새로운 주제 탐지를 통한 지식 구조 갱신에 관한 연구*

A Study on Updating the Knowledge Structure Using New Topic Detection Methods

김 판 준(Pan-Jun Kim)**
정 영 미(Young-Mee Chung)***

초 록

새로운 주제의 탐지를 위한 여러 접근법들을 지식 구조 표현 방법 중 하나인 디스크립터의 부여 및 갱신 과정에 적용하였다. 새로운 주제 탐지는, 특히 특정 학문 분야에서 새로운 주제의 출현 및 성장으로 인하여 지식 구조상의 변화가 발생하는 경우에, 기존의 색인어로는 이를 표현할 수 없거나 표현상의 제한이 따르는 문제를 해결하는 데 응용할 수 있다. 실험 결과, 정보학 내에서 긍정적 측면의 변화가 발생한 것으로 식별된 신흥 주제들은 상당수가 서로 밀접하게 연관되어 있으면서 동시에 성장·발전의 단계에 있는 주제임을 확인하였다. 또한, 새로운 주제 탐지를 통한 후보 디스크립터 리스트의 사용이 색인자의 색인작업을 지원하는 효율적인 도구가 될 수 있다는 가능성을 보여 주었다. 특히, 적절한 디스크립터의 선정과 부여를 위한 후보 디스크립터 리스트의 제공은 색인작업의 효율성과 정확성을 향상시키는 데 기여할 수 있을 것이다.

ABSTRACT

This study utilizes various approaches for new topic detection in the process of assigning and updating descriptors, which is a representation method of the knowledge structure. Particularly in the case of occurring changes on the knowledge structure due to the appearance and development of new topics in specific study areas, new topic detection can be applied to solving the impossibility or limitation of the existing index terms in representing subject concepts. This study confirms that the majority of newly developing topics in information science are closely associated with each other and are simultaneously in the phase of growth and development. Also, this study shows the possibility that the use of candidate descriptor lists generated by new topic detection methods can be an effective tool in assisting indexers. In particular, the provision of candidate descriptor lists to help assignment of appropriate descriptors will contribute to the improvement of the effectiveness and accuracy of indexing.

키워드: 디스크립터, 지식 구조 갱신, 새로운 주제 탐지, 새것 탐지, 신성 주제 탐지, 색인 지원 도구
descriptors, knowledge structure updating, new topic detection, novelty detection, emerging trend detection, indexing aid

---

* 본 연구는 연세대학교 대학원 박사학위논문의 일부를 요약한 것임.

** 연세대학교 대학원 문헌정보학과(dpbluesea@hanmail.net)
*** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr)

■ 논문접수일자 : 2005년 2월 15일
■ 게재확정일자 : 2005년 2월 23일

1. 서 론

최근 여러 학문 분야에서 '새것 탐지'(novelty detection) 혹은 '신성 주제 탐지'(emerging trend detection)에 관련된 여러 접근법이 다양한 목적을 위해 활발하게 응용되고 있다. 예를 들면, 인공지능, 의학, 생물 정보학, 로봇공학 등 여러 분야에서 이와 관련된 기법들을 응용하여 기계학습, 질병 진단, 유전자 분석, 로봇 제어 등 특정 과제를 효율적으로 수행하려는 시도가 이루어지고 있다.

정보학 분야에서도 2002년 TREC-11에서 '새것 탐지' 트랙을 새로이 신설하고, TREC에서 제공하는 질의(topic)와 적합문헌 리스트를 사용하여 적합문장(relevant sentence)과 새로운 문장(novel sentence)을 찾는 공통의 기본 과제를 13개 연구 집단에서 개별적으로 수행한 바 있다. 새것 탐지 혹은 신성 주제 탐지 방법은 이와 같이 문헌정보학을 비롯한 여러 학문 분야에서 다양한 과제를 효율적으로 수행하기 위한 새로운 방법론으로 상당한 가능성을 보여주고 있다.

이 연구에서는 새로운 주제의 탐지를 위한 여러 접근법들을 지식 구조 표현 방법 중 하나인 디스크립터 부여 및 갱신 과정에 적용하였다. 다시 말해서, 특정 학문 분야에서 새로운 주제의 출현 및 성장으로 인하여 지식 구조상의 변화가 발생하는 경우에, 기존의 색인어로는 이를 표현할 수 없거나 표현상의 제한이 따르는 문제를 해결하기 위한 방법으로 '새로운 주제 탐지'(new topic detection) 방법을 응용하였다. 따라서 이 연구의 목적은 특정 분야의 주제를 표현하는 용어들로 구성된 지식 구조에 대해, 새로운 주제 탐지 방법을 적용함으로써 일정 기간 동안 발생한 변화를 탐지하여 현재의 지식 구조를 보다 정확하게 표현하는 데 있다.

새로운 주제 탐지를 통하여 갱신된 지식 구조는 이용자의 탐색과정에서 보다 다양한 접근점을 제공할 수 있을 것이다. 또한, 색인자는 시간의 경과에 따라 변화하는 주제의 양상을 보다 구체적으로 파악하여 적절한 디스크립터를 부여할 수 있으며, 색인작업의 정확성과 효율성을 동시에 향상시킬 수 있는 효과를 기대할 수 있다. 즉, 시스템이 제공하는 구체적인 데이터와 이를 참고한 색인자의 지적 판단의 결과로 색인자가 새로 입력된 문헌에 디스크립터를 부여하는 바로 그 시점에 문헌 데이터베이스를 통해 표현된 지식 구조가 실시간으로 적절하게 갱신되는 효과를 함께 얻을 수 있다.

특정 학문 분야의 문헌에 부여한 디스크립터와 이들 간의 관계, 그리고 디스크립터가 부여된 문헌에 출현한 자연언어 키워드와 디스크립터 간의 관계는 그 자체가 주제 분야에 대한 개념적인 지식 구조를 형성한다고 할 수 있다. 이러한 전제를 바탕으로 각 문헌에 포함된 자연언어 키워드와 디스크립터의 출현 정보 및 동시출현 정보에 기반한 여러 행렬 즉, 디스크립터-키워드 행렬, 디스크립터-디스크립터 행렬, 키워드-키워드 행렬을 생성하여 일종의 지식 구조를 구성하였다. 다음으로, 이러한 지식 구조를 대상으로 새로운 주제 탐지와 이를 통한 지식 구조의 갱신을 위한 다양한 방법을 적용하였다.

2. 새로운 주제 탐지

새로운 주제 탐지를 위한 접근법으로는 크게 새것 탐지와 신성 주제 탐지가 있다. 이 중 새것 탐지는 주로 현재의 시점에서 과거의 데이터를 학습하여 구조화하고, 입력되는 데이터에서 가능한 모든 새로운 정보를 찾는다. 반면, 신성 주제 탐지는 고정된 하나의 시점이 아닌 연속적인 여러 시간대의 데이터에 기반하여 시간의 흐름에 따른 변화를 파악하고, 그 결과 문자 그대로의 새로운 주제인 '신생 주제'(newly occurring topic)를 비롯하여 기존 주제의 발전 및 재조명에 의한 '신흥 주제'(newly developing topic) 등 다양한 유형의 새로운 주제를 세부적으로 식별한다. 따라서 특정 학문 분야의 지적 구조상의 변화 및 발전 상황을 실제적으로 파악할 수 있는 최선의 방법으로는 새것 탐지와 신성 주제 탐지 접근법 양자를 함께 고려하여야 할 것이다.

2. 1 새것 탐지

새것 탐지는 기계학습 시스템이 학습하는 동안 인식하지 못한 새롭거나 알려지지 않은 데이터 혹은 신호를 식별하는 것이며, 응용분야로는 신호 처리, 컴퓨터 시각, 패턴 인식, 데이터 마이닝, 로봇공학(robotics) 등이 있다(Markou and Singh 2003). 새것 탐지는 학문 분야에 따라 여러 가지 다른 용어로 지칭하기도 하는데, 로봇제어나 컴퓨터 보안 혹은 의학 진단 등의 분야에서는 '이상 탐지'(anomaly detection), 기계 감독(machine monitoring) 분야에서는 '고장 탐지'(fault detection), 통계학 분야에서는 '이상치 탐지'(outlier detection)라고 하기도 한다. 본 연구에서 '새것 탐지'는 특정 시점을 기준으로 이전 시기의 주제와 비교하여 최근 시기의 주제에 발생한 성장·발전 등 긍정적인 변화를 탐지하기 위한 목적으로 사용된다. 이러한 변화는 특정 주제 분야에서 새로운 정보의 유입과 성장에 따라 발생하는 것으로 지식 구조 갱신의 필요성에 대한 단서를 제공할 수 있다.

대부분의 새것 탐지는 일반적이거나 정상적인 데이터의 사례들만으로 학습된다. 그런 다음 어떤 입력물이 획득된 모델과 맞지 않는 경우 즉, 새로운 범주에 속하는 데이터(members of the novel class)를 인식하는 것이다. 새것 탐지를 위한 사례들이 불충분한 분야에서는 새것이 아닌데 새것으로 판정하는 것(false positive)보다 새것을 새것이 아닌 것으로 인식하는 것(false negative)이 더 심각한 결과를 초래할 수 있다(Marsland 2003). 지금까지 새것 탐지에 대한 가장 일반적인 접근법은 식별하고자 하는 범주에 속한 어떤 사례도 포함하고 있지 않은 학습집단을 준비하고, 다음으로 이러한 데이터집합의 일반적인 패턴을 학습하는 특정한 학습 시스템을 이용하는 것이다. 새것 탐지는 학습이 이루어진 후에 입력되는 검증집단에 대하여 학습된 모형과 검증집단의 입력물을 비교함으로써 평가될 수 있다.

새것 탐지를 위한 접근법은 크게 통계적 접근법과 신경망 접근법으로 구분할 수 있다. 첫째, 통계적 접근법은 대부분 통계적 특성에 기초한 데이터의 모델링에 기초하고 있으며, 이러한 정보를 검증 사례들이 동일한 분포에서 나온 것인지를 추정하는 데 사용한다(Markou

and Singh 2003). 둘째, 신경망 접근법은 데이터를 2개 이상의 범주로 클러스터링하는 범주화 문제로 간주할 수 있는데, 크게 지도학습(supervised learning)과 비지도학습(unsupervised learning)의 두 가지 방법으로 학습이 이루어진다.

이러한 새것 탐지는 특히 다른 시스템들이 검토하여야 할 입력물의 수를 감소하는 데 사용될 수 있다. 즉, 학습시스템이 이전에 보여지지 않은 혹은 아주 드물게 보여진 사례에만 주의를 집중할 수 있게 하는 사전처리 방법으로 유용하게 사용될 수 있는 것이다. 예를 들면, 새것 탐지 접근법을 이용하여 신경망은 이전에 보지 않은 데이터에 대해서만 학습할 수 있고, 로봇은 이전에 경험하지 못한 입력 자극에만 반응할 수 있다(Marsland 2003).

2. 2 신성 주제 탐지

신성(新成)[1] 주제는 시간이 경과함에 따라 관심과 유용성(utility)이 높아지고 있는 연구 주제를 말한다(Kontostathis et al. 2003). 신성 주제에 대한 탐지는 중장기적 측면에서 시간의 흐름에 따른 변화의 양상을 파악하는 것으로 특정 학문 분야의 경향 혹은 동향에 대한 탐지라고 할 수 있다. 이러한 신성 주제 탐지는 변화의 양상에 따라 각 주제를 신생 주제, 신흥 주제, 기존 주제로 구분할 수 있는 기본적인 틀을 제공하여, 주제 유형별로 지식 구조 갱신에 적용할 수 있는 기초가 된다. 또한, 신성 주제는 크게 '신생(新生) 주제'와 '신흥(新興) 주제'의 두 가지로 구분할 수 있는데, 신생 주제는 이전에는 없었던 어떤 주제가 새롭게 출현한 경우를 말하고, 신흥 주제는 주로 기존에 있던 주제가 성장, 발전 혹은 새롭게 부상한 경우를 말한다.

최근 신성 주제 탐지에 관한 여러 이론적 연구와 상업적인 시스템들을 소개한 논문에서는 신성 주제의 대표적인 예로서 1990년대 중반 새로운 연구 주제로 부상한 XML(eXtensible Markup Language)을 들고 있다. INSPEC 데이터베이스를 대상으로 키워드 탐색을 수행한 결과에 따르면, XML은 1994년 이전에는 검색 결과가 전혀 없다가 1994년부터 1996년 사이에 새롭게 출현하였고 1998년에 이르러 하나의 독립적인 주제로 확립되었다(Kontostathis et al. 2003).

신성 주제 탐지는 일반적으로 탐지의 첫 단계인 입력에서 이용자의 개입 여부에 따라 크게 자동 신성 주제 탐지와 반자동 신성 주제 탐지로 구분할 수 있다. 자동 신성 탐지의 경우에는 이용자에 의한 초기 질의 없이 컴퓨터에 저장된 문헌집단에 기초하여 신성 주제 리스트를 자동 생성하고, 이를 시스템에 의해 발견된 기반 정보와 함께 최종 판정을 맡은 전문가에게 제시한다. 반자동 신성 주제 탐지의 경우에는 시스템에 대한 이용자의 입력 질의에 기초하여 작성된 신성 주제 리스트와 기반 증거를 이용자 친화적인(user-friendly) 인터페이스로 제공한다.

신성 주제 탐지의 수행을 위한 실험집단의

---

1) 신성(新成) 주제: '새롭게 이루(어지)다'의 의미로 특정한 주제의 출현, 성장, 발전, 쇠퇴 등의 변화로 인하여 새롭게 이루어진 주제를 표현하기 위해 본 논문에서 사용되는 용어이다. 한자어 '성(成)'은 영어로는 'accomplish'를 의미하는 것으로 여러 가지 변화의 과정 혹은 결과에 의해 이루어진 상태로 해석할 수 있다. 예를 들면, '성가(成家)'는 '학문이나 기술로 한 체계를 이루다'는 의미로 사용된다.

특성에 따라 자연언어(전문: full-text), 통제언어(디스크립터), 그리고 자연언어와 통제언어 양자를 대상으로 하는 경우로 구분할 수도 있다. 또한, 실험집단의 출처에 따라 기존의 미국과학정보연구소(ISI) 데이터베이스 혹은 TDT (Topic Detection and Tracking) 실험집단을 이용하거나 연구자가 스스로 새로운 실험집단을 구성하는 경우로 구분할 수도 있다. 주의할 점은 특히 자연언어를 대상으로 신성 주제를 탐지할 경우에는 최초로 출현하거나 생성된 단어에 대해서는 우선적인 고려대상으로 삼아야 한다는 것이다. 또한, TDT 실험집단은 자연언어인 기사 전문과 통제언어인 기사 디스크립터로 구성되어 있어 자연언어와 통제언어를 함께 이용할 수 있지만, 새로운 단어의 탐지가 목적일 경우 TDT의 범주가 너무 제한적인 디스크립터로 구성되어 있는 등의 문제를 고려하여야 한다.

신성 주제 탐지의 기본 가설은 컴퓨터 알고리즘이 시간의 경과에 따른 개념 빈도와 개념 사이의 연관성의 변화를 추적함으로써 새로운 정보를 자동적으로 탐지할 수 있다는 것이다 (Pottenger and Yang 2001). 이를 전제로 신성 개념 혹은 주제는 다음과 같은 두 가지 주요 특성을 가진다(Kontostathis et al. 2003).

- 특정 시점의 이전 보다 이후에 의미적으로 더 풍부해진다. 즉, 신성 주제와 동시출현하는 디스크립터와 키워드의 수가 증가한다.
- 해당 개념 혹은 주제와 연관된 항목(문헌)의 수가 증가함에 따라 출현 빈도가 더 높아진다. 다시 말해서, 관련 디스크립터와 키워드 양자의 출현빈도가 증가한다.

3. 새로운 주제 탐지 실험

3. 1 실험 설계

정보학 분야의 개념들과 개념들 간의 관계로 구성된 지식구조를 전제로 하여, 색인 전문가가 학문의 변화 및 발전 양상을 실질적으로 반영할 수 있는 적절한 디스크립터를 색인어로 부여함으로써 기존 지식구조를 적절하게 갱신할 수 있도록 지원하는 방안을 제시하는 것이 본 실험의 목적이다. 이를 위해 색인자가 지식구조를 구성하는 기본적인 요소인 디스크립터를 적절하게 부여 혹은 갱신하는 데 사용될 수 있도록, 새로운 주제 탐지 방법을 응용하여 후보 디스크립터 리스트를 생성하였다.

새로운 주제 탐지 실험을 위한 문헌집단은 LISA에서 정보학 분야의 3개 핵심 저널(JASIST, IPM, JIS)에 수록된 총 16년(1989년~2004년) 동안의 논문을 대상으로 구성하였다. 실험을 위한 기본적인 처리 대상 및 분석 단위는 레코드에 수록된 개별 논문을 대표하는 키워드와 디스크립터이다. 따라서 새로운 주제 탐지의 첫 단계는 데이터베이스에서 표제와 초록 필드를 대상으로 키워드를 추출하고 디스크립터 필드를 대상으로 디스크립터를 추출하는 사전 처리로 시작된다. 즉, 자동색인과 유사한 방법으로 자연언어 색인어인 키워드를 추출하고, 전문가에 의해 주어진 통제언어 색인어인 디스크립터를 추출한 다음, 양자를 기반으로 이후의 실험을 위한 용어들을 선정한다. 이러한 용어들의 출현 시기, 출현 빈도, 동시출현 빈도 등은 이후의 처리를 위한 기반 정보로서 각 용어의 속성이면서 동시에 자질 선정의 기

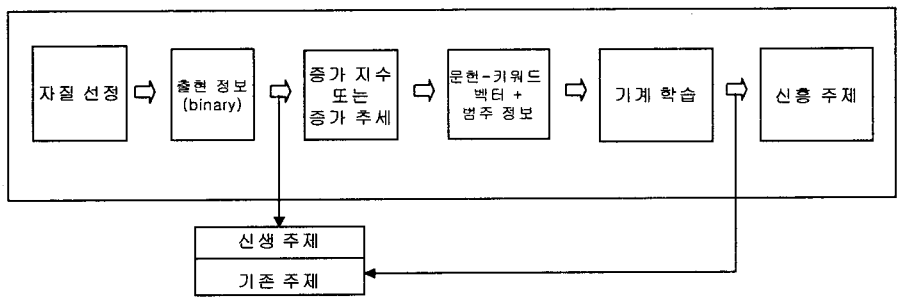
준이 된다.

지금까지 새것 탐지나 신성 주제 탐지에서 새로운 주제 혹은 신흥 주제를 탐지하기 위한 기반 정보로 주로 사용되어 온 것은 주제를 표현하는 용어들의 출현 정보였고, 최근에 와서야 동시출현 정보를 함께 사용하는 시도가 일부 이루어지고 있다. 또한, 기존의 새것 탐지나 신성 주제 탐지에서는 새로 출현한 주제와 기존의 주제 중에서 새롭게 성장하고 있는 주제를 특별히 구분하지 않거나 두 가지 유형의 주제 중 하나에만 초점을 맞추고 있는 경우가 많다. 그러나 본 연구에서는 용어의 출현 정보와 동시출현 정보를 함께 고려하여 새로운 주제의 탐지를 위한 기본적인 자질을 선정하고, 새로운 주제의 유형을 새로 출현한 주제인 신생 주제와 기존 주제이면서 최근 새롭게 부상하고 있는 주제인 신흥 주제를 명백하게 구분하여 식별한다는 점에서 이전의 연구들과는 차별화된다.

본 연구에서 새로운 주제 탐지 과정은 크게 두 단계(1차, 2차)로 나뉘어진다. 1차 새로운 주제 탐지는 각 주제를 세 가지 유형으로 분류하는 것으로, 〈그림 1〉은 이를 위한 흐름도이다. 먼저, 신생 주제의 탐지는 디스크립터의 시기별 출현 정보에 따라 이루어진다. 즉, 이전(1989년~2000년)에는 출현하지 않았다가 최근(2001년~2004년)에 새롭게 출현한 디스크립터는 모두 신생 주제로 추출하였고, 이전에 출현하였지만 최근에는 출현하지 않은 디스크립터는 모두 기존 주제로 분류하였다.

다음으로, 신흥 주제는 이전에도 출현하였고 최근에도 출현한 디스크립터들로부터 출현 정보와 동시출현 정보에 기초하여 신흥 주제 후보 디스크립터들을 선정한 다음, 이들을 대상으로 기계학습 알고리즘에 의한 범주화를 수행하여 추출하였다.

신흥 주제가 되기 위한 디스크립터의 기본적인 조건은 증가 지수가 기준치 이상이거나($I \geq 0$) 선형회귀선의 기울기 값이 기준치 보다 높아야($rs > 0.2$) 한다. 증가 지수는 해당 디스크립터가 문헌에 부여된 빈도(문헌빈도)를 기반으로 이전의 상위 4개년(top 4 year: 1989년~2000년)의 빈도와 최근 4개년(2001년~2004년)의 출현빈도에 기초하여 산출하였고, 선형회귀선의 기울기는 전체 16년(1989년~2004년) 간의 연도별 출현빈도에 기초하여 산출하였다. 여기서 전자는 비교된 두 시기 사이의 증가량을 고려한 것이라 할 수 있고 후자는 전반적인 증가 추세를 반영하는 것이라 할 수 있다.



〈그림 1〉 1차 새로운 주제 탐지의 흐름도

신흥 주제의 식별을 위한 기계 학습 알고리즘의 적용은 각 디스크립터가 특정 학문 분야의 지식 구조를 구성하고 있는 주제를 대표하고 있다는 전제 하에, 이전 문헌집단을 이용하여 해당 디스크립터로의 범주화를 학습한 분류기의 성능이 최근 문헌집단을 이용한 범주화 검증 단계에서 크게 낮아진 경우에는 해당 주제에 최근 어떤 변화가 발생하였다고 가정하는 것이다. 신흥 주제를 식별하기 위한 분류기로는 새것 탐지 선행 연구에서 주로 사용되어온 나이브 베이즈 알고리즘을 사용하였다.

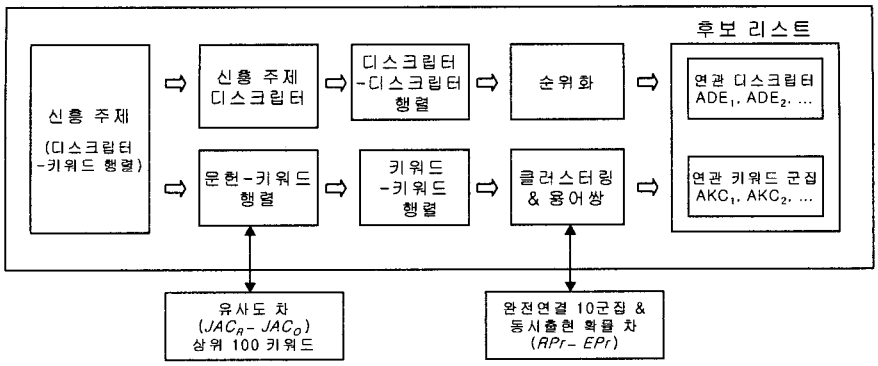
기계학습 알고리즘의 입력물은 각 문헌집단의 문헌과 문헌에 출현한 키워드들로 이루어진 문헌 벡터에 해당 디스크립터의 부여 여부를 구분하는 범주 정보(yes, no)를 추가하여 구성한 문헌-키워드 행렬이 된다. 이에 따라 이전(1989년~2000년) 문헌집단을 학습집단(2,042 개 문헌)으로 사용하여 학습을 수행하고 최근(2001년~2004년) 문헌집단을 검증집단(652개 문헌)으로 사용하여 범주화한 결과, 범주화 성능이 기준치(재현율 0.5) 이하인 디스크립터를 최종적으로 신흥 주제로 추출하였다. 이렇게 한 이유는 해당 디스크립터로 범주화되는 성능의 완전성을 의미하는 재현율(recall)이 50%가 되지 않으면, 최근에 디스크립터가 대표하고 있는 주제에 어떤 변화가 발생하여 검증집단(최근집단)의 문헌 중 절반 이상이 해당 범주로 적절하게 범주화되지 않은 것으로 생각할 수 있기 때문이다. 따라서 일차적인 새로운 주제 탐지의 결과는 전체 디스크립터를 기존 주제, 신생 주제, 신흥 주제로 분류하는 것으로 나타난다.

데이터베이스에 새로운 문헌이 입력되면 1차 새로운 주제 탐지를 통해 세 가지 유형의 주제로 구분된 1차 후보 디스크립터와 관련 정보(동시출현 키워드, 증가 지수)가 유사도순으로 제공될 수 있다. 이 중에서 기존 주제와 신생 주제에 해당하는 디스크립터의 경우에는 색인자가 해당 디스크립터를 문헌에 부여하고 지식 구조를 구성하는 각 행렬에 관련 정보를 추가하는 것으로 갱신이 완료된다. 그러나 신흥 주제의 경우에는 기존의 디스크립터가 해당 분야의 변화를 반영하여 주제를 적절하게 표현하지 못하고 있다는 관점을 반영한 디스크립터의 선정 및 부여 절차가 필요하다. 즉, 새로운 주제 탐지의 결과 식별된 신흥 주제는 기계학습을 통하여 기존의 주제 분야에 어떤 주목할 만한 변화가 발생하여 현재의 디스크립터로는 표현이 부족하거나 부적절한 주제라고 할 수 있으며, 이러한 변화를 반영할 수 있는 추가적인 처리가 요구된다.

2차 새로운 주제 탐지는 이전 단계에서 식별된 신흥 주제에 대한 연관 디스크립터의 추출과 새로운 후보 디스크립터의 도출을 위한 연관 키워드 군집을 생성하는 것으로, 이를 위한 흐름도는 〈그림 2〉와 같다.

먼저, 신흥 주제의 연관 디스크립터는 디스크립터-디스크립터 동시출현 행렬에 기초하여 추출하는 것으로 문헌에 신흥 주제 디스크립터와 동시에 부여된 빈도가 높은 디스크립터를 말한다. 색인자는 문헌에 출현한 키워드 리스트와 연관 디스크립터와 동시출현한 키워드 리스트를 비교하여 신흥 주제에 적절한 연관 디스크립터를 추가 혹은 대체할 수 있다.

다음으로 연관 디스크립터 리스트에서도 신흥 주제를 적절히 표현할 수 있는 디스크립터가 발견되지 않으면, 기존의 세 가지 주제 유형



〈그림 2〉 2차 새로운 주제 탐지의 흐름도

에 속하는 디스크립터와 별도로 새로운 후보 디스크립터 도출을 위한 단서로서 연관 키워드 군집을 생성하여 이용자에게 제시한다. 새로운 후보 디스크립터 도출을 위한 자질로는 신흥 주제가 부여된 문헌에 출현한 키워드를 대상으로 신흥 주제와의 시기별 유사도 차에 기초하여 상위 100개 키워드를 사용한다. 왜냐하면 새로운 후보 디스크립터 생성을 위한 주요 자질은 이전에는 해당 디스크립터와의 유사도가 낮았던 반면, 최근에는 유사도가 높아진 키워드이어야 하기 때문이다. 이에 따라 선정된 상위 100개 키워드를 대상으로 키워드-키워드 동시출현 행렬을 이용한 클러스터링과 키워드 쌍 순위화를 수행하여 연관 키워드 군집을 생성하였는데, 결과적으로 이러한 연관 키워드 군집에 포함된 키워드들이 신흥 주제의 새로운 측면을 표현하는 후보 디스크립터 생성을 위한 단서를 제공하게 되는 것이다.

3. 2 1차 새로운 주제 탐지 실험

1차 새로운 주제 탐지의 목적은 전체 디스크립터를 주제 유형별로 구분하는 것이다. 즉, 주제 분야의 변화 양상을 반영하는 다양한 정보에 기초하여 모든 디스크립터를 신생 주제, 신흥 주제, 기존 주제의 세 가지 주제 유형으로 분류하였다.

먼저, 디스크립터의 시기별 출현여부에 따라 최근에 새롭게 출현한 신생 주제를 분리하였다. 즉, 이전 문헌집단(1989년~2001년)에는 출현하지 않고 최근 문헌집단(2001년~2004년)에만 출현한 디스크립터(409개)를 신생 주제로 분류하였다. 또한, 최근 문헌집단에는 출현하지 않고 이전 문헌집단에만 출현한 디스크립터(1,265개)는 기존 주제로 분류하였다. 다음으로, 학습집단(이전집단)과 검증집단(최근집단) 양자에 출현한 디스크립터(503개)에 대하여 최근 집단의 문헌빈도가 기준치 이상(DF 〉 8)인 디스크립터만을 대상으로, 증가량(증가 지수: I)과 증가 추세(선형회귀선의 기울기: rs)의 두 가지 기준을 적용하여 신흥 주제 후보를 식별하였다. 이전집단과 최근집단 간의 디스크립터 문헌빈도의 증가량을 나타내는 증가 지수(I: increase index)는 다음의 공식으로 산출하고, 사전 실험에 따라 증가 지수가 $I \geq 0$인 디스크립터를 우선적으로 신흥 주제 후보로 선정하였다.

$$I = ln\left(\frac{DF_R}{top4DF_O}\right)$$

DF_R: 디스크립터의 최근 문헌집단 내 문헌빈도

$top4DF_O$: 디스크립터의 이전 문헌집단 내 문헌빈도(상위 4개년)

또한, 디스크립터의 증가 추세 측면에서는 실험집단의 전체 기간(1989년~2004년)을 고려하였다. 즉, 16년(1989년~2004년) 동안의 디스크립터 증가 추세를 나타내는 선형회귀선(Linear Regression Line), $y = a + bx$의 기울기(slope: $b = rs$)를 다음 공식으로 산출하고(안광호, 임병훈 2004), 사전 실험에 따라 기울기 값이 b〉 0.2인 디스크립터를 신흥 주제 후보로 추가하였다. 공식에서 독립변수인 x는 연도를, 종속변수인 y는 출현빈도를 나타낸다. 여기서 신흥 주제 후보로 식별된 디스크립터 이외에 위의 두 가지 기준을 만족하지 못한 디스크립터들은 기존 주제로 추가하였다.

$$a = \frac{(\sum y_i - b\sum x_i)}{n} = \bar{y} - b\bar{x}$$

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = rs$$

신흥 주제 후보를 대상으로 실제적인 신흥 주제를 식별하기 위한 방법으로는 기존의 지식 구조를 학습하고 이에 기초하여 새로운 정보의 식별 및 분류를 시도하는 연구들에서 최근 주로 사용되고 있는 나이브 베이즈(Naive Bayes) 분류기를 사용하였다. 문헌 범주화를 위한 대표적인 기계학습 알고리즘 중의 하나인 나이브 베이즈 알고리즘은 독립성 가설에 기초한 것으로 해당 범주에 문헌이 할당될 경우와 할당되지 않을 경우로 구분하여 문헌을 범주화하는 것이다. 문헌 범주화를 위한 공식은 다음과 같다(Jackson and Moulinier 2002).

$$P(C_i|D) = \frac{P(D|C_i)P(C_i)}{P(D)}$$

$P(C_i|D)$: 문헌 D가 범주 Ci에 분류될 확률

$D = (t_1, ..., t_n)$

$$P(D|C_i) = \prod_{j=1}^{n} P(t_j|C_i)$$

$P(t_j|C_i)$: 범주 Ci에 분류된 문헌이 단어 tj를 포함할 확률

3.3 2차 새로운 주제 탐지 실험

1차 새로운 주제 탐지가 전체 디스크립터를 기존 주제, 신생 주제, 신흥 주제로 구분하여 후보 디스크립터 리스트를 생성하는 것을 목적으로 하였다면, 2차 새로운 주제 탐지는 신흥 주제로 분류된 디스크립터를 위한 2차 후보 디스크립터 리스트를 생성하는 것이 목적이다. 즉, 신흥 주제를 대표하고 있는 기존의 디스크립터가 시간의 경과에 따른 해당 주제의 변화를 반영하지 못하고 있거나 부적절하게 표현하고 있다는 가정 하에, 신흥 주제를 위한 두 가지 유형의 후보 디스크립터 리스트를 생성하였다.

신흥 주제를 위한 첫 번째 후보 디스크립터 리스트에 포함되는 연관 디스크립터는 다음과

같은 절차에 따라 추출하였다.

(1) 신흥 주제로 분류된 디스크립터에 대한 문헌-디스크립터 행렬 생성
(2) 신흥 주제를 표현하는 디스크립터와 다른 디스크립터 간의 동시출현 행렬(디스크립터-디스크립터 행렬) 생성
(3) 각 디스크립터별로 동시출현 빈도에 따른 순위화
(4) 상위 10개 디스크립터를 연관 디스크립터로 추출

신흥 주제를 위한 두 번째 후보 디스크립터 리스트로서 색인 전문가의 새로운 디스크립터 도출을 위한 단서가 될 용어 군집을 생성하였다. 이를 위해, 먼저 신흥 주제 디스크립터가 부여된 문헌에 출현한 모든 키워드 중에서 신흥 주제 디스크립터와 이전보다 최근의 유사도가 높은 상위 100개의 키워드를 선정하였다. 즉, 이전 시기(1989년~2000년)에는 신흥 주제와의 유사도가 낮았지만 최근(2001년~2004년)에 와서 유사도가 높아진 키워드가 새로운 후보 디스크립터로서 더욱 가치가 있을 것으로 보고 두 시기의 유사도 간의 차이를 순위화의 기준으로 삼았다. 자카드 유사계수(Sneath and Sokal 1973)에 기초한 키워드 선정을 위한 공식은 다음과 같다.

$$JAC_{SD} = JAC_R - JAC_O$$

$$JAC_R = \frac{DF_R(NDDE, KW)}{(DF_R(NDDE) + DF_R(KW) - DF_R(NDDE, KW))}$$

$$JAC_O = \frac{DF_O(NDDE, KW)}{(DF_O(NDDE) + DF_O(KW) - DF_O(NDDE, KW))}$$

JAC_{SD}: 유사도 차, JAC_R: 최근 유사도, JAC_O: 이전 유사도
$DF_R(NDDE, KW)$: 키워드의 신흥 주제 디스크립터와의 최근 동시출현 빈도
$DF_R(NDDE)$: 디스크립터의 최근 문헌빈도
$DF_R(KW)$: 키워드의 최근 문헌빈도
$DF_O(NDDE, KW)$: 키워드의 신흥 주제 디스크립터와의 이전 동시출현 빈도
$DF_O(NDDE)$: 디스크립터의 이전 문헌빈도
$DF_O(KW)$: 키워드의 이전 문헌빈도

위의 공식에 따라 선정된 100개 키워드에 기초한 연관 키워드 군집은 다음과 같은 절차에 따라 생성하였다.

(1) 신흥 주제 디스크립터가 부여된 문헌에 출현한 키워드 중에서 신흥 주제 디스크립터와의 유사도순 상위 100개 키워드 선정
(2) 신흥 주제 디스크립터와의 유사도가 높은 상위 100개 키워드에 대하여 키워드-키워드 동시출현 행렬 생성
(3) 키워드-키워드 동시출현 행렬을 사용한 완전연결 클러스터링
(4) 신흥 주제와의 유사도가 높은 상위 100개 키워드에 대하여 키워드 쌍 리스트를 생성하고, 신흥 주제와의 동시출현 확률 차에 따라 순위화
(5) (3)의 결과에서 30군집 내 키워드와 (4)의 결과에서 상위 50위 내 키워드 쌍을 상호 비교하여 일치하는 군집(대표어+소속 키워드)들을 새로운 후보 디스크립터 생성을 위한 연관 키워드 군집으로 추출

4. 새로운 주제 탐지 실험의 결과 분석

본 연구에서 새로운 주제 탐지 실험은 1차로 전체 디스크립터를 신생 주제, 기존 주제, 신흥 주제의 세 가지 주제 유형으로 구분하고, 이 중

에서 신흥 주제를 기존의 주제에 긍정적 측면의 변화가 발생하여 새로운 정보가 추가된 주제로 파악하였다. 또한, 2차로 신흥 주제로 분류된 디스크립터의 적절한 처리를 위한 추가적인 디스크립터 선정을 위해 사용될 연관 디스크립터를 추출하고 새로운 디스크립터 도출을 위한 연관 키워드 군집을 생성하였다.

디스크립터들이 학문 분야의 지식구조를 반영하고 있다고 가정한다면, 신흥 주제는 최근에 와서 기존의 주제에 성장 혹은 확장 등 어떤 긍정적인 측면의 변화가 발생하여 이전의 디스크립터로는 표현하기에는 부족하거나 부적합한 주제라고 할 수 있다. 따라서 신흥 주제 후보로 식별된 디스크립터에 대하여 새로운 주제 탐지를 위한 방법론 중의 하나인 기계학습을 적용하여 이러한 신흥 주제의 특성을 식별하였다.

기계학습 프로그램 WEKA에서 제공하는 나이브 베이즈 알고리즘을 이용한 범주화 결과의 성능은 재현율과 정확률로 측정하였으며, 〈표 1〉과 같은 각 범주에 대한 분류 결과를 표현하는 분할표를 이용하여 산출하였다(Yang 1999).

〈표 1〉 범주화 성능 산출을 위한 분할표

	범주 적합문헌	범주 부적합문헌
범주에 할당	a	b
범주에 할당되지 않음	c	d

a: 해당 범주로 정확하게 할당된 문헌수
b: 해당 범주로 부정확하게 할당된 문헌수
c: 해당 범주로 부정확하게 할당되지 않은 문헌수
d: 해당 범주로 정확하게 할당되지 않은 문헌수

$$재현율 = \frac{a}{a+c}$$

$$정확률 = \frac{a}{a+b}$$

42개 신흥 주제 후보 디스크립터에 대하여 나이브 베이즈 알고리즘을 이용한 범주화 성능을 최고값, 최저값, 평균으로 구분하여 산출한 결과는 〈표 2〉와 같다.

범주화 성능의 해석 측면에서, 학습집단(이전 문헌집단)의 데이터로 학습하고 검증집단(최근 문헌집단)의 데이터로 검증한 결과에서 해당 디스크립터로 범주화되는 성능의 완전성을 의미하는 재현율(recall)이 50%가 되지 않으면, 최근에 디스크립터가 대표하고 있는 주제에 어떤 변화가 발생하여 검증집단(최근 집단)의 문헌 중 절반 이상이 해당 범주로 적절하게 범주화되지 않은 것으로 간주하였다. 즉, 이전 단계에서 신흥 주제 후보를 선정하는 기준이 디스크립터의 증가량과 증가 추세를 반영한 것이라는 측면에서 긍정적인 측면의 변화라고 할 수 있으므로, 해당 주제의 성장 혹은 발전에 따라 기존 주제에 어떤 변화가 발생하여 해당 범주로의 할당에 실패한 것으로 고려할 수 있을 것이다. 따라서 나이브 베이즈 학습 알고리즘을 이용한 범주화 재현율이 50% 이하인 디스크립터(nb_R≤0.5)는 기존 주제에

〈표 2〉 나이브 베이즈 알고리즘을 이용한 범주화 성능 요약

구분 \ 성능	최고값	최저값	평균
나이브 베이즈 재현율(nb_R, yes)	0.735	0.000	0.258
나이브 베이즈 정확률(nb_P, yes)	0.556	0.000	0.188

긍정적인 측면의 변화가 발생하여 성장·발전의 단계에 있는 신흥 주제로 분류하였다. 이에 따라, 전체 42개 신흥 주제 후보 중에서 재현율이 50%를 넘는 'automatic text analysis', 'searching', 'scientometrics', 'world wide web' 등은 기존 주제로 분류되었고, 재현율이 50% 이하인 나머지 디스크립터(38개)는 신흥 주제로 분류되었다. 이와 같이 식별된 신흥 주제 디스크립터들은 새로운 주제를 표현하기에 부족하다고 보고, 각 디스크립터에 대한 연관 디스크립터와 새로운 디스크립터 생성을 위한 연관 키워드 군집을 생성하였다.

먼저, 신흥 주제 디스크립터에 대한 연관 디스크립터를 추출한 결과의 예는 〈표 3〉과 같다.

여기서 신흥 주제 'algorithms'의 상위 10개 연관 디스크립터 가운데 굵은 활자로 표시된 디스크립터는 모두 스스로도 신흥 주제로 분류된 디스크립터이다. 이처럼 신흥 주제 디스크립터들에 대하여 이와 밀접하게 관련된 상위 10개의 연관 디스크립터를 추출하였을 때, 연관 디스크립터 그 자체가 신흥 주제 디스크립터인 경우가 많았다(평균 45.3%). 특히, 신흥 주제 디스크립터인 'interactive systems', 'internet', 'students', 'search strategies', 'online information retrieval', 'end users' 등은 연관 디스크립터 중의 70% 이상이 신흥 주제이면서 서로 상대적인 연관 디스크립터 관계를 구성하고 있었다. 따라서 신흥 주제들은 서로 간에 밀접한 연관 관계를 가지고 있으며 긍정적 측면의 변화로 성장·발전하고 있는 주제들이라는 것을 유추할 수 있다.

또한, LISA의 디스크립터 필드에서 추출한 디스크립터와 LISA 시소러스의 디스크립터를 대조해 본 결과, LISA의 디스크립터 필드에는 시소러스에 수록된 실제적인 디스크립터 이외에도 새롭게 문헌에 출현한 인명, 기관명, 시스템명, 회의명 등이 포함되어 있었다. 본 연구에서 신흥 주제로 식별된 38개 디스크립터 중에서 시소러스에 수록되지 않은 용어는 유일하게 'trec'으로 밝혀졌다. 시소러스에 수록된 나머지 37개 신흥 주제 디스크립터 간의 관계를 LISA 시소러스를 검색하여 살펴 본 결과, 이 중에서 11개 디스크립터(29.7%)가 시소러스의 상위어(BT), 하위어(NT), 관련어(RT) 관계로 서로 연결되어 있었다. 예를 들면, 자신이 신흥

〈표 3〉 신흥 주제 디스크립터의 연관 디스크립터 추출 결과의 예

신흥 주제	순위	연관 디스크립터
algorithms	1	automatic text analysis
	2	**online information retrieval**
	3	**automatic classification**
	4	clustering
	5	databases
	6	**evaluation**
	7	**interactive systems**
	8	**models**
	9	natural language porcessing
	10	**serach engines**

주제이면서 다른 신흥 주제의 상위어인 디스크립터는 'information seeking behaviour', 'user surveys' 등이었고, 반면 다른 신흥 주제의 하위어인 디스크립터는 'online information', 'end users', 'search strategies' 등이었다. 또한, 신흥 주제 디스크립터가 다른 신흥 주제 디스크립터의 관련어인 경우가 가장 많았는데, 'web sites', 'digital libraries', 'evaluation', 'surveys', 'web pages', 'research' 등이 여기에 속한다.

다음으로, 신흥 주제 디스크립터를 위한 연관 키워드 군집을 생성하였다. 신흥 주제 디스크립터와의 시기별 유사도 차이에 따라 선정된 100개 키워드를 대상으로 완전연결 클러스터링과 키워드 쌍 순위화를 적용하여 생성한 관련 연관 키워드 군집은 색인 전문가가 색인작업시 새로운 디스크립터를 생성할 수 있는 단서로서 제공될 수 있다.

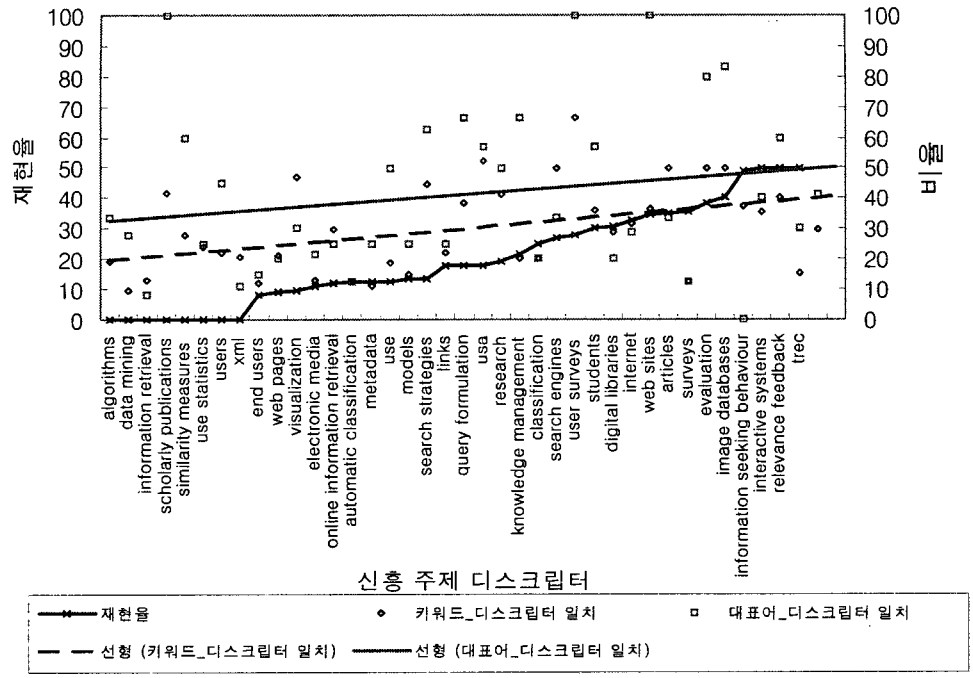
완전연결 클러스터링과 키워드 쌍 순위화의 결과인 30군집 내 키워드와 상위 50위 내 키워드 쌍을 구성하는 키워드를 함께 고려하여 최종적으로 생성한 연관 키워드 군집 리스트는 색인 전문가의 새로운 디스크립터 도출을 위한 단서를 제공하는 역할을 한다. 이러한 연관 키워드 군집의 예는 〈표 4〉와 같다.

신흥 주제 디스크립터 38개 모두에 대하여 이러한 군집을 생성한 결과, 각 디스크립터 당 1개에서 10개까지(평균 4.3개)의 군집이 형성되었다. 이 중 군집 대표어로 추출한 용어가 전체 디스크립터 집합에 포함된 디스크립터와 일치하는 경우는 평균 41.1%였고, 군집 대표어로 추출한 용어가 해당 신흥 주제 디스크립터와 일치한 경우는 31.6%였다. 또한, 군집을 형성한 소속 키워드 중에서 전체 디스크립터 집합의 디스크립터와 일치하는 경우는 29.9%였다.

〈그림 3〉은 신흥 주제 디스크립터를 기계학습을 통한 범주화 성능(재현율)에 따라 배열하고, 이러한 신흥 주제 디스크립터와 동시출현한 키워드 집합을 대상으로 생성한 연관 키워드 군집에서 군집 대표어와 군집 내 소속 키워드가 전체 디스크립터 집합의 디스크립터인 비율을 산출하여, 이를 선형회귀선으로 표시한

〈표 4〉 신흥 주제 디스크립터 'user surveys'의 새로운 디스크립터 도출을 위한 연관 키워드 군집

순위	군집 대표어(키워드쌍)	군집 내 키워드	군집
1	university : library	behaviour library make online relationship university	9
2	collect : data	collect data service	14
3	access : internet	access internet source	1
4	electronic : mail	electronic mail	27



〈그림 3〉 연관 키워드 군집의 대표어 혹은 소속 키워드가 신흥 주제 디스크립터인 비율

것이다. 그림에서 보는 바와 같이 신흥 주제 디스크립터와 최근 자주 동시출현한 키워드들로 생성한 연관 키워드 군집에서 대표어로 선정한 용어들이 소속 키워드들 보다 신흥 주제 디스크립터가 되는 비율이 더 높다는 것을 알 수 있다.

5. 결 론

본 연구에서는 특정 학문 분야의 개념들과 개념들 간의 관계로 구성된 지식구조를 바탕으로 색인 전문가가 학문의 변화 및 발전 양상을 반영하여 주제를 표현하는 디스크립터를 색인어로 부여하고, 이를 통해 지식구조를 적절하게 갱신할 수 있도록 지원하는 방안을 제시하였다. 특히, 정보학 분야를 대상으로 새로운 주제의 출현 및 성장으로 인하여 지식 구조상의 변화가 발생하였을 때, 기존의 색인어로는 이를 표현할 수 없거나 표현상의 제한이 따르는 문제를 해결하기 위한 방안으로 새로운 주제 탐지 방법을 응용한 후보 디스크립터 리스트의 사용을 제안하였다.

지금까지 이용자의 정보검색을 지원하기 위한 도구에 관련된 연구는 활발히 이루어져 왔지만, 색인 전문가에 의한 색인작업의 효과적인 지원을 위한 연구는 거의 찾아볼 수 없었다. 그러나 학문 분야의 지식 구조를 구성하는 적절한 통제언어 색인어의 선정 및 부여는 해당 분야의 지식에 대한 체계적인 구조화와 이를

통한 검색의 성능 향상에도 상당한 영향을 미칠 수 있다. 따라서 이 연구에서는 지식 구조를 구성하는 기본적인 요소로서 통제언어 색인어인 디스크립터를 적절하게 부여 혹은 갱신할 수 있는 효율적인 지원 도구로서 새로운 주제 탐지 방법을 응용한 후보 디스크립터 리스트의 생성을 위한 실험을 수행하고 그 결과를 분석하였다.

또한, 새로운 주제 탐지 실험의 결과인 두 가지 후보 디스크립터 리스트를 사용하는 지식 구조 갱신 과정을 모의실험하고 그 결과를 평가하였다. 즉, 문헌정보학 대학원 석·박사 과정 10명을 대상으로 색인자 집단을 구성하고 서로 다른 조건에서 디스크립터를 부여하게 한 다음, 이들이 부여한 디스크립터와 LISA에서 문헌에 부여한 디스크립터 간의 일관성을 일관성 척도인 CP(Consistency Pair)를 사용하여 평가하였다(Lancaster 1978, 197).

색인자의 색인작업을 지원하기 위한 도구로서 후보 디스크립터 리스트를 생성하기 위한 새로운 주제 탐지 실험 및 지식 구조 갱신 실험을 수행한 결과 밝혀진 사실은 다음과 같다.

첫째, 연관 디스크립터 리스트 생성을 위해 신흥 주제인 디스크립터에 대하여 상위 10개의 연관 디스크립터를 추출하였을 때, 연관 디스크립터 자체가 신흥 주제인 비율이 평균 45.3%로 상당히 높았다. 따라서 특정 학문 분야 내에서 신흥 주제들은 상당수가 서로 밀접하게 연관되어 있으면서 동시에 성장·발전의 단계에 있는 주제라는 것을 확인할 수 있었다.

둘째, 실험을 위하여 LISA의 디스크립터 필드에서 추출한 디스크립터와 실제 LISA 시소러스의 디스크립터를 대조해 본 결과, LISA의 디스크립터 필드에는 시소러스에 수록된 실제적인 디스크립터 이외에도 새롭게 문헌에 출현한 인명, 기관명, 시스템명, 회의명 등이 다수 포함되어 있었다. 그러나 신흥 주제로 식별된 38개 디스크립터 중에서 시소러스에 수록되지 않은 용어는 유일하게 'trec'뿐인 것으로 확인되었다. 또한, 나머지 37개 신흥 주제 디스크립터 간의 연관관계를 실제 LISA 시소러스를 검색하여 살펴 본 결과, 이 중에서 11개 디스크립터(29.7%)가 시소러스의 상위어(BT), 하위어(NT), 관련어(RT) 관계로 서로 긴밀하게 연결되어 있었다.

셋째, 새로운 디스크립터 도출을 위한 연관 키워드 군집 리스트를 생성한 결과, 신흥 주제 디스크립터 당 1개에서 10개까지(평균 4.3개) 군집이 형성되었다. 이 중 군집 대표어로 추출한 용어가 전체 디스크립터 집합에 포함된 디스크립터와 일치하는 경우는 평균 41.1%이었고 특히, 군집 대표어로 추출한 용어가 해당 신흥 주제 디스크립터와 일치한 경우는 31.6%였다. 또한, 군집을 형성한 소속 키워드 중에서 전체 디스크립터 집합의 디스크립터와 일치하는 경우는 29.9%였다.

넷째, 신흥 주제 디스크립터와 동시출현한 키워드 집합을 대상으로 생성한 연관 키워드 군집에서 군집 대표어와 군집 내 소속 키워드가 전체 디스크립터 집합의 디스크립터와 일치하는 비율을 산출하여 선형회귀선으로 표시한 결과, 군집의 대표어로 선정한 용어들이 군집 내 소속 키워드보다 디스크립터와의 일치율이 더 높다는 사실을 알 수 있었다.

다섯째, 지식 구조 갱신 실험 및 결과 평가에 따르면 전반적으로 기본정보만을 사용한 경우

(평균 0.105) 보다 새로운 주제 탐지를 통하여 생성한 리스트를 사용한 것이 더 좋은 결과를 보여주었다. 특히, 단순히 주제 유형정보와 함께 유사도순으로 디스크립터를 제시한 1차 후보 디스크립터 리스트만을 사용한 결과(평균 0.175) 보다 1차 후보 디스크립터 리스트와 신흥 주제와 관련한 연관 디스크립터와 연관 키워드 군집을 제시한 2차 후보 디스크립터 리스트를 함께 사용한 결과(평균 0.211)가 가장 좋은 결과를 보였다.

이상의 실험 결과는 새로운 주제 탐지를 통한 후보 디스크립터 리스트의 사용이 색인자의 색인작업을 적절하게 지원함으로써 성장 및 발전 등 긍정적 측면의 변화를 적시에 적절하게 반영하는 지식 구조의 갱신을 위한 효율적인 방안이 될 수 있다는 가능성을 보여주었다. 색인자의 색인작업을 지원하기 위한 도구가 거의 없는 상황에서 문헌에 적합한 디스크립터의 선정 및 부여를 위한 다양한 기반 정보를 제공할 수 있는 후보 디스크립터 리스트의 제공은 색인작업의 효율성과 정확성을 향상시킬 수 있을 것이다. 뿐만 아니라 색인자가 새로 입력되는 문헌에 대하여 기존 주제의 변화를 적절하게 반영하여 디스크립터를 부여하는 행위 자체가 결과적으로 실시간으로 지식 구조를 적절하게 갱신하는 효과를 가져 올 수 있을 것이다.

참 고 문 헌

김태수. 2000. 『분류의 이해』. 서울: 문헌정보처리연구회.

안광호, 임병훈. 2004. 『SPSS를 활용한 사회과학조사방법론』. 서울: 학현사.

정영미. 1993. 『정보검색론』. 서울: 구미무역(주) 출판부.

Allan, J., R. Papka, and V. Lavrenko. 2001. "Topic Models for Summarization Novelty". In *Proceedings of the Workshop on Language Modeling in Information Retrieval*, 66-71.

Dkaki, Taoufiq, Josiane Mothe, and Jerome Auge. 2002. "Novelty Track at IRIT-SIG." In *Proceeding of Eleventh Text REtrieval Conference (TREC-11)*. [cited 2004.4.5]. 〈http://trec.nist.gov/pubs/trec11/papers/irit.novelty.pdf〉.

Harman, Donna. 2002. "overview of the TREC 2002 Novelty Track." In *Proceeding of Eleventh Text Retrieval Conference(TREC-11)*, [cited 2004.4.5]. 〈http://trec.nist.gov/pubs/trec11/papers/NOVELTY.OVER.pdf〉.

Holtzman, L. E. et al. 2004. "A Software Infrastructure for Research in Textual Data Mining." [cited 2004.4.2]. 〈http://hddi.cse.lehigh.edu/docs/AITools.pdf〉.

Jackson, P. and I. Moulinier. 2002. *Natural Language Processing for Online Ap-*

plications: Text Retrieval, Extraction and Categorization. John Benjamins Publishing Company.

Japkowicz, N., C. Myers, and M. Gluck. 1995. "A Novelty Detection Approach to Classification." In *Proceedings of the 14th International Conference on Artificial Intelligence*, 518-523.

Kontostathis, A. et al. 2003. "A Survey of Emerging Trend Detection in Textual Data Mining." In *Survey on text mining : clustering, classification, and retrieval.* ed. by Michael W. Berry. New York: Springer_Verlag.

Lancaster, F. W. 1978. *Information Retrieval Systems : Characteristics, Testing, and Evaluation.* N. Y.: John Wiley & Sons.

Manevitz, L. and M. Yousef. 2001. "One-Class SVMs for Document Classification." *Journal of Machine Learning Research*, 2: 139-154.

Markou, M. and S. Singh. 2003. "Novelty Detection: A Review, Part I: Statistical Approaches." *Signal Processing* (under submission, 2003), [cited 2003. 5. 21].
〈http://www.dcs.ex.ac.uk/research/pann/pdf/pann_SS_086.PDF〉.

Marsland, S. 2003. "Novelty Detection in Learning Systems." In *Neural Computing Surveys* 3, 157-195. [cited 2004. 7].
〈http://www.icsi.berkeley.edu/~jagota/NCS〉.

Pottenger, W. M. and T. Yang. 2001. "Detecting Emerging Concepts in Textual Data Mining." In *Computational Information Retrieval.* ed. by Michael W. Berry. Philadelphia: SIAM.

Roy, Soma, Gevry David, and William M. Pòttenger. 2002. "Methodologies for Trend Detection in Textual Data Mining." In *Proceedings of the Textmine '02 Workshop, Second SIAM International Conference on Data Mining.*

Sneath, Peter H. A. and Robert R. Sokal. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification.* San Francisco: W. H. Freeman and Company.

Soboroff, I. and Donna H. 2003. "overview of the TREC 2003 Novelty Track." In *Proceedings of the Twelfth Text REtrieval Conference(TREC 2003).* [cited 2004.9.20].
〈http://trec.nist.gov/pubs/trec12/papers/NOVELTY.OVERVIEW.pdf〉.

Yang, Y. 1999. "An Evaluation of Statistical Approaches to Text Categorization." *Information Retrieval*, 1: 69-90.

Zhang, Yi, J. Callan, and M. Thomas. 2002. "Novelty and Redundancy

Detection in Adaptive Filtering." In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research an Development in Information Retrieval*, 81-88.