# 데이터의 카테고리 연관성을 이용한 색인어 자동 추출

우영호*, 허태성**, 허웅***, 박영배****, 민홍기*
*인천대학교 정보통신공학과
*인하공업전문대학 컴퓨터시스템공학부
**명지대학교 전자공학과
***경희대학교 한의학과

# Automated Keyword Extraction using Category Correlation of Data

Young-ho Woo*, Tae-sung Hur**, Woong Her***, Young-bae Park****, Hong-Ki Min*
*Dept. of Information and Telecommunication Eng., University of Incheon
**School of Computer, System Eng., Inha Technical College.
***Dept. of Electronic Eng., Myong-Ji University.
****College of Oriental Medicine, Kyung-Hee University.

## 요 약

본 논문에서는 특정 영역에서 나타날 수 있는 데이터를 카테고리별로 저장한 시소러스를 이용하여 색인어 후보를 추출한다. 그리고 각 데이터의 카테고리 간의 상호 연관성을 고려하여 검출되는 색인어의 정확도를 향상시킬 수 있는 연관 중요도를 적용한 색인어 자동 추출 시스템을 제안하였다. 제안된 시스템은 출현빈도를 고려한 방법보다 47% 시소러스를 이용한 방법보다 18% 향상된 성능을 보였다.

## I. 서 론

정보 검색에서 다루는 데이터는 자연어로 구성된 비정형 데이터이므로 컴퓨터가 이를 직접 처리할 수 없다. 문서를 컴퓨터가 직접 처리할 수 있는 형태로서 정형적인 데이터로 변환하는 것이 문서의 인덱스이다[1]. 인덱스를 통해 검출되는 단어를 색인어라 한다. 특정 정보 항목을 대변하는 색인어 집합은 그 정보 항목의 내용을 충실히 반영하면서 동시에 다른 항목과의 차별화가 되도록 선정되어야 한다.

웹 검색 엔진을 비롯해서 모든 정보 검색 시스템의 검색 결과에 원하지 않는 데이터가 포함되어 있는 근본적인 원인은 크게 두 가지로 볼 수 있다. 첫 번째는 색인어가 원천적으로 문서 내용을 그대로 대변할 수 없다는 것이고, 두 번째는 시스템이 처리해야할 질의가 사용자의 정보 요구를 제대로 표현하지 못하기 때문이다[2]. 정보 검색 분야 뿐 아니라 분류, 관리 등의 정보 핸들링 분야에서 보다 정확한 결과를 제시하기 위해서는 시스템이 비정형성을 띄는 문서에서 해당 문서의 내용을 충실히 표현할 수 있는 색인어를 추출해낼 수 있어야 한다.

본 논문에서는 색인어 추출의 기본적인 형태로 볼 수 있는 출현 빈도를 이용함과 동시에 서로 유관한 카테고리를 포함한 시소러스를 구축하여, 이를 복합적으로 적용한 출현 빈도와 상호 연관성을 고려하는 하는 색인어 추출 방법을 제안한다. 제안된 시스템은 정확도 및 재현율에서 기존의 시스템보다 향상된 성능을 보였다.

## II. 카테고리별 연관성의 적용

전문 분야의 내용을 포함하는 문서의 경우 색인어 추출 시, 정확도의 향상을 위해 출현 가능성이 있는 단어나 어휘가 저장된 시소러스를 이용한다. 본 논문에서 제안하는 시스템의 시소러스의 데이터는 그 종류에 따라 카테고리가 분류되어 있다. 임의의 연구 문서에서 색인어 추출 시 연구의 특징이 저장된 카테고리와 연구 대상이 저장된 카테고리의 연관도는 매우 높다. 이를 기초로 하여 두 카테고리의 데이터 연관성의 중요도를 예상할 수 있다. 예를 들면, '사과'와 '당도 증가'의 조합이 '사과'와 '탄수화물 대사'의 조합보다 높은 중요도 점수를 부여받을 수 있다. 상호 연관성을 이용한 색인어 추출을 위해서 다음과 같

은 과정을 수행한다. 상호 연관도를 가질 수 있는 카테고리를 각각 A, B라 하고 각 카테고리의 색인어 후보로 선정된 데이터를 A1 ~A3라 B1 ~B5라 하면, A와 B의 쌍을 각각 모든 경우의 수에 따라 시소러스에 질의한다. 검색되어 나오는 데이터의 건수가 해당 쌍의 중요도 점수가 부여되고 점수가 높은 색인어 후보 쌍이 색인어로 추출된다. 만약 점수가 동일한 색인어 후보 쌍이 존재할 경우 더 많은 형태소가 결합된 데이터가 포함된 색인어 후보 쌍을 색인어로 선택한다. 상호 연관도를 조사하여 색인어를 선택하는 과정을 그림 1에 나타나있다.

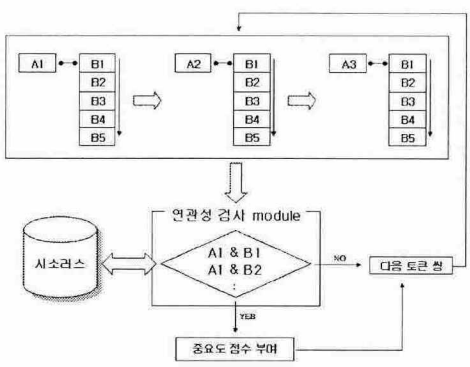


그림 1. 연관성을 이용한 중요도 점수 적용의 예

## Ⅲ. 제안된 시스템

출현빈도를 사용하여 색인어 추출을 하는 경우 정확도와 일관성에서의 문제점이 나타난다. 또한 역문서 빈도, 역단어 빈도를 이용하는 방법은 특정 연구에 관한 문건이 다량 존재하는 경우 특정 단어가 대부분의 표본 문서에서 존재할 수 있으므로 정확도가 낮아지는 문제점을 갖고 있다. 또한 자연어에서는 같은 주제라도 문헌 생산자 마다 그 표현하는 용어가 달라질 수 있어 문헌의 분석이나, 색인 작성 시에 많은 어려움이 야기된다. 따라서 필요한 정보를 찾으려고 하는 이용자는 하나의 검색어만으로 해당 주제를 전부 검색할 수 없으므로 그 검색어에 관련된 개념의 관련어, 동의어 등을 모두 검색하여야하는 번거로움이 있다.

이에 그 해당하는 주제 분야에서 필요한 용어를 수집, 분류하여 해당 용어를 대표할 수 있는 데이터를 저장하여 시소러스를 구축하여 색인어 추출 시 이용하면 이러한 문제점을 해결할 수 있다. 또한 시소러스 내에서 용어가 분류될 때 나타나는 연관성을 고려하면 보다 정확한 색인어 추출을 기대할 수 있다.

본 논문에서는 출현빈도와 함께 시소러스를 이용하여 색인어를 각각의 카테고리에 분류하고 분류된 데이터의 상호 연관성을 고려한 색인어 추출 방법을 제안한다. 문서가 입력되면 우선 전처리 과정을 통해 토큰화와 불용어 필터링을 수행한다. 문서의 내용이 문자열 형태로 입력되게 되는데 효율적인 처리를 위하여 의미를 가진 가장 작은 단위인 어절 형태로 분할하게 되는데 이를 토큰화라 한다. 이렇게 분할된 토큰에는 전치사, 관사 등 색인어로 추출될 경우 무의미한 용어가 포함되어 있으므로 이를 제거하는 불용어 필터링을 수행한다. 불용어 필터링 과정은 제안된 방법에서의 필수 요소라기보다는 처리 속도 개선을 위한 필요 요소라 볼 수 있다.

전처리 과정을 완료한 토큰은 시소러스에 질의되고 매칭되는 데이터가 존재하는 경우 해당 카테고리로 분류된다. 그러나 해당 토큰이 시소러스에 나타나지 않는 경우 빈도측정을 위한 테이블에 저장된다. 복합어, 축약형 등의 처리를 위해서 시소러스에서 토큰을 검색하여 분류할 때 카테고리 분류 규칙을 적용한다. 이렇게 카테고리 분류가 완료된 토큰들은 색인어 후보로 지정된다. 상호 연관성을 가지는 카테고리에 포함된 색인어 후보들은 이를 고려하여 중요도 점수를 부여받게 되고 높은 점수를 받은 색인어 후보가 각 카테고리의 색인어로 추출되게 된다. 제안된 시스템의 전체 흐름을 나타내는 다이어그램이 그림 2에 나타나있다.

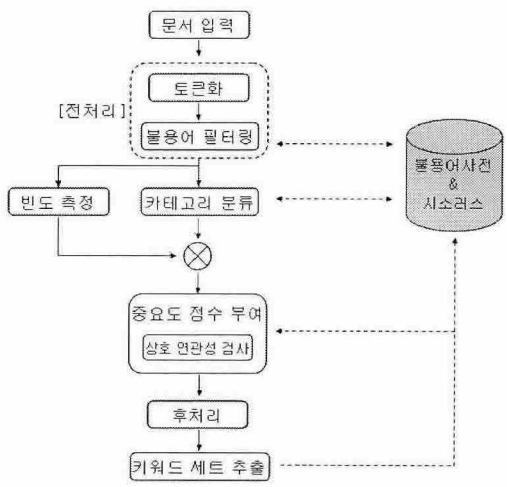


그림 2. 제안된 시스템의 블록다이어그램

## Ⅳ. 실험 결과 및 분석

### 1. 실험 자료

성능평가를 위해 30개의 Living Modified Organism (LMO) 관련 연구 문서를 수집하였다. 최근 이슈가 되고 있는 LMO 즉, 유전자 조작 생물체 분야는 많은 연구 데이터가 존재하지만 데이터의 비정형성으로 인해 정보의 분류, 관리 및 검색이 어려운 실정이다. LMO 연구와 관련한 영어 원문을 포함하는 실험 데이터는 www.agbios.com 사이트에서 무작위로 추출하였다. 해당 사이트는 본문과 함께 본문의 색인어 또는 요약 문장을 함께 제공하므로 시스템의 데이터 정확도를 평가하기에 매우 적합하다. 수집된 문서는 평균 36.42개의 문장을 가지며 각 문장은 약 24어절로 이루어져 있다. 문서 내 불

용어 포함률은 평균 36.3%였다. 표본 문서의 통계적인 특성은 표 1과 같다.

표 1. 표본 문서의 통계적 특성

| 대상 영역 | LMO 연구 문헌 |
|---|---|
| 문서 개수 | 30 건 |
| 문서의 평균 길이 | 34.7 문장 |
| 문장의 평균 길이 | 24.17 어절(단어) |
| 불용어 포함율 | 35.7 (%) |

## 2. 시소러스의 구축

시소러스를 이용한 색인어 자동 추출에서는 시소러스의 데이터 및 구성이 전체 시스템의 정확도 및 수행 시간에 막대한 영향을 준다. 본 논문은 색인어를 카테고리 별로 분류하므로 시소러스 또한 모든 데이터가 객체명, 연구특징, 수행국가, 연구년도의 네 가지의 카테고리에 분류되어 테이블에 저장되어 있다.시소러스를 위한 데이터는 LMO 관련 연구 내용을 색인어 형태로 제공하는 사이트인 www.binas.com, www.isb.vt.edu 와 www.biotrack.com에서 추출하였다. 시소러스는 23,645개의 데이터를 보유하고 있다.

## 3. 실험 결과

제안된 방법을 사용하여 추출한 표본 문서 10건의 색인어와 해당 문서의 실제 색인어가 표 2에 나타나 있다.

표 2. 실제 색인어와 추출된 색인어의 예

| | 시스템을 통해 추출된 색인어 | | | | 실제 색인어 | | | |
|---|---|---|---|---|---|---|---|---|
| | org_name | trait | country | year | org_name | trait | country | year |
| text 001 | brassica napus | glyphosate herbicide tolerance | Canada | 2004 | brassica napus | glyphosate herbicide tolerance | Canada | 2003 |
| text 002 | soybean | glufosinate ammonium tolerance | United States | 2004 | soybean | glufosinate ammonium tolerance | United States | 1998 |
| text 003 | cotton | insect resistance | United States | 2004 | cotton | lepidopteran pests resistance | United States | 2003 |
| text 004 | creeping bentgrass | glyphosate tolerance | United States | - | creeping bentgrass | glyphosate herbicide tolerance | United States | 2003 |
| text 005 | canola | glyphosate tolerance | Canada | 2004 | argentine canola | glyphosate herbicide tolerance | Canada | 2004 |
| text 006 | canola | gluphosinate ammonium tolerance | Canada | 2004 | argentine canola | glufosinate ammonium tolerance | Canada | 1998 |
| text 007 | chicory | glufosinate herbicide tolerant | - | 1990 | chicory | glufosinate ammonium tolerance | United States | 1997 |
| text 008 | cotton | insect resistance | United States | 2004 | cotton | lepidopteran pests resistance | United States | 2004 |
| text 009 | squash | virus resistance | United States | 2004 | squash | viral infection resistance | United States | 1998 |
| text 010 | soybean | seed fatty acid | United States | 2004 | soybean | modified seed fatty acid content | United States | 2001 |

## 4. 실험 결과 분석

본 논문에서 제안된 시스템의 객관적인 성능 측정을 위해 출현 빈도를 고려한 방법과 데이터의 상호 연관성을 배제한 시소러스 이용 방법으로 동일 데이터에 대한 각각의 결과를 생성하였다. 단순히 시소러스를 이용한 색인어 추출 방법은 제안된 색인어 추출 방법이 사용하는 시소러스를 그대로 사용하여 시소러스의 규모 및 정확도에서 오는 편차를 방지하였다. 시스템의 성능을 평가하기 위한 기준은 전통적으로 사용되는 정확도와 재현율을 사용하였다. 정확도와 재현율을 구하기 위해 아래 식을 사용하며 P는 정확도(Precision), R은 재현율(Recall)을 의미한다.

$$P(\mathrm{Pr}ecision) = \frac{\text{검출된 키워드 중 실제 키워드 토큰의 수}}{\text{검출된 키워드 토큰의 수}}$$

$$R(Recall) = \frac{\text{검출된 키워드 중 실제 키워드 토큰의 수}}{\text{전체 키워드 토큰의 수}}$$

출현빈도를 사용한 키워드 추출 방법을 Method_A, 상호 연관성을 배제한 단순한 시소러스 참고 방법을 Method_B라 했을 때, 각각의 방법을 사용한 경우에 30개의 표본 문서에서 추출된 색인어의 정확도와 재현율은 표 3과 같고 이를 그림 2에 그래프로 나타내었다.

표 3. 각 시스템의 정확도와 재현율

| | Method_A | Method_B | 제안된 방법 |
|---|---|---|---|
| 정확도(P) | 0.32 | 0.61 | 0.81 |
| 재현율(R) | 0.41 | 0.52 | 0.71 |

제안된 방법을 사용한 경우 전체 30개의 문서에서 추출된 색인어의 정확도는 81%로 Method_A의 32%, Method_B의 61% 보다 좋은 성능을 보였다. 재현율에서는 제안된 방법이 71%로 Method_A와 Method_B의 41%, 52%보다 높은 수치를 기록하였다.

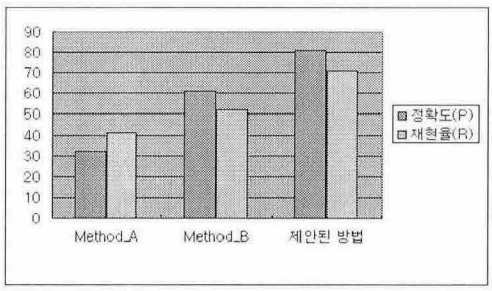


그림 3. 각 시스템의 정확도 그래프

분석 결과, 그림 3에서 나타난 것과 같이 본 논문에서 제안한 시스템이 출현 빈도만을 고려한 Method_A와 상호 연관성을 배재한 단순한 시소러스 참고 방법을 이용한 Method_B 보다 정확도 및 재현율 모두에서 우수한 성능을 보였다. 제안된 시스템을 통해 색인어를 추출한 결과의 각 카테고리 별 평균 정확도는 그림 4와 같다.

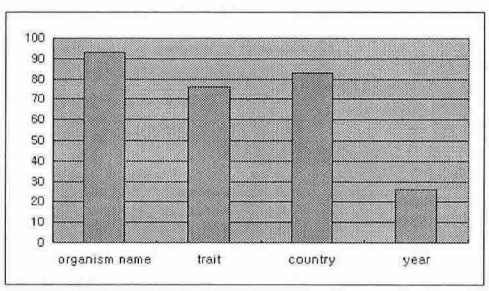


그림 4. 추출된 색인어의 카테고리 별 평균 정확도

수행년도 카테고리의 경우 원문에서 연구 년도를 표기하는 경우가 비교적 적어 아예 검출이 안 되거나 이전 연구의 데이터를 참고하여 그 연구 기간 등을 원문에서 이용한 경우 때문에 정확도가 27%로 다른 카테고리의 데이터에 비해 그 정확도가 낮았다. 그러나 연구 문헌에서 해당 연구의 수행 연도보다 더 중요한 요소로 평가받는 객체명, 수행국가 카테고리의 색인어 정확도는 각각 93%, 83%의 정확도를 나타내었다. 또한 객체명 카테고리와의 상호 연관성을 고려한 연구특징 카테고리의 정확도는 76%로 나타난 것을 확인할 수 있었다.

## V. 결 론

본 논문에서는 특정 영역에서 나타날 수 있는 데이터를 카테고리별로 저장한 시소러스를 이용하여 색인어 후보를 추출하고, 각 데이터의 카테고리 간의 상호 연관성을 고려하여 검출되는 색인어의 정확도를 향상시킨 색인어 자동 추출 방법을 제안하였다. 시소러스는 23,645개의 데이터를 포함하고 있으며 상호 연관성을 고려하는 대상이자 연구의 주제가 되는 객체명과 연구특성의 대표값이 각각의 필드에 나뉘어 저장되어 있다.

제안된 방법을 사용하여 생명공학분야의 LMO 관련 문서 30건에 대상으로 색인어 추출 실험을 하였다. 색인어 추출 결과 제안된 방법을 사용했을 경우 정확도(P)가 76%로 출현빈도만을 사용한 방법의 29%보다 약 47%의 향상을 보였으며 상호 연관성을 고려하지 않고 시소러스만을 사용한 경우의 58%보다 약 18%의 향상을 보였다. 재현율(R)에서는 제안된 방법이 67%로 시소러스만을 이용한 방법의 52%와 출현빈도를 이용한 방법의 37%보다 높은 재현율을 보임을 확인할 수 있었다.

제안된 방법을 사용하면 출현빈도를 고려한 방법과 상호 연관성이 적용되지 않고 시소러스만을 사용하는 방법에서 나타나는 단점이 보완 가능하여 보다 신뢰도 높은 색인어 추출을 기대할 수 있는 것을 확인할 수 있었다. 그러나 색인어의 정확도 향상을 위해 사용한 시소러스의 반복적인 접근에 의해 수행 시간이 매우 길어지는 단점이 발견되었다. 색인어 추출 시 색인어의 정확도 향상과 수행 시간의 단축은 반비례 관계라고 볼 수 있다. 특히 시소러스를 사용하는 경우, 정확도 향상을 위한 하나의 매개변수 변경으로 인해 수행 시간이 두 배 이상 걸리는 경우도 나타났다. 이는 시소러스의 규모가 클수록 더욱 심화될 것으로 예상된다.

제안된 색인어 추출 방법은 그 성능 평가의 지표를 정확도에 두었기 때문에 좋은 성능을 발휘한 것으로 평가될 수 있으나, 빠른 시간에 응답을 보여야하는 시스템에서는 적합하지 않으리라 사료된다. 따라서 시스템의 목적에 따라 수행시간과 정확도 간의 적절한 타협점을 설정하여야 한다.

본 논문에서 제안된 색인어 추출 방법은 특정 분야의 문서 내 색인어 추출 시 해당 분야의 시소러스를 이용하므로 높은 정확도를 요구하는 연구 문헌 정보 검색 시스템이나 문서 요약 시스템에 적용될 수 있을 것으로 기대된다.

*본 연구는 인천대학교 멀티미디어연구센터와 보건복지부 한방치료기술연구개발사업의 일부 지원에 의하여 수행되었음.

## 참 고 문 헌

[1] 조태호, 서정현, "문서의 키워드 추출에 대한 신경망 접근"

[2] 맹성현, "정보검색 기술의 현황과 발전방향", 정보과학회지, 제22권, 제4호, pp. 6-14 ,2004. 4.

[3] William B. Frakes, Richard Baeza-Yates, "Information Retrieval : Data Structures & Algorithms", Prentice-Hall, 1992.

[4] H.P. Luhn, "The Automatic Creation of Literature Abstracts", IBM JOURNAL, pp. 159-165, 1958. 4.

[5] Gerard Salton, "Automatic Text Processing" Addison Wesley, 1989.