133-138, 1994.

[30] Li, Y., Bandar, Z.A., and McLean, D., "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources," IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, pp. 871-882, 2003.

[31] Leacock, C., and Chodorow, M., "Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet," In An Electronic Lexical Database. MIT Press, pp. 265-283, 1998.

[32] Richardson, R., Smeaton, A., and Murphy, J., "Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words," Techn. Report Working paper CA-1294, School of Computer Applications, 1994.

[33] Lord, P., Stevens, R., Brass, A., and Goble, C., "Investigating Semantic Similarity Measures across the Gene Ontology: the Relationship between Sequence and Annotation," Bioinformatics, Vol. 19, No. 10, pp. 1275-1283, 2003.

[34] Resnik, O., "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language," Journal of Artificial Intelligence Research, Vol. 11, pp. 95-130, 1999.

[35] Lin, D., "Principle-Based Parsing Without Overgeneration," In: Annual Meeting of the Association for Computational Linguistics (ACL'93), pp. 112-120, 1993.

[36] Jiang, J., and Conrath, D., "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," In: Intern. Conf. on Research in Computational Linguistics, 1998.

[37] Seco, N., Veale, T., and Hayes, J., "An Intrinsic Information Content Metric for Semantic Similarity in WordNet," Techn. report, University College Dublin, Ireland, 2004.

[38] Tversky, A., "Features of Similarity," Psycological Review, Vol. 84, No. 4, pp. 327-352, 1977.

[39] Wikipedia - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Wikipedia, 3 Nov., 2010.

[40] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S., "DBpedia - A Crystallization Point for the Web of Data," Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 7, pp. 154-165, 2009.

[41] Suchanek, F.M., Kasneci, G., and Weikum, G., "YAGO: A Core of Semantic Knowledge - Unifying WordNet and Wikipedia," In proceedings of 16th International World Wide Web Conference, pp. 697-706, 2007.

[42] Reed, S., and Lenat, D.B., "Mapping Ontologies into Cyc," In AAAI 2002 Conference Workshop on Ontologies For The Semantic Web, July 2002.

[43] Kong, H.J., Hwang, M.G., Hwang, G.S., Shim, J.H., and Kim, P.K., "Topic Selection of Web Documents Using Specific Domain Ontology," MICAI 2006: Advances in Artificial Intelligence, LNCS 4293, pp. 1047-1056, 2006.

[44] Chang, J.H., Lee, J.W., Kim, Y.S., and Zhang, B.T., "Topic Extraction from Text Documents Using Multiple-Cause Networks," In Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligences: Trends in Artificial Intelligence, LNCS 2417, pp. 434-443, 2002.

[45] Ahmad, R. and Khanum, A., "Document Topic Generation in Text Mining by Using Cluster Analysis with EROCK," International Journal of Computer Science and Security, pp. 176-182, 2010.

[46] Hwang, M.G., Kong, H.J., Baek, S.K., and Kim, P.K., "TSM. Topic Selection Method of web Documents," In Proceedings of the First Asia International Conference on Modelling & Simulation, pp. 369-374, 2007.

[47] 황명권, 배용근, 김판구, "문서 내용의 계층화를 이용한 문서 비교 방법", 한국해양정보통신학회논문지, 제10권 12호, pp.2335-2342, 2006년 12월.

[48] Chen, L. and Roberts, C., "Semantic Tagging for Large-scale Content Management," In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, pp. 478-481, 2007.

[49] Parry, D.T. and Tsai, T.C., "Crowdsourcing techniques to create a fuzzy subset of SNOMED CT for semantic tagging of medical documents," In Proceedings of 2010 IEEE International Conference on Fuzzy Systems, pp. 1-8, 2010.

[50] Martin, T., Shen, Y., and Azvine, B., "Automated semantic tagging using fuzzy grammar fragments," In Proceedings of IEEE International Conference on Fuzzy Systems, pp. 2224-2229, 2008.

[51] Jang, H.J., Song, S.K., and Myaeng, S.H., "Semantic Tagging for Medical Knowledge Tracking," In Proceedings of Engineering in 28th Annual International Conference of the IEEE

Medicine and Biology Society, pp. 6257-6260, 2006.

[52] Jang, H.J., Jin, Y., and Myaeng, S.H., "Integration of Low Level Linguistic Information for Clinical Document Semantic Tagging," In Proceedings of IEEE International Conference on Information Reuse and Integration, pp. 292-297, 2006.

[53] Yelloz, J. and Taehyung Wang, "Optimization and Load Balancing of the Semantic Tagging and Searching System," In Proceedings of IEEE International Workshop on Semantic Computing and Applications, pp. 43-50, 2008.

[54] Hope, G., Wang, T.H., and Barkataki, S., "Convergence of Web 2.0 and Semantic Web: A Semantic Tagging and Searching System for Creating and Searching Blogs," In Proceedings of the First IEEE International Conference on Semantic Computing, pp. 43-50, 2007.

[55] Yang, H.C., "Bridging the WWW to the Semantic Web by Automatic Semantic Tagging of Web Pages," In Proceedings of The Fifth International Conference on Computer and Information Technology, pp. 238-242, 2005.

[56] Chandramouli, K., Kliegr, T., Svatek, V., and Izquierdo, E., "Towards semantic tagging in collaborative environments," In Proceedings of 16th International Conference on Digital Signal Processing, pp. 1-6, 2009.

[57] Zadeh., L.A., "Fuzzy Sets," Information Control, Vol. 8, pp. 338-353, 1965.

[58] Rabiner, L.R. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, Vol. 77, No. 2, pp. 257-286, 1989.

[59] Chandramouli, K., Kliegr, T., Nemrava, J., Svatek, V., and Izquierdo, E., "Query Refinement and User Relevance Feedback for Contextualized Image Retrieval," In Proceedings of the 5th International Conference on Visual Information Engineering, pp. 453-458, 2008.

[60] Kaiser, F., "Using Wikipedia-based conceptual contexts to calculate document similarity," In Proceedings of Third International Conference on Digital Society, pp. 322-327, 2009.

[61] Schonhofen, P., "Identifying document topics using the Wikipedia category network," In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, pp. 456-462, 2006.

[62] Huang, A., Milne, D., Frank, E., and Witten, I.H., "Clustering Documents with Active Learning Using Wikipedia," In Proceedings of Eighth IEEE International Conference on Data Mining, pp. 839-844, 2008.

[63] Milne, D. and Witten, I.H., "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," In Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence, 2008.

[64] Schonhofen, P. "Annotating Documents by Wikipedia Concepts," In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 461-467, 2008.

[65] Ruiz-casado, M., Alfonseca, E., and Castells, P., "Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets," In Proceedings of the Atlantic Web Intelligence Conference, LNCS 3528 of Lecture Notes in Computer Science, pp. 380-386, 2005.

[66] Strube, M. and Ponzetto, S.P., "WikiRelate! computing semantic relatedness using wikipedia," In Proceedings of the 21st national conference on Artificial intelligence, Vol. 2, pp. 1419-1424, 2006.

[67] Banerjee, S., Ramanathan, K., and Gupta, A., "Clustering short texts using wikipedia," In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 787-788, 2007.

[68] Gabrilovich, E. and Markovitch, S., "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis," In Proceedings of the 20th international joint conference on Artifical intelligence, pp. 1606-1611, 2007.

[69] Adafre, S.F. and Rijke, M., "Discovering Missing Links in Wikipedia," In Proceedings of the 3rd international workshop on Link discovery, pp. 90-97, 2005.

[70] Hu, J., Fang, L.J., Cao, Y., Zeng, H.J., Li, H., Yang, Q., and Chen, Z., "Enhancing text clustering by leveraging Wikipedia semantics," In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 179-186, 2008.

[71] Chemov, S., Iofciu, T., Nejdl, W., and Zhou, X., "Extracting semantic relationships between

wikipedia categories," In 1st International Workshop: SemWiki2006 - From Wiki to Semantics, pp. 153-163, 2006.

[72] Hepp, M., Bachlechner, D., and Siorpaes, K., "Harvesting Wiki Consensus-Using Wikipedia Entries as Ontology Elements," First Workshop on Semantic Wikis, pp. 124-138, 2006.

[73] Cucerzan, S., "Large-scale named entity disambiguation based on Wikipedia data," In Proceedings of Empirical Methods in Natural Language Processing, 2007.

[74] Milne, D. and Witten, L.H., "Learning to Link with Wikipedia," In Proceeding of the 17th ACM conference on Information and knowledge management, pp. 509-518, 2008.

[75] Gregorowicz, A. and Kramer, M.A., "Mining a Large-Scale Term-Concept Network from Wikipedia," Mitre Technical Report 06-1028, October 2006.

[76] Gabrilovich, E. and Markovitch, S., "Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge," In Proceedings of the 21st national conference on Artificial intelligence, pp. 1301-1306, 2006.

[77] Volkel, M., Krotzsch, M., Vrandecic, D., Haller, H., and Studer, R., "Semantic Wikipedia," In Proceedings of the 15th International Conference on World Wide Web, pp. 585-594, 2006.

[78] Kaiser, F., Schwarz, H., and Jakob, M., "Using Wikipedia-Based Conceptual Contexts to Calculate Document Similarity," In Proceedings of the 2009 Third International Conference on Digital Society, pp.322-327, 2009.

[79] Minier, Z., Bodo, Z., and Csato, L., "Wikipedia-based Kernels for Text Categorization," In Proceedings of the Ninth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, pp. 157-164, 2007.

[80] Toutanova, K. and Manning, C., "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger," In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70, 2000.

[81] Toutanova, K., Klein, D., Manning, C., and Singer, Y., "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language

Technology, pp. 173-180, 2003.

[82] Computer - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Computer, 3 Aug., 2010.

[83] A* search algorithm - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/A-star_search_algorithm, 3 Aug., 2010.

[84] Java (software platform) - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Java_(software_platform), 3 Aug., 2010.

[85] Apple Inc. - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Apple_Inc, 3 Aug., 2010.

[86] Amit Sheth - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Amit_Sheth, 3 Aug., 2010.

[87] Choi, D.J., Hwang, M.G., and Kim, P.K., "Semantic Context Extraction from Wikipedia Document," In Proceedings of The 2010 International Conference on Semantic Web & Web Services, pp. 38-41, July 2010.

[88] Jung, J.J. "Reusing Ontology Mappings for Query Routing in Semantic Peer-to-peer Environment," Information Science, Vol. 180, No. 17, pp. 3248-3257, Sep., 2010.

[89] Hwang, M.G., Choi, C., Youn, B.S., and Kim, P.K., "Word Sense Disambiguation based on Relation Structure," In Proceedings of the 2008 International Conference on Advanced Language Processing and Web Information Technology, pp. 15-20, 2008.

저작물 이용 허락서

학 과	컴퓨터공학과	학 번	20067327	과 정	박사
성 명	한글: 황 명 권 한문: 黃 明 權 영문: HWANG, MYUNG GWON				
주 소	광주광역시 남구 진월동 고운하이플러스 102동 807호				
연락처	E-mail: mg.hwang@gmail.com				
논문제목	[한글] 지능적인 웹 검색을 위한 의미적 문서 태깅 방법 연구				
	[영문] Semantic Document Tagging Methods for Intelligent Web Retrieval				

본인이 저작한 위의 저작물에 대하여 다음과 같은 조건 아래 조선대학교가 저작물을 이용할 수 있도록 허락하고 동의합니다.

- 다 음 -

1. 저작물의 DB구축 및 인터넷을 포함한 정보통신망에의 공개를 위한 저작물의 복제, 기억장치에의 저장, 전송 등을 허락함.
2. 위의 목적을 위하여 필요한 범위 내에서의 편집과 형식상의 변경을 허락함. 다만, 저작물의 내용변경은 금지함.
3. 배포·전송된 저작물의 영리적 목적을 위한 복제, 저장, 전송 등은 금지함.
4. 저작물에 대한 이용기간은 5년으로 하고, 기간종료 3개월 이내에 별도의 의사 표시가 없을 경우에는 저작물의 이용기간을 계속 연장함.
5. 해당 저작물의 저작권을 타인에게 양도하거나 출판을 허락을 하였을 경우에는 1개월 이내에 대학에 이를 통보함.
6. 조선대학교는 저작물 이용의 허락 이후 해당 저작물로 인하여 발생하는 타인에 의한 권리 침해에 대하여 일체의 법적 책임을 지지 않음.
7. 소속 대학의 협정기관에 저작물의 제공 및 인터넷 등 정보통신망을 이용한 저작물의 전송·출력을 허락함.

동의여부 : 동의(○) 반대( )

2011년 2월 25일

저작자: 황 명 권 (인)

조선대학교 총장 귀하

감사의 글

인생에 구체적인 목표가 없었던 저에게 큰 꿈을 향해 달릴 수 있도록 조언해 주시고 박사학위를 가질 수 있도록 도와주신 분들이 계십니다. 그동안 직접적으로 표현하지 못했던 말들을 영원히 남을 이 페이지에 기록하고자 합니다.

가장먼저 저의 불투명한 미래에 대해 저의 계획과 의지만을 믿고 평생 함께할 약속을 맺고 항상 저의 즐거움과 행복이 되고 있는 저의 집사람 혜준에게 사랑과 고마움을 표현하고 싶습니다. 그리고 저희를 믿고 지긋이 지켜봐주시며 사위를 친아들처럼 대해주시는 장인어르신과 장모님께 감사의 말씀을 전합니다.

저의 학업 수행 과정에서 지도교수님으로써 따끔한 훈계와 훈훈한 관심과 애정으로 저를 지도해주시고, 어려운 상황에 대해 저의 편에 서서 조언 및 지지해주시며, 미래의 진로에 대해 함께 고민해주신 김판구 교수님께 감사의 말씀을 전합니다.

그리고 저의 석사 박사과정 동안 연구의 성취감과 즐거움을 공유한 선배, 후배 및 동기들이 있습니다. 따뜻한 마음으로 항상 저를 대해 주시며, 심지어는 양말까지 선물해주셨던 신주현 선생님께 감사의 말씀을 드립니다. 익살스러움과 진지함을 함께 겸비한 김원필 선배님, 무뚝뚝하지만 진심으로 후배들을 이끌어주시는 최준호 선배님, 저에게 직접적으로 논문작성, 프로젝트 수행에 대해 가르쳐주며 그 성취감과 즐거움을 가장 많이 나누었으며 국내외 여행을 함께 하며 많고 다양한 추억을 나눈 공현장 선배님과 백선경 선배님, 차분하게 얘기하는 믿음직스러움과 간혹 엉뚱함으로 웃음을 주신 정관호 선배님, 작은 체구임에도 지치지 않는 연구 수행 에너지를 보여주었던 조미영 박사, 연구실의 살림을 도맡으며 많은 업무를 배분하고 관리하는 최창, 중국 유학생으로써 한국 문화에 대해 잘 적응하며 외국어 학습에 대한 동기부여를 해주었던 송단과 유해도, 컴퓨터의 다양한 부분에 깊은 관심과 지식을 가진 김성석, 큰 등치에 맞는 믿음직스러움과 업무 수행에 책임감을 보여주었던 황광수, 네팔 유학생으로써 다양한 문화에 대해 말해주며 저의 간접경험을 쌓게 해주었던 디페시 가우탐, 교수님과 닮은 외모를 가지며 저의 잔소리로 가장 고생한 윤병수, 일에 대한 빠른 이해력과 꼼꼼한 업무 수행 능력을 가진 최동진, 다양한 프로그래밍 능력을 보여 연구의 진척에 도움이 되고 여러 운동도 함께 했던 김동철, 항상 조용하게 연구실을

지키는 이효갑, 막내 석사과정으로 연구실의 다양한 업무를 수행하느라 고생하는 고병규, 그리고 학부생 김정인과 김하영, 위 모든 선배, 후배 및 동기들에게 고마움을 전합니다.

그리고 저와 함께 일상에서 즐겁게 지내고 있는 가족이 있습니다. 먼저 저와 어린시절을 함께 보내며 많은 추억을 공유하고 친구와 같은 편안함을 지닌 형님(황춘근), 철없는 형님을 따뜻함과 애정으로 가르치시느라 고생하시며 저와 저의 처 그리고 부모님께도 늘 한가족처럼 대해주시는 형수님(백형옥), 앞으로 큰 인재가 될 예비조카(복땡이), 작으면서도 강인함을 가지며 철없는 형부와 남편의 행동을 지긋이 지켜봐주는 처제(문혜수), 아직 이루고자 하는 일을 위해 끊임없이 도전하는 동서(김하섭), 사회 초년생으로써 바쁜 나날을 보내며 간혹 철든 행동에 가족들에게 뿌듯함을 안겨주는 처남(문종현), 탄생과 동시에 처갓집에서 큰사위인 나의 존재감을 앗아가며 그 대신 새생명에 대한 신비감을 알려주며 사랑과 기쁨을 함께 선물해준 조카(김재준)에게 감사의 말씀을 전합니다.

이 글을 통해 표현하지는 않았지만 저의 삶에 활력이 되는 친구들, 교수님들, 선배님들, 후배들이 있습니다. 이 순간에도 떠오르는 여러 얼굴들이 있지만 따로 표기하지 않아도 저의 마음을 알 수 있기에 생략하도록 하겠습니다.

마지막으로 위의 모든 사람들과 인연을 맺으며 인생의 즐거움과 행복을 느낄 수 있도록 저를 키워주시고 진로에 대한 저의 결정을 믿고 적극적으로 지지해주신 부모님께 감사의 말씀을 드립니다.