집단지성을 활용한 시소러스 갱신에 관한 연구: 위키피디아를 중심으로*

Thesaurus Updating Using Collective Intelligence: Based on Wikipedia Encyclopedia

한승희(Seung-Hee Han)**

초 록

이 연구에서는 위키피디아를 활용하여 시소러스를 갱신하고, 그 결과를 평가함으로써 시소러스 갱신에 있어 집단지성의 활용가능성에 대해 확인하고자 하였다. ASIS&T 시소러스를 대상으로 시소러스를 갱신한 결과, 용어 포괄성의 측면에서 ASIS&T 시소러스에 비해 위키 시소러스가 우수한 것으로 나타났다. 또한, 갱신된 시소러스를 평가한 결과, 위키피디아가 시소러스 갱신에 활용될 수 있음이 증명되었다. 특히, 리디렉션, 카테고리, 상호 링크로 요약되는 위키피디아의 구조적 특성은 시소러스의 의미관계를 추출하는 데 있어 적합하다는 것을 확인하였다. 이 연구의 결과를 일반화하기 위해 다국어 시소러스를 포함한 다양한 시소러스를 대상으로 적용해 볼 필요가 있다.

ABSTRACT

The purpose of this study is to suggest how the classic thesaurus structure of terms and links can be mined and updated from Wikipedia encyclopedia, which is the best practice of collective intelligence. In a comparison with ASIS&T thesaurus, it was found that Wikipedia contains a substantial coverage of domain-specific concepts and semantic relations. Furthermore, it was resulted that the structural characteristics of Wikipedia, such as redirects, categories, and mutual links are suitable to extract semantic relationships of thesaurus. It is needed to apply to update various thesauri, including multilingual thesaurus, in order to generalize the results of this research.

키워드: 시소러스 갱신, 시소러스 유지관리, 시소러스, 집단지성, 위키피디아
thesaurus update, thesaurus maintenance, thesaurus, collective intelligence, Wikipedia

---

* 이 논문은 2007학년도 서울여자대학교 사회과학연구소 교내학술연구비의 지원을 받았음.
** 서울여자대학교 사회과학대학 문헌정보학과 조교수(hanshee@swu.ac.kr)
■ 논문접수일자 : 2009년 6월 23일 ■ 최초심사일자 : 2009년 6월 26일 ■ 게재확정일자 : 2009년 7월 1일
■ 정보관리학회지, 26(3): 25-43, 2009. 〔DOI:10.3743/KOSIM.2009.26.3.025〕

1. 연구의 목적 및 필요성

오늘날 지식정보사회의 핵심 역량은 방대한 정보 속에서 정확한 정보를 검색해내는 능력이라고 할 수 있다. 그러나 정보탐색 및 검색 과정에서 이용자가 직면하는 가장 큰 과제 중 하나는 바로 질의를 작성하기 위해 탐색어를 선정하는 일일 것이다. 일반적으로 정보검색 시스템의 정확률 향상을 위해서는 색인어와 같은 정보표현의 품질뿐만 아니라, 이용자의 정보요구가 질의의 형태로 얼마나 효과적으로 표현되었는가의 여부도 중요한 영향을 미친다. 그렇기 때문에, 정보 시스템의 규모가 커질수록 효율적인 정보검색 도구에 대한 필요성이 증대되었으며, 정보검색 효율을 높이기 위해 용어 간의 관계를 정의한 시소러스의 중요성이 크게 부각되었다.

시소러스는 색인 시에는 색인어의 선정에, 탐색 시에는 질의어와 색인어 간의 관계를 연결해주는 역할을 한다(최석두 2000). 특히 정보검색에서 시소러스는 질의 구축과 확장에 있어 탐색어 선정을 위한 하나의 중요한 도구가 된다. 정보환경이 변화하면서 시소러스의 이용목적 역시 초기의 색인도구에서 검색도구로 변화함에 따라, 많은 정보검색 시스템에서 시소러스를 개별 혹은 통합하여 이용하고 있으며, 웹 기술의 발전으로 인해 VisualThesaurus (http://www.visualthesaurus.com/) 등과 같은 다양한 온라인 시소러스가 등장하였다(김지훈 2002).

앞에서 언급한 바와 같이, 정보검색 환경에서 시소러스의 필요성에 대해 인지하고 있음에도 불구하고, 구축에 많은 비용과 시간이 소요되기 때문에 일부 분야에 한해서만 시소러스가 구축되어 있다. 특히 오늘날 사용되고 있는 대부분 시소러스의 내용과 구조는 시소러스 구축에 관한 국가 및 국제적 지침에 따라 특별한 변화 없이, 1960년대에 만들어진 이후로 지금까지 지속되고 있어, 시간의 흐름에 따른 해당 분야의 변화를 신속하게 반영하여 개정하지 못하고 있는 실정이다(김지훈, 김태수 2006).

시소러스를 구축하는 데 많은 시간과 비용이 소요된다면, 그 대안으로 기존 시소러스를 갱신하는 접근법을 고려해 볼 필요가 있다. 일반적으로 시소러스는 도서관 분류체계보다 범위가 좁기 때문에 분류표보다 더 자주 갱신되어야 하며(최석두 2000), 일정한 검색효율을 유지하기 위해 유지보수 혹은 갱신작업이 지속적으로 수반되어야 한다(남영준, 이영주 2002). 결국, 시소러스의 갱신도 일회성으로 그치는 것이 아니라, 지속적인 작업을 통해 용어의 최신성을 유지해야만 한다.

일반적으로 전문가에 의해 이루어지는 시소러스 갱신에 대해 다양한 접근법이 있으나, 지속적인 갱신에 소요되는 시간 및 비용 문제를 해결할 뿐만 아니라, 용어의 최신성을 유지할 수 있다는 측면에서 협력적 태깅(collaborative tagging) 등과 같은 집단지성(collective intelligence)을 활용하는 것도 하나의 방법이 될 수 있다. 이러한 관점에서, 이 연구의 목적은 집단지성을 활용하여 시소러스를 갱신하는 방법을 제안하고, 이 방법이 시소러스를 갱신하는 데 있어 효과적인가를 확인하는 것에 있다. 구체적으로는 집단지성의 모범적 사례라고 할 수 있는 위키피디아(http://www.wikipedia.org)를 활용하여 시소러스를 갱신하고, 그 결과를 평가

함으로써 시소러스 갱신에 있어 집단지성의 활용가능성에 대해 확인하고자 한다.

2. 집단지성과 지식관리

2.1 집단지성의 개념

집단지성이란 다수의 개체들이 서로 협력 혹은 경쟁을 통하여 얻게 되는 지적 능력에 의한 결과로 얻어진 집단적 능력을 말한다. 대중의 지혜(wisdom of crowds) 또는 군중지혜(swarm intelligence)라고도 불리는 이 개념은 원래 1910년대 하버드 대학 교수이자 곤충학자인 William Morton Wheeler가 개미의 사회적 행동을 관찰하면서 처음 제시한 것으로, 이후 사회학, 컴퓨터과학, 심리학 등의 세부영역으로 박테리아, 식물, 동물, 인간 사회의 행동까지 다양한 대상에 응용되어 왔다(위키백과 2009).

집단지성이 오늘날의 웹 2.0을 이해하기 위한 개념으로 널리 알려진 데에는 사회학적 영향력이 컸다고 할 수 있는데, 특히 사회학자인 Pierre Levy는 디지털 기술의 문화적·인식론적 영향과 사회적 활용에 대해 연구하였다. 그는 집단지성을 '어디에서나 분포하고, 지속적으로 가치가 부여되며, 실시간으로 조정되어, 역량의 실제적 동원에 이르는 지성'이라고 정의하고, 과학기술을 통해 인류사회는 공동의 지적 능력과 자산을 서로 소통하면서 집단지성을 쌓아왔으며, 이를 통해 시공간의 제약을 극복한 인류의 진정한 통합과 새로운 진화가 완성단계에 이를 수 있다고 주장하였다(피에르 레비 2002).

집단지성은, 기본적으로 보통 사람들은 이성적 판단을 내릴 수 있는 존재이며, 집단은 어떤 개인이 가진 능력의 합이나 똑똑한 몇몇의 전문가들보다 나은 판단을 할 수 있을 것이라고 가정한다. 이와 관련하여 James Surowiecki는 집단의 현명함을 보장하는 네 가지의 중요한 속성을 제시하고 있다. 첫째는, 다양성이 있어야 하고, 둘째는, 집단이 분권화되어 있어야 하고, 셋째는, 구성원의 의견이 정리되고 모아져 하나의 결정을 만들어 낼 수 있는 방법론을 가지고 있어야 하며, 마지막으로, 다른 구성원의 의견에 영향받지 않도록 구성원이 상호 독립적이어야 한다는 것이다(제임스 서로위키 2004).

2.2 웹 2.0과 집단지성

앞에서 언급한 사회학적 관점에서의 집단지성은 웹 환경의 변화를 주도하기 위한 방법론으로 논의되기 시작하였다(박재천, 신지웅 2007). 집단지성은 웹 2.0과 동일한 사상적 바탕을 가지고 있다. 즉, 웹 2.0 구현을 위한 참여, 공유, 개방, 협력은 집단지성에서도 필수적인 요소라고 할 수 있는데, 집단이 자유롭게 참여할 수 있고, 집단이 생산한 지식이 집단 내에서 자유롭게 공유되며, 지식 생산을 위한 협력이 아무런 제약 없이 이루어질 수 있는 개방적 환경이 조성되어야 집단지성이 발현될 수 있다.

웹 2.0 환경에서 집단지성 플랫폼을 구현하기 위해서는 무엇보다도 지식이 발전할 수 있는 형태를 고려해야 한다. 이를 가능하게 하기 위해 집단지성의 기술적 구현에 필요한 두 가지 기능이 필요하다(Heylighen 1999). 우선, 피드백 기능은 최초에 발의되었던 지식이 집단의 활동을 통해 발전하도록 돕는다. 즉, 피드백 기

능은 이용자 간 토론을 통해 얻어진 지식을 반영하는 기술과 최초 지식에서 토론을 거쳐 수정된 지식이 다시 논의를 거쳐야 하는 여부를 판단해 줄 수 있다. 또한, 집단지성이 최대한의 지식을 끌어내기 위해서는 이용자의 활동을 기술적으로 모니터링해야 한다. 그 이유는 집단 내의 구성원은 모두 순기능적 역할을 하는 것이 아니며, 악의적인 활동을 하는 이용자도 존재하기 때문이다. 모니터링은 악의적인 이용자에게는 일정한 제약을, 적극적인 참여자에게는 일정한 인센티브를 주어 플랫폼의 활성화를 이루게 하는 필수적인 요소라고 할 수 있다.

오늘날 인터넷에서는 웹 2.0 환경에서 집단지성을 활용한 사례들을 쉽게 찾아볼 수 있다. 구글(http://www.google.com)이 세계 최대 검색 사이트가 될 수 있었던 이유 중 하나는 웹 문서의 중요도를 판별하는 '페이지 랭크(Page Rank)'라는 구글의 알고리즘이 집단 지성, 대중의 지혜를 반영했기 때문으로 볼 수 있다. 구글은 링크가 많을수록, 특히 권위 있는 사이트의 링크가 많을수록 높은 점수를 배정하는 식으로 검색결과를 순위화하고 있다. 또한, del.icio.us (http://del.icio.us)나 CiteULike(http://www.citeulike.org) 등과 같은 소셜 북마킹(social bookmarking) 서비스는 이용자가 개인적인 목적으로 보관하고 분류해 둔 북마크에서 '대중의 지혜'를 이용해 정보의 중요도를 결정하고, 분류하고, 탐색하게 해 준다. 이러한 사례들은 링크와 이용자 선호도라는 이용자 행위를 이용하여 집단지성이 갖추어야 할 피드백 기능을 잘 설명해주고 있다.

집단 지성이 웹 2.0의 핵심으로 떠오르면서 올바른 집단지성을 만들어내는 시스템의 중요성도 커지고 있다. 앞에서 설명한 구글의 페이지 랭크나 아마존(http://www.amazon.com)의 추천 시스템과 같이 이용자 각각의 행위(링크나 상품 구매)에 의해 집단 지성이 만들어지는 경우는 큰 문제가 되지 않는다. 그러나 위키피디아와 같이 온라인 백과사전을 만드는 데 이용자가 직접 참여하는 경우에는 부정확하거나 악의적인 정보가 집단지성의 폐해로 작용할 수 있기 때문에, 이러한 폐해를 효과적으로 걸러줄 수 있는 필터링과 편집 기능이 필요하다. 위키피디아는 이러한 문제를 해결하기 위해 사실이 아닌 정보에 대해서는 공동의 작업에 의해 필터링되어 삭제되도록 함으로써 컨텐츠의 품질을 관리하고 있다.

2.3 협업에 의한 지식관리

집단지성의 개념은 웹 2.0 환경에 응용되기 전, 이미 지식관리의 영역에서 협업기반 지식관리(collaboration-based knowledge management)라는 개념으로 널리 적용되어 왔다. 전통적 관점의 지식관리에서는 지식을 상품으로 여기는 반면, 최근의 지식관리는 지식노동자간 협업, 커뮤니케이션 및 사회적 네트워크의 개발을 지원하는 방향으로 발전하고 있다 (Fischer 2003).

〈표 1〉에서 보는 바와 같이, 기존의 지식생성은 소수의 전문가 집단이 주도하고 그 내용을 평가함으로써 단방향으로 진행되어 왔다. 이러한 지식생성의 패턴은 관련 주제를 바탕으로 한 학회나 협회, 학술연구단체 등과 같은 단체의 권위를 바탕으로 지식과 정보의 소유에 있어 독과점의 형태를 지니게 되며, 생산 형태

〈표 1〉 전통적 지식관리와 협업기반 지식관리 비교(Fischer 2003)

구 분	전통적 지식관리	협업기반 지식관리
지식생성	전문가, 전문활동에 의해	누구나, 협업활동에 의해
지식통합	시스템 설계 시	지식 이용 시
지식배포	강의, 방송, 수업 등	온디맨드, 학습과 업무의 통합, 개인화형태
학습 패러다임	지식전이	협업기반 지식구축
업무	시스템 중심	이용자/업무 중심
사회구조	계층적, 상향식 커뮤니케이션	커뮤니티 기반 피어 투 피어 커뮤니케이션
업무 패턴	표준화	경험기반
정보환경	폐쇄형, 정적	개방형, 동적
정보접근	푸시	출판 및 구독

의 특성상 그 영향력은 웹에 비해 시·공간적으로 제한적일 수밖에 없다. 또한 강의, 방송, 학습과 같은 지식배포의 패러다임은 지식의 수정과 재배포를 매우 느리게 하는 특징을 갖는다(김학래, 최재화, 김홍기 2006).

이에 반해 협업을 통한 지식의 생산은 개인 중심의 협업 활동을 통해 이루어짐으로써 정보의 생산과 소비가 동시적으로 진행되는 양방향성을 지닌다. 게시판, 댓글, 트랙백, RSS와 같은 웹 2.0 기술은 정보 생산자의 확대를 가져왔으며, 정보 가치의 판단에 있어 대중적 가치를 우선시하는 방향으로 바꾸어 놓았다. 또한 네트워크라는 기술적 특징 때문에 상호연계성이 중시되며, 지식의 수정과 재배포가 실시간으로 가능해지게 되었다(석영희 2006).

또한, 전통적인 지식관리 환경에서 이용자는 자신의 분야와 연관해서만 문제를 해결하려고 하는 경향이 있는데, 이는 복합적인 문제의 해결에 있어 문제점으로 작용할 수 있다. 또한 전문적인 식견을 가지고 있더라도 다른 사람간의 상호교류가 없다면 해결방안에서 오류를 가질 수 있다. 이러한 점을 해결하기 위한 한 방안으로 지식관리에서 집단지성의 활용이 부각되었는데, 다양한 계층이 동적으로 문제해결에 참여하고 각자의 의견을 제시함으로써 다양한 문제해결방법을 도출할 수 있다(Noubel 2007).

협업에 의한 지식관리는 다음과 같은 관점에서 의미가 있다(Mimul's Developer World, 2007년 11월 17일 기사). 첫째, 동료에 의한 지식의 검토가 가능하다. 둘째, 등록된 지식을 순위화함으로써 집단지성에 의한 이용자의 선호도와 관심을 표현할 수 있다. 셋째, 리뷰를 통해 경험을 공유할 수 있다. 넷째, 등록된 지식 및 지식 생산자간 관계를 나타내는 지식 네트워크의 분석을 통해 보다 효율적인 지식의 검색과 활용이 가능하다. 마지막으로, 다섯째, 새로운 지식의 창조가 가능하다.

3. 위키피디아를 활용한 시소러스 갱신

3.1 위키피디아

3.1.1 위키피디아 개요

위키피디아는 2001년 Jimmy Wales와 Larry

Sanger에 의해 시작된 위키(Wiki) 기반 무료 백과사전 프로젝트로, 현재 인터넷 상에서 가장 잘 활용되고 있는 참고정보원 중 하나이며, 동시에 인터넷 기술을 기반으로 새로이 등장한 집단지성의 좋은 사례로 손꼽히고 있다(Wikipedia 2009).

위키피디아는 다양한 배경의 지성적 존재들이 단순히 지적 결과물을 만들어 내는 데 그치지 않고, 상호작용을 통한 내용의 확인과 확대, 그리고 연결을 유기적으로 진행하고 있다. 위키피디아 이용자는 서로 독립적이며, 다양한 배경을 가지고 있고, 기존의 사회와는 달리 위계질서가 명확하지 않다. 위키피디아는 내용의 접근 및 편집에 대해서는 거의 제한이 없으나, 데이터의 품질은 비교적 엄격하게 내용을 검사하고 상호 비판을 통해 정리해 나가는 방식으로 평판과 동료평가에 의해 관리·운영되고 있다(위키백과 2009). 〈표 2〉에서는 2009년 6월 현재 영문판 위키피디아 통계를 보여주고 있다. 3억 회 이상의 편집횟수 및 18회 이상의 페이지당 평균 편집횟수를 볼 때, 위키피디아에서는 집단지성을 활용함으로써 자연스럽게 데이터 품질 유지가 이루어지고 있음을 알 수 있다.

3.1.2 위키피디아의 구조적 특성

위키피디아의 구조적 특성은 리디렉션(redirects), 카테고리(category), 용어 간 상호 링크(mutual links)와 같이 크게 세 가지로 요약될 수 있다. 이러한 구조적 특성은 전문가에 의해 생성되는 것이 아니라, 이용자의 지속적 참여에 의한 수정과 갱신 등의 협업을 통해 유지·관리되는 것이 그 특징이라 할 수 있다.

(1) 리디렉션

일반적으로 시소러스는 문헌 집단 내에 존재하는 다양한 개인어(ideolects)와 전문용어를 연결하는 통제어휘로서의 역할을 수행한다. 이와 유사하게, 위키피디아에서는 리디렉션을 이용하여 대·소문자, 철자법, 약어, 동의어, 구어체, 학술용어 등의 다양한 이형에 대해 우선어(preferred term) 역할을 하는 문서의 제목으로 연결시켜주고 있다. 예를 들어, 이용자가 검색창에 'UK'를 입력했다면, 검색결과의 상단에 'redirected from UK'라는 표현과 함께 'United Kingdom'에 대한 문서를 볼 수 있도록 자동으로 연결시켜주고 있다.

또한, 위키피디아는 시소러스의 범위주기(scope note)와 같이, 이용자가 의도한 문서를 선정할 수 있도록 다양한 의미를 보여줌으로써 용어의 의미모호성 해소(disambiguation)를 돕고 있다.

(2) 카테고리

위키피디아에서 분류(categorization)란 이용

〈표 2〉 2009년 6월 현재 영문판 위키피디아 통계(Wikipedia 2009)

문 서		편 집		이용자	
문서페이지 수	2,921,449	페이지 편집횟수	314,633,151	등록이용자 수	9,922,754
위키 전체 페이지 수	17,149,376	페이지 당 평균 편집횟수	18.35	실제이용자 수 (최근 30일간)	155,524

〈표 3〉 위키피디아의 용어 의미모호성 해소 예시(book 항목)

- *book* is a set or collection of written, printed, illustrated, or blank sheets, made of paper, parchment, or other material, usually fastened together to hinge at one side.
- *Book(graph theory)*, a split graph consisting of p triangles sharing a common edge
- *Book(law school)*, a common award given by some law schools
- *Book(musical theatre)*, the spoken dialogue of a stage musical

자가 관련 주제에 대한 문서집합을 찾을 수 있도록 문서를 작성한 사람이 직접 문서를 카테고리에 부여하는 행위를 의미한다. 문서의 작성과 수정에 관여한 이용자들이 문서에 카테고리를 부여하고 카테고리 정보를 수정하는 등의 행위는 일종의 협력적 태깅으로 이해할 수 있다(Voss 2006). 그러나 기존의 태깅 시스템과는 달리, 위키피디아 카테고리 시스템은 계층적 형태를 띤다는 차이점이 있다.

위키피디아의 카테고리 시스템에서 한 카테고리는 다른 카테고리의 하위 카테고리가 될 수 있다. 즉, 상위 카테고리(supercategory)와 하위 카테고리(subcategory) 간의 계층관계를 인정함으로써 기본적으로 트리구조를 제공한다. 그러나 이러한 계층구조는 기존의 분류체계에서 제공하고 있는 것과 다른 점이 있다. 위키피디아에서는 하나의 카테고리가 두 개 이상의 상위 카테고리에 분류될 수 있기 때문에, 일반적인 분류체계에서 볼 수 있는 엄격한 계층구조가 아닌 일종의 방향성 비순환 그래프(directed acyclic graph)와 유사하다. 예를 들면, 카테고리 'music'은 'entertainment', 'performing arts', 'sound', 'arts', 'sound production'과 같은 5개의 상위 카테고리에 포함된다(Wikipedia 2009). 이러한 형태로 볼 때, 위키피디아의 카테고리 시스템은 개념의 트리구조에서 부모노드와 자식노드간의 일대일 관계만을 허용하는 분류체계보다는 시소러스 구조에 가깝다고 할 수 있다(Voss 2006).

(3) 용어 간 상호 링크

문서에 존재하는 하이퍼링크는 문서와 링크된 용어의 주제적 연관성을 나타낸다. 일반적으로 하이퍼링크 문서에는 두 가지 형태의 링크, 즉 단방향 링크와 상호 링크가 있는데, 이 중 상호 링크를 활용하면 연관성 높은 어휘를 얻을 수 있다. Ito et al.(2008)의 연구에서는 위키피디아 문서 내에 존재하는 링크의 동시출현 정보를 활용하여 링크 간 연관성을 계산하기 위해 다양한 통계적 분석을 시도하였는데, 용어 간 상호 링크가 연관성 높은 어휘를 추출하는 데 일정 수준 이상의 정확성을 보인 것으로 나타났다. 이러한 결과는 위키피디아 문서 내에 존재하는 링크가 용어 간의 주제적 연관성을 나타낼 수 있는 것으로 해석할 수 있다. 특히, 위키피디아에서는 문서작성자를 위한 링크 지침을 마련함으로써 링크의 품질을 유지하고 있는데, 일반적으로 많은 어휘보다는 연관된 핵심 용어에 링크를 표현할 것을 제안하고 있다.

(4) 시소러스와 위키피디아 구조의 비교

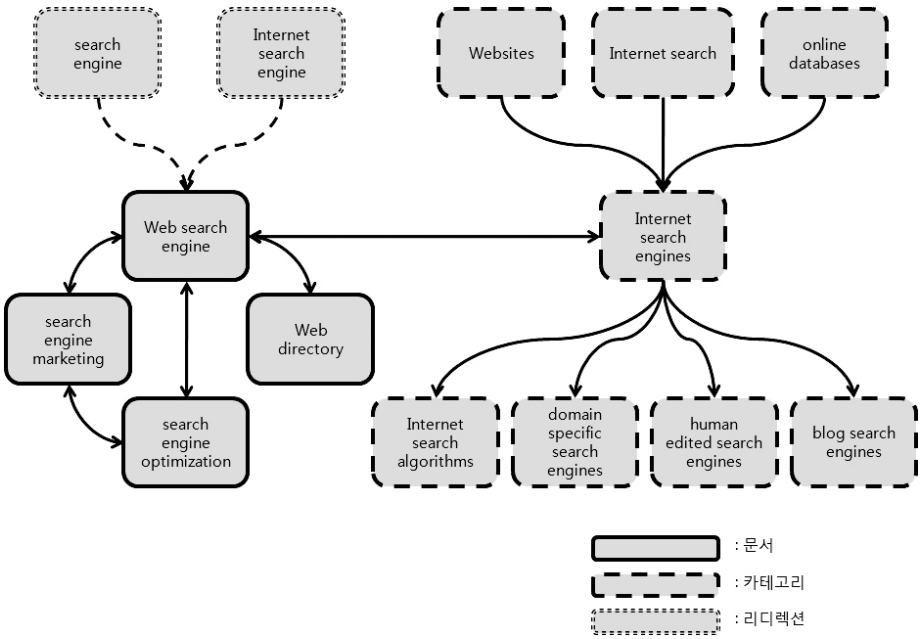
앞에서 언급한 바와 같이, 위키피디아의 구

조는 시소러스와 유사하다고 볼 수 있다. 위키피디아의 표제는 간명하고 잘 정의된 단어/구로 표현되어 있어 전통적인 시소러스의 용어와 비교될 수 있으며, 리디렉션, 카테고리 시스템, 하이퍼링크와 같은 구조가 기존의 시소러스에서 제공하고 있는 용어 간 관계와 유사하다. 〈표 4〉에서 보는 바와 같이, 위키피디아의 리디렉션은 시소러스의 동등관계, 카테고리 시스템은 계층관계, 그리고 용어 간 상호 링크는 연관관계로 매칭될 수 있다(Voss 2006).

이러한 내용을 근거로 하여 위키피디아의 'Web search engine' 항목 중 일부를 표현하면 〈그림 1〉과 같다. 'Web search engine'은 'search engine'과 'Internet search engine'등과 같은 리디렉션 페이지를 갖고 있으며, 이 항목의 카테고리에 해당하는 'Internet search engines'는 'Websites' 등과 같은 상위 카테고리를, 'Internet search algorithm' 등과 같은 하위 카테고리를 갖고 있다. 또한 'Web directory'와는 상호 링크로 연결된다. 이 그림으로 볼 때, 위키피디아는 시소러스와 유사한 구조로 표현됨을 확인할 수 있다.

〈표 4〉 시소러스와 위키피디아 구조 비교

관 계	관계기호	위키피디아 구조
동등관계	USE/UF	페이지 리디렉션
계층관계	BT/NT	상위 카테고리/하위 카테고리
연관관계	RT	용어 간 상호 링크



〈그림 1〉 위키피디아의 구조 예시(Web search engine 항목)

3.2 시소러스 갱신

시소러스의 갱신 과정은 시소러스 구축 과정과 거의 동일하다. 다만 구축에 투입되는 인력과 용어선정 과정에서 약간의 차이가 발생한다. 일반적인 시소러스 갱신을 위해 채택할 수 있는 방법은 기존 용어를 수정, 삭제하거나 신규 용어를 추가하는 것과 같이 용어를 대상으로 하는 방법과, 디스크립터에서 비디스크립터로 용어 간 관계를 수정하거나, 오래된 관계를 삭제, 또는 새로운 관계를 설정하는 것과 같이 의미관계를 대상으로 하는 방법이 있다(Aitchison, Gilchrist, and Bawden 2000).

Aitchison, Gilchrist, and Bawden(2000)은 위에서 제시한 방법 가운데 신규 용어를 추가하고 그에 적절한 관계를 부여하는 것이 가장 경제적으로 시소러스를 갱신할 수 있는 방법이라고 하였다. 시소러스의 갱신에 대한 경제성은 상대적으로 중요하게 고려되어야 한다. 왜냐하면, 시소러스 갱신의 범위를 너무 폭넓게 설정할 경우 시소러스 구축과 비슷한 규모의 인력, 시간 및 비용이 소요될 수 있기 때문이다.

전통적으로 시소러스의 갱신은 전문가 집단에 의해 일방적으로 이루어져 왔다. 그러나 이러한 유형의 갱신은 효율적이지 못한 결과를 낳게 된다. 이용자의 견해를 반영하여 시소러스를 갱신해야 하는 이유는 다음과 같은 두 가지 관점에서 설명할 수 있다. 첫째, 커뮤니케이션 관점에서, 시소러스는 저자의 언어를 이해하기 위한 색인자와 이용자의 커뮤니케이션 수단이라 할 수 있다. 그러므로 색인자와 이용자가 동시에 시소러스 갱신에 참여함으로써 저자와 이용자가 효율적으로 커뮤니케이션하는 것이 가능하다. 둘째, 언어학적 관점에서 볼 때, 언어는 늘 변화하는 불안정적인 객체이므로, 시소러스는 신조어 등과 같은 새로운 언어를 지속적으로 추가해주어야 한다(Kim 1973).

시소러스를 갱신하기 위해 신규 용어를 추출하는 방법에 대해 이용자 로그파일 분석(Schirmer 1967), 시소러스 통합(Kramer, Nicolai, and Habeck 1997), 의미 매핑(Doerr 2001), 단어 동시출현빈도에 근거한 연관어휘 추출(Curran and Moens 2002)과 동적 시소러스의 활용(이재윤 1994), 이용자 그룹 지식의 활용(장준국 2004), 단어연상검사에 의한 연상어휘 분석(한승희 2006), 서지 데이터를 이용한 용어 추출(Wang 2006) 등 다양한 방법이 연구되어 왔다. 이 중, 용어 간 동시출현빈도와 같은 통계나 확률정보를 이용하는 자동구축 시소러스는 구축 비용을 현저하게 감소시킬 수 있다는 장점이 있다. 그러나 개념 간 의미관계를 단순히 연관관계 정도로만 표현하고 있기 때문에 해당 분야의 관계정보를 충분히 반영할 수 없다는 문제점이 있다.

시소러스의 구축 및 갱신에 위키피디아를 활용한 연구에는 위키피디아의 링크에 대해 동시출현 정보를 분석하여 대규모의 연관 시소러스를 구축한 연구(Ito et al. 2008)와 DDC와 위키피디아의 구조를 비교 분석함으로써 위키피디아의 협력적 시소러스 태깅(collaborative thesaurus tagging) 구조를 소개한 연구(Voss 2006), Agrovoc 시소러스와 위키피디아의 용어를 비교 분석한 연구(Milne, Medelyan and Witten 2006) 등이 있다. 이들 연구 결과에서는 공통적으로 위키피디아가 시소러스 구축 및 갱신에 있어 효율적인 어휘자원으로 활용될 수

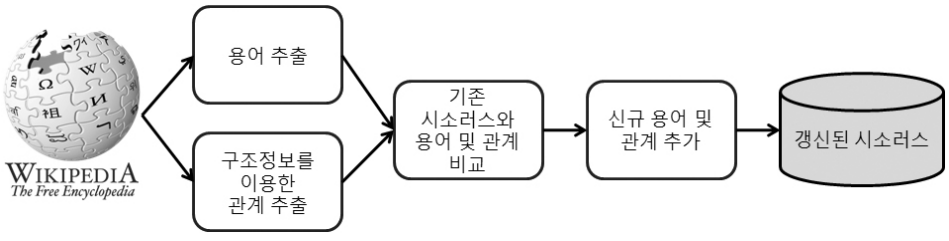
있음을 언급하였다. 즉, 해당분야의 텍스트, 전문용어사전, 백과사전 등과 같이 기존의 시소러스 구축 및 갱신에 필요한 용어를 획득하기 위한 외부자원과 비교해 볼 때, 위키피디아가 전문용어를 어느 정도 포함하고 있음을 증명하면서, 위키피디아가 시소러스 구축 및 갱신에 있어 효과적으로 활용될 수 있다는 점을 강조하였다.

3.3 위키피디아를 이용한 시소러스 갱신 실험

시소러스 갱신을 위한 어휘의 자동추출에 대한 대안적 접근법으로 협력적 태깅을 생각해볼 수 있다(Medelyan and Milne 2008). 협력적 태깅은 이용자 입장에서 특정 주제에 대한 다양한 개념을 얻을 수 있으며, 무료로 이용가능하고 전자적인 형태로 표현되어 있기 때문에 용어 추출이 쉽다. 협력적 태깅에는 다양한 사례가 있으나, 시소러스의 갱신을 위해 문서의 구조정보가 있는 위키피디아를 활용할 수 있다. 시소러스 갱신을 위한 플랫폼으로 위키피디아를 활용하는 것은 전통적인 방식으로 특정 주제영역의 시소러스를 구축하는 방식이 갖는 기본적인 장점을 가지고 있다. 즉, 비용적 측면에서의 경제성뿐만 아니라, 용어의 최신성을 유지할 수 있다(Milne, Medelyan, and Witten 2006).

이러한 관점의 타당성을 확인하기 위해 본 연구에서는 위키피디아를 활용하여 특정 주제영역의 시소러스를 갱신하는 실험을 수행하였다. 구체적으로는 ASIS&T 시소러스 제3판에서 20개의 디스크립터를 추출한 후 영문판 위키피디아를 활용하여 시소러스를 갱신하는 실험을 실시하였다. 실험과정은 〈그림 2〉와 같다. 먼저 추출된 20개의 디스크립터를 대상으로 위키피디아를 활용하여 연구자가 수작업으로 용어와 동등, 계층, 연관관계 등의 관계정보를 추출하였는데, 관계 추출에는 각각 위키피디아의 리디렉션, 카테고리, 용어 간 상호 링크를 이용하였다. 이후, 추출된 용어와 관계를 ASIS&T 시소러스와 비교하여 새롭게 추가된 용어와 관계를 확인한 후, 이를 ASIS&T 시소러스에 추가함으로써 시소러스 갱신작업을 완료하였다.

실험대상이 되는 20개의 디스크립터는 문헌정보학 분야에서 오랫동안 연구되고 있는 주제와 새롭게 연구되고 있는 주제를 골고루 반영하여 선정하였다. 왜냐하면, 위키피디아가 최신의 주제 혹은 인터넷 정보기술과 같은 특정 주제영역과 관련된 개념을 추가하는 데 유리한지 살펴보기 위해서이다.



〈그림 2〉 위키피디아를 활용한 시소러스 갱신 과정

4. 결과 분석

4.1 갱신 용어의 특성 분석

ASIS&T 시소러스와 위키피디아에서 추출한 시소러스(이하 위키 시소러스)를 대상으로 의미관계별 용어의 분포를 비교하면 〈표 5〉와 같다. 위키 시소러스의 디스크립터 별 평균 용어 수는 15.4개, ASIS&T 시소러스는 평균 6.65개의 용어를 포함하고 있는데, 이를 통해, 용어 포괄성의 측면에서 ASIS&T 시소러스에 비해 위키 시소러스가 우수하다는 것을 알 수 있다.

특히 위키 시소러스에는 연관관계 용어가 한 디스크립터 당 평균 9.35개, 전체 용어의 60.6%로, 다른 관계의 용어에 비해 많은 비율을 차지하고 있는데, 이것은 위키피디아 문서를 작성하거나 수정하는데 관여한 이용자의 다양한 지식구조를 반영하였기 때문에, 전문가 집단에 의해 관리되는 ASIS&T 시소러스에 비해 다양한 연관관계 용어를 포함할 수 있었던 것으로 분석된다. 동등관계의 경우에는 위키 시소러스(0.45)에 비해 ASIS&T 시소러스(1.05)가 평균적으로 더 많은 수의 용어를 포함하며, 비율 측면에서도 15.8%라는 높은 비율을 나타냈는데, 이것은 기존의 통제어휘가 동등관계 어휘를 효과적으로 잘 표현하고 있다는 것을 의미한다.

두 시소러스 간 용어 중복도를 측정한 결과, 〈표 6〉에서와 같이 평균적으로 0.9개의 용어만이 중복되어, 두 시소러스 간 용어 일치도가 매우 낮은 것으로 나타났다. 특히 NT의 경우에는 중복되는 용어가 20건 사례 중 단 하나도 없었다. 이러한 결과는 전문가 집단에 의해 생성된 통제어휘와 이용자 집단에 의해 생성된 용어 간 특성이 일치하지 않는다는 것을 의미한다. 기존의 통제어휘는 어휘의 제어 측면에 큰 관심을 두고 구축되어 왔는데, 시소러스가 정보검색에 적극적으로 활용되거나 이용자 친화적인 형태가 되려면 집단지성의 형태로 표현되는 실제 용어의 사용 패턴 등과 같은 영역지식(domain knowledge)을 효과적으로 반영할 필요가 있다.

〈표 5〉 ASIS&T 시소러스와 위키피디아의 의미관계별 용어 수 비교

비교 / 관계 / 디스크립터	ASIS&T 시소러스					위키 시소러스				
	UF	BT	NT	RT	합계	UF	BT	NT	RT	합계
archives	2	1	1	9	13	0	5	17	15	37
citation analysis	1	1	9	3	14	0	1	0	1	2
collection development	1	1	2	0	4	1	0	0	0	1
controlled vocabularies	0	1	7	6	14	0	6	0	16	22
digital libraries	2	1	0	2	5	1	11	14	23	49
faceted classification	0	1	0	6	7	0	2	0	5	7
information needs	2	0	0	6	8	0	2	0	1	3
information professionals	3	1	11	3	18	1	3	0	1	5
knowledge management	1	1	0	0	2	0	3	0	24	27

library management	1	1	1	1	4	0	0	0	3	3
MARC formats	1	1	0	1	3	1	2	0	5	8
OPACs	2	3	0	0	5	1	2	0	6	9
open access publications	0	1	0	2	3	1	5	0	9	15
public libraries	1	1	0	3	5	1	1	10	15	27
query expansion	0	1	0	0	1	0	2	0	1	3
relevance	0	0	0	10	10	0	1	0	1	2
usability	2	0	0	6	8	2	3	1	19	25
semantic Web	0	1	0	0	1	0	4	8	16	28
text mining	0	1	0	0	1	0	3	0	8	11
Zipf's law	2	1	0	4	7	0	6	0	18	24
합 계	21	19	31	62	133	9	62	50	187	308
평 균	1.05	0.95	1.55	3.10	6.65	0.45	3.10	2.50	9.35	15.40
비율(%)	15.8	14.3	23.3	46.6	100.0	3.2	20.0	16.1	60.6	100.0

〈표 6〉 의미관계별 중복된 용어의 수와 추가된 용어의 수 비교

항목 / 관계 / 디스크립터	중복된 용어 수					ASIS&T 시소러스에 추가된 용어 수				
	UF	BT	NT	RT	합계	UF	BT	NT	RT	합계
archives	0	0	0	4	4	0	5	17	11	33
citation analysis	0	1	0	0	1	0	0	0	1	1
collection development	0	0	0	0	0	1	0	0	0	1
controlled vocabularies	0	0	0	1	1	0	5	0	15	20
digital libraries	1	0	0	1	2	0	10	14	22	46
faceted classification	0	1	0	0	1	0	1	0	5	6
information needs	0	0	0	0	0	0	2	0	1	3
information professionals	1	0	0	0	1	0	3	0	1	4
knowledge management	0	0	0	0	0	0	3	0	24	27
library management	0	0	0	0	0	0	0	0	3	3
MARC formats	1	0	0	0	1	1	2	0	5	8
OPACs	1	0	0	0	1	0	2	0	7	9
open access publications	0	0	0	0	0	1	5	0	7	13
public libraries	0	1	0	0	1	1	0	10	15	26
query expansion	0	0	0	0	0	0	2	0	1	3
relevance	0	0	0	0	0	0	1	0	1	2
usability	0	0	0	1	1	2	3	1	18	24
semantic Web	0	1	0	0	1	0	3	8	16	27
text mining	0	0	0	0	0	0	3	0	8	11
Zipf's law	0	0	0	3	3	0	6	0	15	21
합 계	4	4	0	10	18	6	56	50	176	288
평 균	0.20	0.20	0.00	0.50	0.90	0.30	2.80	2.50	8.80	14.40
비율(%)	22.2	22.2	0.0	55.6	100.0	2.1	19.4	17.4	61.1	100.0

또한, 중복된 용어를 제외하고 디스크립터별로 ASIS&T 시소러스에 추가된 용어의 수는 평균 14.4개로 나타났다. 〈표 6〉에서 보는 바와 같이, 모든 디스크립터에 대해 용어가 추가되었으며, 'digital libraries(46)', 'archives(33)', 'knowledge management(27)', 'semantic Web(27)', 'public libraries(26)', 'usability(24)', 'Zipf's law(21)', 'controlled vocabularies(20)' 등과 같은 8개의 디스크립터에서는 20개 이상의 용어가 추가된 것을 확인하였다. 이 용어들은 'public libraries'와 'Zipf's law'를 제외하고는 정보기술 환경에서 널리 응용되는 개념을 나타내고 있다. 특히, 'digital libraries', 'archives', 'knowledge management', 'semantic Web' 등과 같은 개념은 문헌정보학 영역에서 비교적 최근에 발생한 연구 주제로, 예를 들면, 'digital libraries'의 경우에는 'library 2.0(NT)'과 같은 최신의 용어가 추가되었다. 또한 'semantic Web'과 같은 경우에는 ASIS&T 시소러스에 1개의 BT 용어(표 5 참조)만이 표현되어 있는 반면, 위키피디아를 통해 추가된 용어들을 살펴보면, 'folksonomy(NT)', 'Resource Description Framework(RT)' 등 시맨틱 웹의 개념을 설명하고 있는 여러 관련 개념들이 추가된 것을 볼 수 있었다.

또한, 'knowledge management', 'Zipf's law' 등과 같은 용어는 문헌정보학 이외에 각각 경영학과 수학과도 밀접한 관계를 가지고 있는데, 위키피디아를 통해 추가된 용어 역시 해당 분야의 용어를 포함하고 있어, 용어의 추가를 통해 해당 개념의 학제적 특성을 표현할 수 있음을 확인하였다.

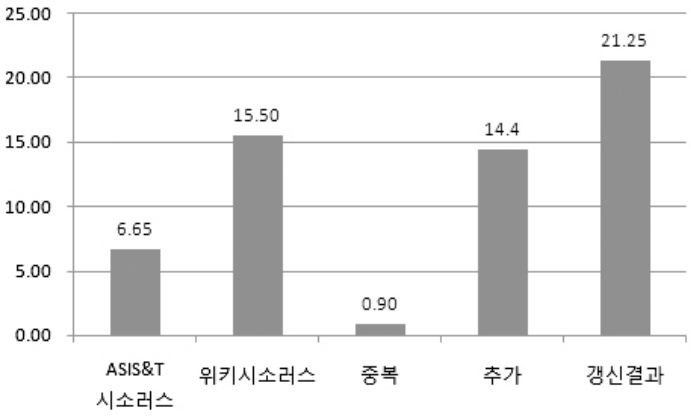
이외에, 'public libraries'는 문헌정보학에서 오래전부터 연구되고 있는 영역이나, 추가된 용어의 수가 많은 것으로 나타났는데, 이는 전 세계의 유명한 공공도서관 사례가 위키피디아 문서 내에 표현되어 관련 용어로 추가되었기 때문인 것으로 보인다. 또 다른 흥미로운 결과로, 'public libraries'와 'semantic Web', 'faceted classification' 등과 같은 일부 용어에 각각 'Andrew Carnegie'나 'Timothy John Berners-Lee', 'Shiyali Ramamrita Ranganathan' 등과 같은 해당 주제의 연관인물이 RT관계의 용어로 추가되었는데, 이것은 기존 시소러스에서 보기 어려운 새로운 유형의 용어로, 본 연구에서의 용어 추출 대상이 백과사전이라는 특성을 반영했기 때문인 것으로 분석할 수 있다.

〈표 7〉과 〈그림 3〉에서는 중복된 용어를 제외하고 추가된 ASIS&T 시소러스의 갱신 결과를 나타낸다. 표와 그림에서 알 수 있듯이, ASIS&T 시소러스의 갱신 전 디스크립터 당 평균 용어 수는 6.65개였으나, 위키피디아를 활용하여 갱신한 후의 평균 용어 수는 21.25개로, 디스크립터 당 평균 14.4개의 용어가 증가한 것을 알 수 있다. 이러한 결과는 〈표 5〉의 분석결과와 마찬가지로 용어 포괄성의 측면에서 ASIS&T 시소러스에 비해 위키 시소러스가 우수하다는 것을 보여준다.

또한, 용어의 의미관계별 비율의 변화를 살펴보면, UF를 제외한 나머지 관계에서 모두 비율이 증가한 것으로 나타났다. 특히 RT가 전체 용어의 56.5%, UF가 6.4%로 나타났는데, 이러한 현상은 기존의 시소러스가 UF 관계의 용어에 대해서는 충분한 포괄성을 갖고 있으나, RT 관계의 용어는 충분히 반영하지 못하고 있는 것으로 해석할 수 있다.

〈표 7〉 위키피디아를 활용한 ASIS&T 시소러스 갱신 결과

관계 / 디스크립터	UF	BT	NT	RT	합계
archives	2	6	18	20	46
citation analysis	1	1	9	4	15
collection development	2	1	2	0	5
controlled vocabularies	0	7	7	21	35
digital libraries	2	12	14	24	52
faceted classification	0	2	0	11	13
information needs	2	2	0	7	11
information professionals	3	4	11	4	22
knowledge management	1	4	0	24	29
library management	1	1	1	4	7
MARC formats	2	3	0	6	11
OPACs	2	5	0	7	14
open access publications	1	6	0	11	18
public libraries	2	1	10	18	31
query expansion	0	3	0	1	4
relevance	0	1	0	11	12
usability	4	3	1	24	32
semantic Web	0	4	8	16	28
text mining	0	4	0	8	12
Zipf's law	2	7	0	19	28
합계	27	77	81	240	425
평균	1.35	3.85	4.05	12.00	21.25
비율(%)	6.4	18.1	19.1	56.5	100.0



〈그림 3〉 평균 용어 수 비교

4.2 용어 및 관계의 적합성 평가

외부의 언어자원에 의해 시소러스가 갱신되는 경우에 반드시 고려해야 할 두 가지 사항이 있다. 하나는 추가된 용어가 해당 디스크립터에 적합한가의 여부이고, 또 다른 하나는 추가된 용어에 부여된 관계가 적합한가의 여부이다. 본 연구에서는 시소러스가 얼마나 효과적으로 갱신되었는가를 평가하기 위해 다음과 같은 두 가지 평가항목을 근거로 연구자가 직접 갱신된 시소러스를 평가하였다.

먼저, 용어 적합성(term relevance)은 추가된 용어가 해당 디스크립터에 적합한가의 여부를 평가하기 위한 것으로, 추가된 용어 중 디스크립터의 주제에 적합한 용어의 수의 비율로 계산되며, 이를 공식화하면 다음과 같다.

$$\text{용어적합성}(TR) = \frac{\text{디스크립터의 주제에 적합한 용어의 수}}{\text{추가된 용어의 수}}$$

또 다른 평가 척도인 관계 적합성(relation relevance)은 추가된 용어에 부여된 관계가 적합한가의 여부를 평가하기 위한 것으로, 디스크립터의 주제에 적합한 용어 중 디스크립터와의 관계가 적합한 용어의 수의 비율로 계산되며, 이를 공식화하면 다음과 같다.

$$\text{관계적합성}(RR) = \frac{\text{디스크립터와의 관계가 적합한 용어의 수}}{\text{디스크립터의 주제에 적합한 용어의 수}}$$

이러한 두 가지 평가 척도를 통해 갱신된 ASIS&T 시소러스의 평가 결과는 〈표 8〉과 같다. 전체 평균을 보았을 때, 용어 적합성과 관계 적합성의 비율은 각각 87.2%와 79.2%로, 이러한 결과는 ASIS&T 시소러스에 추가된 용어와 관계는 일정 수준 이상으로 적합하다는 것을 의미한다. 관계 적합성에 있어 UF는 100%의 적합성을 보여주었는데 이것은 두 가지 관점에서 해석이 가능하다. 하나는, UF의 경우 다른 의미관계에 비해 위키피디아를 통해 용어가 상대적으로 적게 추가가 되었기 때문에 의미관계가 잘못 부여될 확률이 적으며, 다른 하나는 위키피디아의 리디렉션이 동등관계 용어를 정확하게 제시해주고 있기 때문인 것으로 보인다.

이러한 결과는, 앞에서 언급한 연구들과 마찬가지로, 위키피디아가 시소러스 갱신에 적합한 외부 언어자원으로 활용될 수 있다는 것을 증명한다. 특히, 리디렉션, 카테고리, 용어 간 상호 링크로 요약되는 위키피디아의 구조적 특성은 시소러스의 의미관계를 유추하는 데 있어 적합한 것임을 확인할 수 있다.

5. 결론 및 제언

본 연구에서는 위키피디아를 활용하여 특정 주제영역의 시소러스를 갱신하는 실험을 수행하였다. 먼저, 추출된 20개의 디스크립터를 대상으로 위키피디아를 활용하여 용어와 의미관계를 추출한 후, ASIS&T 시소러스와 비교하여 새롭게 추가될 용어와 관계정보를 확인하였고, 이를 ASIS&T 시소러스에 추가함으로써 시소러스를 갱신하였다. 또한 시소러스가 얼마나 효과적으로 갱신되었는가를 평가하기 위해

〈표 8〉 ASIS&T 시소러스에 추가된 용어 및 관계 적합성 평가

항목 / 관계 / 디스크립터	용어 적합성					관계 적합성				
	UF	BT	NT	RT	합계	UF	BT	NT	RT	합계
archives	-	5	12	9	26	-	5	7	6	18
citation analysis	-	-	-	1	1	-	-	-	0	0
collection development	1	-	-	-	1	1	-	-	-	1
controlled vocabularies	-	4	-	15	19	-	3	-	12	15
digital libraries	-	9	13	21	43	-	4	9	19	32
faceted classification	-	1	-	5	6	-	1	-	4	5
information needs	-	2	-	1	3	-	2	-	1	3
information professionals	-	3	-	1	4	-	3	-	1	4
knowledge management	-	2	-	22	24	-	2	-	14	16
library management	-	-	-	3	3	-	-	-	2	2
MARC formats	1	2	-	5	8	1	2	-	4	7
OPACs	-	2	-	7	9	-	2	-	6	8
open access publications	1	4	-	7	12	1	2	-	5	8
public libraries	1	-	10	13	24	1	-	10	8	19
query expansion	-	2	-	0	2	-	2	-	0	2
relevance	-	1	-	1	2	-	1	-	0	1
usability	1	3	1	17	22	1	3	1	14	19
semantic Web	-	3	7	14	24	-	3	7	13	23
text mining	-	3	-	7	10	-	3	-	6	9
Zipf's law	-	4	-	4	8	-	4	-	3	7
합 계	5	50	43	153	251	5	42	34	118	199
비율(%)	83.3	89.3	86.0	86.9	87.2	100.0	84.0	79.1	77.1	79.3

('-' 표시는 해당하는 용어가 없어 적합성 판정이 이루어지지 않았음을 의미함.)

용어 적합성과 관계 적합성의 두 가지 평가 항목을 근거로 갱신된 시소러스를 평가함으로써, 시소러스 갱신에 있어 집단지성의 활용가능성에 대해 살펴보았다.

실험 결과, ASIS&T 시소러스와 위키 시소러스 간 용어 중복도는 매우 낮은 것으로 나타났으며, 모든 디스크립터에서 용어가 추가된 것으로 보아 용어 포괄성의 측면에서 ASIS&T 시소러스에 비해 위키 시소러스가 우수하다는 결론을 얻을 수 있었다. 특히, 디스크립터 중 문헌정보학 분야에서 비교적 새로운 주제 분야의 경우, 다른 용어에 비해 다양한 최신의 용어가 추가되었는데, 이것은 위키피디아가 주제의 최신성을 효과적으로 반영해주고 있다는 것을 의미한다. 또한 추가된 용어의 의미관계를 분석한 결과, 특히 RT관계의 용어가 두드러지게 높은 비율을 차지하고 있는 것으로 나타났는데, 이것은 다양한 배경지식을 가진 집단지성의 개입으로 인해 다양한 연관 개념이 시소러스에 추가될 수 있었기 때문인 것으로 해석할 수 있다. 이러한 결과를 통해, 위키피디아에 의해 시소러스가 갱신될 경우, 시소러스가 본질적으로 갖고 있는

어휘의 통제기능 이외에 탐색 보조도구로서의 기능이 더욱 강화될 것으로 기대할 수 있다.

또한, 갱신된 시소러스의 용어 적합성과 관계 적합성을 측정한 결과, ASIS&T 시소러스에 추가된 용어와 의미관계는 일정 수준 이상의 적합성을 보여주었는데, 이러한 결과는 곧, 위키피디아가 시소러스 갱신에 적합한 외부 언어자원으로 활용가능하다는 것을 증명한다. 특히, 리디렉션, 카테고리, 용어 간 상호 링크로 요약되는 위키피디아의 구조적 특성은 시소러스의 의미관계를 유추하는 데 있어 적합하다고 결론지을 수 있다.

시소러스 갱신에 있어 위키피디아의 활용가능성은 단일어 시소러스에만 머무르지 않는다. 주지하다시피, 위키피디아는 현재 세계 125개국의 언어로 제공되고 있다. 현재의 각 언어별 버전은 지속적으로 증가하고 있으며, 각 언어별 문서는 같은 개념을 다루고 있는 경우가 많다. 이러한 유의미한 중복은 다국어 시소러스 갱신을 위한 출발점이 된다. 그러므로 다국어 시소러스의 갱신을 위해 위키피디아를 활용하는 방안에 대해 관심을 가질 필요가 있다. 또한, 계층구조로 표현되지 않는 소셜 북마크나 협력적 태깅과 같은 집단지성 역시 풍부한 용어은행으로서의 기능을 수행할 수 있다. 그러므로 이들을 활용하여 관계정보를 추출하고 시소러스를 갱신할 수 있는 방안에 대한 연구가 필요하다. 마지막으로, 이 연구의 결과를 일반화하기 위해 더 많은 디스크립터와 다양한 시소러스를 대상으로 적용해보고, 갱신된 시소러스가 탐색 보조도구로서의 기능을 제대로 수행하고 있는지를 확인하기 위해 갱신된 시소러스를 대상으로 탐색성능에 대해 평가해 볼 필요가 있다.

참 고 문 헌

김지훈. 2002. 디지털 환경에서 시소러스의 동향과 연구과제. 『國會圖書館報』, 39(6): 3-27.

김지훈, 김태수. 2006. 용어정의와 관계추출을 통한 시소러스 확장에 관한 연구. 『한국문헌정보학회지』, 40(1): 293-314.

김학래, 최재화, 김홍기. 2006. 사회적 소프트웨어(Social Software)와 지식관리 프로세스. 『한국지능정보시스템학회 2006 춘계학술대회 논문집』, 316-323.

남영준, 이영주. 2002. 시소러스 구축과 유지보수 방안에 대한 연구. 『國會圖書館報』, 39(6): 49-68.

박재천, 신지웅. 2007. 웹2.0플랫폼에서의 집단지성 활용방안 연구: 교육분야에서의 적용을 중심으로. 『인터넷정보학회지』, 8(2): 15-20.

석영희. 2006. 『집단지성(Collective Intelligence)을 통한 정보생산의 사회적 의미』. 석사학위논문, 영남대학교 대학원 사회학과.

『위키백과』. 〈http://ko.wikipedia.org/〉.

이재윤. 1994. 『동적 시소러스 구축에 관한 실험

적 연구』. 석사학위논문, 연세대학교 대학원 문헌정보학과.

장준국. 2004.『사용자 그룹 지식을 활용한 온라인 시소러스 구축 시스템』. 석사학위논문, 고려대학교 교육대학원 컴퓨터공학전공.

제임스 서로위키 지음. 홍대운, 이창근 옮김. 2005.『대중의 지혜: 시장과 사회를 움직이는 힘』. 서울: 랜덤하우스.

최석두. 2000. 시소러스 標準開發에 대한 硏究.『지식처리연구』, 1(2): 1-38.

피에르 레비 지음. 권수경 옮김. 2002.『집단지성: 사이버 공간의 인류학을 위하여』. 서울: 문학과지성사.

한승희. 2006. 단어연상검사법을 이용한 탐색 시소러스 구축에 관한 실험적 연구.『한국문헌정보학회지』, 40(3): 289-304.

Aitchison, Jean, Alan Gilchrist, and David Bowden. 2000. *Thesaurus Construction and Use: A Practical Manual*, 4th ed. Chicago: Fitzroy Dearborn Publishers.

Curran, James R. and Marc Moens. 2002. "Improvements in automatic thesaurus extraction." *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, 9: 59-66.

Doerr, M. 2001. "Semantic Problems of Thesaurus Mapping." *Journal of Digital Information* [online], 1(8). [cited 2009. 4.7]. 〈http://journals.tdl.org/jodi/article/view/31/32〉.

『Enterprise 2.0 기반의 지식 경영 시스템 - 집단 지성 활용』. 2007. In *Mimul's Developer World* [personal blog]. 2007년 11월 17일. [cited 2009.5.13]. 〈http://www.mimul.com/pebble/default/2007/11/17/1195290240000.html〉.

Fischer, Gerhard. 2003. "Designing Social Networks in Support of Social Creativity." Position Paper for the Workshop on *Moving from Analysis to Design: Social Networks in the CSCW Context*, ECSCW '03: 14-18.

Heylighen, Francis. 1999. "Collective Intelligence and its Implementation on the Web: Algorithms to Develop a Collective Mental Map." *Computational & Mathematical Organization Theory*, 5(3): 253-280.

Ito, Masahiro, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2008. "Association Thesaurus Construction Methods based on Link Co-occurrence Analysis for Wikipedia." *Proceeding of the 17th ACM conference on Information and knowledge management*, 817-826.

Kim, Chai. 1973. "Theoretical Foundations of Thesaurus-Construction and Some Methodological Considerations for Thesaurus Updating." *Journal of the American Society for Information Science*, 24(2): 148-156.

Kramer, R., R. Nicholai, and C. Habeck. 1997 "Thesaurus Federations: Loosely Integrated Thesauri for Document Re-

trieval in Networks Based on Internet Technologies." *International Journal of Digital Libraries*, 1: 122-131.

Medelyan, O. and D. Milne. 2008. Augmenting Domain-specific Thesauri with Knowledge from Wikipedia. In *Proceedings of the NZ Computer Science Research Student Conference(NZCSRSC 2008)*, Christchurch, New Zealand. [online]. [cited 2009.5.2]. 〈http://www.cs.waikato.ac.nz/~dnk2/publications/NZCSRSC08-AugmentingDomainS pecificThesauri.pdf〉.

Milne, D., O. Medelyan, and I. H. Witten. 2006. Mining Domain-Specific Thesauri from Wikipedia: A case study. In *Proceedings of the International Conference on Web Intelligence(IEEE/WIC/ACM WI 2006)*, Hong Kong. [online]. [cited 2009.4.30]. 〈http://www.cs.waikato.ac.nz/~dnk2/publications/WI06-MiningDomainS pecificThesauriFromWikipedia.pdf〉.

Noubel, Jean-François. 2007. *Collective Intelligence, The Invisible Reolution*. [pdf file]. [cited 2009.5.8]. 〈http://www.thetransitioner.org/Collective_Intelligence_Invisible_Revolution_JFNoubel.pdf〉.

Redmond-Neal, Alice and Marjorie M. K. Hlava ed. 2005. *ASIS&T Thesaurus of Information Science, Technology and Librarianship*. New Jersey: Information Today, Inc.

Schirmer, Robert F. 1967. "Thesaurus Analysis for Updating." *Journal of Chemical Documentation*, 7(2): 94-98.

Voss, Jakob. 2006. "Collaborative Thesaurus Tagging the Wikipedia Way." *Wikimetrics* [online], 1(1). 〈http://arxiv.org/abs/cs.IR/0604036〉

Wang, Jun. 2006. "Automatic Thesaurus Development: Term Extraction from Title Metadata." *Journal of the American Society for Information Science and Technology* 57(7): 907-920.

Wikipedia. 〈http://www.wikipedia.org/〉.