이기종 데이터 간 상호운용적 분류체계 관리를 위한 분류체계 자동화 방안

이원구* · 황명권** · 이민호** · 신성호** · 김광영** · 윤화묵** · 성원경** · 정도헌***

The Automatic Management of Classification Scheme with Interoperability on Heterogeneous Data

Won-Goo Lee* · Min-ho Lee** · Sung-Ho Shin** · Kwang-Young Kim** · Hwa-Mook Yoon** · Won-Kyung Sung** · Do-Heon Jeon***

요 약

과학기술의 융·복합현상은 21세기 지식 기반 경제하에서 더욱 활발하게 진행됨에 따라 과학기술 분야를 적절히 분류해내고, 미래의 신성장 분야까지 포용할 수 있는 체계를 만드는 것이 결코 쉽지 않다. 특히, 이기종 도메인 간 상호운용성 확보는 정보표준화, 정보서비스 분야와 같이 복잡하고 다양하게 구성된 시스템과 콘텐츠를 운영하는 영역에서 매우 중요한 사항이다. 이에, 본 연구에서는 각 콘텐츠 관리·서비스 기관이 분류체계 간 상호운용성을 갖을 수 있도록 분류체계를 유연적으로 수용·확장하기 위한 시스템적 해결방안을 제시하고자 한다. 특히 두 개 이상의 상이한 학술정보 자원의 주제분류간에 자동화된 매칭기법을 적용하여 상호운용을 가능케 하는 방법을 제시하였다.

ABSTRACT

Under the knowledge-based economy in 21C, the convergence and complexity in science and technology are being more active. Interoperability between heterogeneous domains is a very important point considered in the field of scholarly information service as well information standardization. Thus we suggest the systematic solution method to flexibly extend classification scheme in order for content management and service organizations. Especially, This paper shows that automatic method for interoperability between heterogeneous scholarly classification code structures will be effective in enhancing the information service system.

키워드

분류체계, 주제분류, 주제어 자동분류, 상호운용성

Key word

Classification Scheme, Topic Classification, Automatic Keyword Classification, Interoperability

* 정회원 : 한국과학기술정보연구원 정보기술연구실(주저자, wglee@kisti.re.kr)
** 정회원 : 한국과학기술정보연구원 정보기술연구실
*** 정회원 : 한국과학기술정보연구원 정보기술연구실(교신저자, heon@kisti.re.kr)

접수일자 : 2011. 10. 11
심사완료일자 : 2011. 11. 24

Ⅰ. 서 론

과학기술의 융·복합현상은 21세기 지식기반경제 하에서 더욱 활발하게 진행되어져 왔고, 국가별·연구주체별로 다양한 과학기술의 융·복합 관련 연구들이 진행되면서 매우 복잡한 양상으로 다변화되어 오늘에 이르고 있으며, 향후에도 더욱 다양한 분야로 분화되고 융합될 것으로 쉽게 예측할 수 있다. 따라서 이러한 과학기술 분야를 적절히 분류해내고, 미래의 신성장 분야까지 포용할 수 있는 체계를 만드는 것이 결코 쉽지 않은 일임은 자명하다[1][2]. 즉, 기존의 과학기술 콘텐츠의 분류체계는 매우 빠르게 증가하고, 변화하는 융·복합 기술 및 신기술을 모두 담아내는데 그 한계를 가지고 있다[3].

이에, 많은 연구에서는 과학기술의 융·복합 현상을 동적인 흐름으로 파악하고, 기술 융합형태에 따른 특성을 분석하여 종합적인 관점에서 새로운 분류체계를 제안하는 것에서부터, 다양한 과학기술분류체계를 아우를 수 있는 기본 개념체계를 만들고, 이를 통해서 분류체계 간의 상호운용성을 점검할 수 있는 방법론을 개발하는 것, 특정 분류체계에 대한 분류규정 및 분류기법의 개선방향을 모색하는 것 등 분류체계 문제를 해결하기 위한 다양하고도, 많은 노력을 기울여 왔다. 이것은 과학기술계 뿐만 아니라 정보서비스 분야의 전반적인 요구이기도 하고, 정보유통 개발 및 서비스 업무를 보다 체계적으로 수행하기 위해서는 분류체계가 일관된 체계로 재편성하거나 새롭게 개발할 필요가 있다는 것을 충분히 인식하고 있다. 하지만 이러한 필요가 발생할 때마다 새로운 분류체계를 만든다면 그것은 또 하나의 분류체계가 추가되고, 관리·적용할 대상이 하나 더 늘어나는 결과를 초래하여 문제를 더 어렵게 만들 것이다[4][5][6].

이에 본 연구에서는 분류체계 문제의 해결을 분류체계에 대한 새로운 체계 또는 개선된 체계를 제안하는 것이 아니라, 다른 각도(시스템을 통한 상호 교환 방식(체계)에서 분류체계 관리방식을 조명하고자 한다. 즉, 위와 같은 노력에도 불구하고 과학기술 콘텐츠 분류체계는 지속적으로 개선되고, 새로운 분류체계가 공표될 것임은 자명하다. 다만, 이는 수용하는 기관이 이를 어떻게 적용할 것인가의 문제로 봉착될 것이며, 본 연구는 각 기관이 분류체계 간 상호운용성을 갖을 수 있도록 분류체계를 유연적으로 수용·확장하기 위한 시스템적 적용(수용)방안 내지는 해결방안을 제시하고자 한다.

아울러, 대용량 학술정보를 구축하고 이를 서비스할 경우에 흔히 발생할 수 있는 문제로, 여러 종류의 분류체계가 상호운용성이 확보되지 않은 채 별도로 구축되고 운영될 경우, 학술정보의 통합서비스 품질에 영향을 주게 된다. 특히, 학술정보의 주제 분류체계와 같이 항목이 많고 상호 관계가 복잡한 경우에는 상이한 분류체계간의 의미해석이 매우 어려우므로, 이를 해석하기 위해 자동화된 기법을 적용할 수 있다면 매우 의미가 있을 것이다. 이를 위해 두 개의 이질적인 분류체계로 구축된 학술논문 정보를 이용해 분류 체계간 매칭 테이블을 자동적으로 작성하는 방법을 통해 이기종 도메인간의 상호운용을 위한 자동화 방안을 제시하고자 한다. 추가적으로 대용량의 문서를 학습함에 있어 자질(저자키워드) 축소기법에 의존하지 않고 대량의 문서를 자유롭게 학습하고 부분적인 자질추가 변경 시에 변경요소만을 효과적으로 반영할 수 있는 범용적이고 일반적인 분류기의 구조설계 방법을 보이고자 한다.

Ⅱ. 기존 분류체계

일반적으로 분류체계에 대한 관리는 각 문헌(article)의 메타데이터에 각 분류체계의 항목을 기술하는 정도의 수준이다. 이러한 관리 형태는 분류체계의 개선(version-up) 시, 버전 이력 내지는 버전별 증적자료 관리를 어렵게 하며, 이는 기존의 분류체계 형태로 구분지어진 문헌에 대한 서비스(검색)를 더욱더 어렵게 하고 있다. 더욱이 새로운 분류체계의 도입은 기존 분류체계 적용 문헌에 대한 관리 및 서비스에 치명적인 악영향을 미치게 된다.

우선, 표 1과 같이 기존의 분류체계 관리, 즉 메타데이터 상의 분류체계 표기에 대해 살펴보도록 한다. 본 예제는 한국과학기술정보연구원에서 최근에 개발 중인 관리시스템에 포함된 학술논문/학위논문/특허/연구보고서/동향·분석/산업·표준 콘텐츠의 메타데이터 상에 나타난 분류체계 관련 항목을 나타낸 것이다. 학술논문은 분류체계가 너무나 많아 기술하지는 않지만, 단순히 분류코드로만 관리하고 있는 상황은 다른 콘텐츠와 마찬가지였다.

특허(이미 국제적으로 버전관리 활동을 수행 중임) 콘텐츠에 한해 <분류명>과 <분류버전명>을 통해 버전(이력)관리를 하고 있는 것으로 나타났지만, 대부분의 문헌의 분류체계는 <분류명>과 <분류코드> 형태로만 관리되고 있었으며, 이러한 형태의 관리는 본 조사기관뿐만 아니라 대부분의 콘텐츠 관리 및 서비스 기관에서도 다음과 같은 형태로 관리되고 있다[2]. 이러한 관리의 문제점을 요약하면 다음과 같다.

표 1. 기존의 분류체계
Table. 1 An existing Classification Scheme

학위논문 DB : 주제분류코드, CC (ex, C64, S56, S12 등) XML : article-meta/article-categories/subj-group@subj-group-type/subject@content-type, <classification, dimd>
특허 DB : 대표IPC분류코드, RPRST_IPC_CLSFC_CD. IPC분류버전명, IPC_CLSFC_VRSN_NM XML : bibliographic-data/classification-ipc/main-classification, further-classification. bibliographic-data/classification-national@country/main-classification.bibliographic-data/classification-ipc/main-classification.bibliographic-data/classification-ipc/edition
국가 과학기술 표준 분류 DB : 과학기술표준분류명, SCTEC_STND_CLSFC_CD XML : article-meta/article-categories/subj-group@subj-group-type/subject@content-type, <classification, stcd>
산업표준 DB : 규격분야코드, STND_FLD_CD XML : article-meta/article-categories/subj-group@subj-group-type/subject@content-type
동향분석 DB : 기술산업분류코드, TREND_ANLS_SUBJ_CLSFC_CD. 규격분야코드, 차세대분류코드, NXGEN_TECH_CD XML : article-meta/article-categories/subj-group@subj-group-type/subject@content-type, <classification, trend_theme | tech | ngtech ...>

① 다른 분류체계/같은 분류체계의 새로운 버전의 입수에 따른 기관 표준 분류체계와의 새로운 매핑 작업을 상황이 발생할 때마다 수동으로 진행해야 하며, 일괄적으로 관리 데이터베이스를 업데이트 또는 마이그레이션을 통한 변환을 수행해야 한다.

② 다른 분류체계/같은 분류체계의 새로운 버전의 입수에 따른 기관 표준분류체계로의 적용이 어려우며(매핑의 어려움) - 기관 표준분류체계는 과거의 분류체계임을 감안 - 이를 해결하기 위한 기관 표준분류체계의 변경/확장 또한 어렵고(임의적 분류 또는 기타 분류로 처리해야 함), 마찬가지로 임의 분류된 문헌에 대한 서비스는 더욱 어렵다.

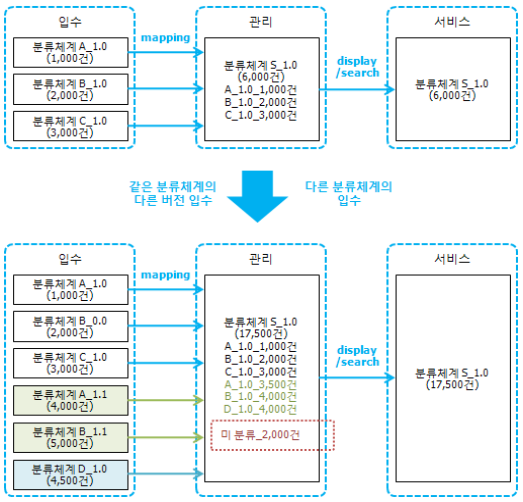


그림 1. 입수 분류체계 변경의 문제점
Fig. 1 Issues on modification of CS to acquisition

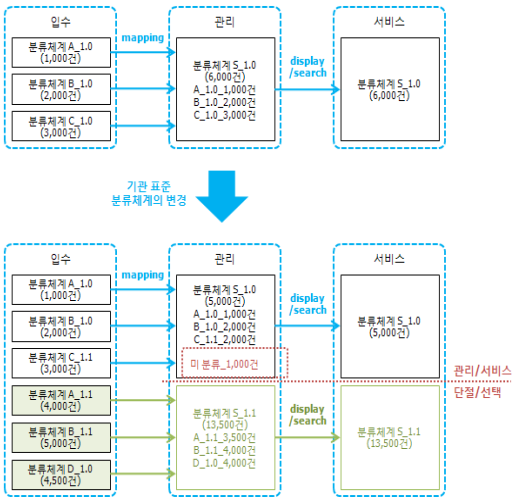


그림 2. 관리 분류체계 변경의 문제점
Fig. 2 Issues on modification of CS to management

표준 분류체계의 변경에 따른 기존 관리방식에서 나타날 수 있는 문제점은 다음과 같다.

① 우선, 관리 분류체계가 변경됨에 따라 입수 분류체계가 같은 문헌에 대해서도 다른 분류체계(새로운 관리 분류체계)로의 매핑(mapping)이 발생할 수 있으며, 기존 관리 분류체계로 분류된 문헌과 새로운 관리 분류체계로 분류된 문헌을 동시에 관리함으로써 관리체계의 혼선을 야기할 수 있다. 또한, 상기 방식의 문제점과 마찬가지로 - 기관 표준분류체계는 과거의 분류체계임을 감안 - 이를 해결하기 위한 기관 표준분류체계의 변경/확장 또한 어렵고(임의적 분류 또는 기타 분류로 처리해야 함), 마찬가지로 임의 분류된 문헌에 대한 서비스는 더욱 어려울 것으로 전망된다.

② 관리 분류체계의 상이함은 서비스에도 영향을 미치게 된다. 즉, 기존 관리 분류체계로 분류된 문헌에 대해 서비스(검색) 또한 같은 서비스 분류체계를 가지고 있다. 하지만, 새로운 관리 분류체계가 적용된다면, 서비스 또한 새로운 관리 분류체계를 준용하여 서비스 체계를 가지고 갈 것이다. 이 때, 기존의 관리 분류체계로 분류된 문헌에 대한 서비스는 미실행되거나, 실행을 위해 새로운 분류체계로의 일괄 업데이트(대량의 데이터의 경우 매우 어려움)를 수행해야만 할 것이다.

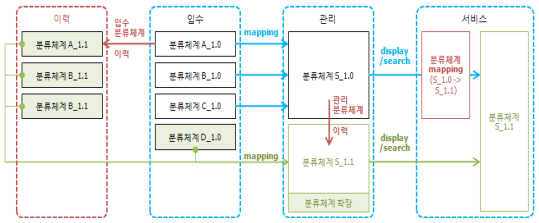


그림 3. 입수/관리 분류체계 개념도
Fig. 3 Diagram on acquisition/management CS

이에, 본 연구에서는 같은 입수 분류체계의 다른 버전 입수 및 다른 분류체계의 입수 문제, 그리고 다른 입수 분류체계의 입수 문제, 기관 표준 관리 분류체계의 변경 문제 등 상기에 기술된 다양한 문제에 대해 다음에 기술하는 바와 같은 형태로 해결책을 제시하고자 한다.

Ⅲ. 자동 분류체계

1. 단위 분류기 기반 대용량 데이터 자동 분류

상기와 같이 기존 관리체계는 많은 문제점을 드러내고 있다. 하지만 더 큰 문제는 관리 분류체계 코드가 부여되지 않은 채 관리되고 있는 수천만건의 학술정보 콘텐츠에 대한 관리 분류체계 코드를 일관되게 부여하는 것이다. 이는 기존의 수동적인 분류체계 부여 방식(기사 주제분류 DB구축 사업 등으로 진행)으로는 해결할 수 없는 사안이다. 하지만, 관리 분류체계 코드가 수천만건(2011년 9월 기준, 약 7천만건 수준)에 달하는 모든 콘텐츠에 부여된 후에야, 입수・관리 분류체계 이력관리 및 분류체계 간 매핑에 대해 논할 수 있을 것이다. 이에 본 논문에서는 자동 분류를 위해 대량의 데이터를 학습함에 있어 자질축소 기법에 의존하지 않고, 자유롭게 학습하고 부분적인 자질추가 변경 시에 변경요소만을 효과적으로 반영할 수 있는 범용적이고 일반적인 분류방안을 제시하고자 한다.

1.1 단위 분류기 생성과 동적 결합

효율적인 데이터(문서) 분류처리를 위해 자질선택 기법을 사용하는데, 이는 정보량의 축소뿐만 아니라 성능의 향상을 위해서도 필요한 과정으로 알려져 있다. 그러나 대용량의 문서학습을 하는 작업에서는 과도한 비율 이상으로 자질을 제거하는 과정이 성능에 영향을 끼치게 되므로 자질선택 및 축소기법의 적용 역시 한계가 존재하게 된다. 일반적으로 학습문서의 수나 자질의 수에 대한 고려는 분류기의 생성효율과 관련이 있다. 최적화된 문서와 자질을 이용할 때 빠른 처리와 함께 높은 성능을 낼 수 있기 때문에 이와 관련한 여러 기법을 연구하고 적용하고 있지만, 정보량을 축소하는 보다 근본적인 이유는 분류기 생성에 소요되는 시간과 실제 메모리 점유의 문제로 인해 대량의 문서를 학습할 수 없는 제한점이 존재하기 때문이다.

문서의 기계학습에 되도록 많은 문서와 많은 자질을 이용하면 성능이 좋아진다는 기본 가설에는 변함이 없으므로, 본 기술을 통해 대용량 정보의 기계학습 시 고민해야하는 정보의 차원축소 문제라는 제약사항을 해소할 수 있는 방법을 제공하고자 한다. 즉, 이 기법은 정보를 축소하는 것이 아니라, 실제로 대용량의 매트릭스를 생성하는 것과 작은 용량의 매트릭스를 다수 생

성하여 동적으로 결합하는 두 가지 분류기 생성방법에 있어 학습결과의 수치상 차이가 전혀 없도록 하는 방법이다.

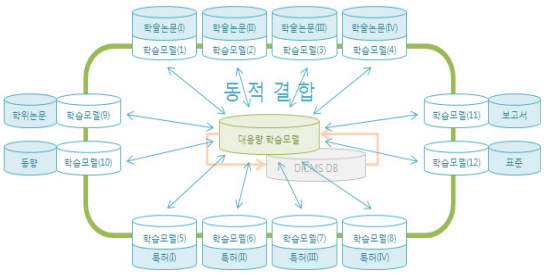


그림 4. 대량 분류기 생성 개념도
Fig. 4 Diagram on classification of a large data

그림 4는 데이터베이스별로 여러 개별 분류기를 조합하는 예시이며, 개별 데이터베이스도 구성문서의 수가 많을 경우, 여러 개의 분할된 복수의 분류기로 구성하여 동적으로 결합하여 최종 분류기를 생성할 수 있다.

1.2. 단위 분류기의 생성과정

단위 분류기의 생성을 위해 아래와 같은 전처리 과정을 포함한 일련의 과정을 거친다.

(1) 자질 추출

자질을 추출하기 위해 아래의 두가지 타입을 고려할 수 있다. 타이틀, 초록 등으로부터 정보를 추출하는 경우에는 스테밍(영문) 또는 형태소분석(한글)을 거쳐 자질 집합을 생성한 후 자질축소의 과정을 고려하는 것이 좋다. 또한, 전체문서 집합에서 저빈도(CF=1) 자질은 제거한다.

① 키워드, 디스크립터 : 논문 저자의 키워드 필드나 통제어휘인 디스크립터 필드를 이용한다.
② 용어 추출(Info Extraction) : 타이틀, 초록 등의 비구조적인 정보로부터 명사구를 포함한 주요 정보를 추출한다.

(2) 문헌별 자질정보 추출 및 생성

문헌을 구성하는 개별자질에 범주코드를 부여한다.
*주요 생성필드 : 문헌고유ID, 자질, 범주코드

(3) 자질 특성 매트릭스 생성

최종 자질 벡터를 연산하기 위한 매트릭스 정보를 생성하여 DB나 바이너리 파일로 적재한다.
* 주요 생성필드 : 자질고유ID, 자질, 범주코드, TP, TN, FP, FN, CF, IDF 등 (표 2)

표 2. 자질-범주간 출현관계 분할표
Table. 2 emergence relation between feature-category

	범주 cj 소속	범주 cj 미소속
자질 fi 출현	TP	TN
자질 fi 미출현	FP	FN

1.3. 단위 분류기 결합을 통한 대용량 분류기 생성

대용량 분류기 생성의 핵심은 단위 분류기 생성 프레임워크 단계 중 3단계에서 생성된 자질특성 매트릭스를 결합하는 방법을 이용해 분류기의 동적결합의 수행하는 것이다. 단위 분류기는 일반적으로 도메인별로 생성될 수 있으나, 학습할 대상문헌이 많은 도메인의 경우에는 적당한 크기로 자유롭게 생성하여 필요한 경우 동적으로 결합해 거대한 매트릭스를 재생산할 수 있다.

(1) 매트릭스 동적결합 수행

① 우선 복수개의 결합 대상 '자질특성 매트릭스'를 메모리에 상주하여, 모든 매트릭스에 출현한 자질값의 고유한(distinct) 전체 셋을 만든다.
② 개별 자질에 결합 대상 매트릭스들을 참조하여 정보를 가져온다. 이때, 자질이 모든 자질특성 매트릭스에서 출현하지 않으므로 자질의 개수, 전체 문헌의 수 등 각 매트릭스의 통합정보를 동적으로 산출하여 TP, TN, FP, FN과 IDF, CF 등 주요 정보를 재계산한다.

(2) 개별 자질에 대한 주제-가중치 벡터를 생성

통합된 자질 특성 매트릭스로부터 거리계수 및 Cosine, LOR 등 유사척도를 이용해 최종 투표분류기에 적합한 자질 벡터형태를 생성하여 DB나 바이너리 파일로 적재한다. LogTF*IDF*Cosine 계수를 이용한 자질벡터는 아래와 같이 표현이 가능하다[7].

$$vs(ficj) = (1 + logtf) \times log(N/df) \times cos(ficj)$$

(3) 문헌범주화 수행

통합 매트릭스에서 생성된 자질벡터를 이용해 투표형 분류기법으로 분류를 수행한다. 자질값 투표형 분류기(Feature-value Voting Classifier: FVC)는 여러 관련연구를 통해 수행되었다[7][8][9]. 생성된 자질 벡터를 메모리에 상주한 후, 대량의 입력문헌에 대해 고속의 다원분류를 수행하여 입력문서를 분류한다.

최종 생성된 분류기는 최종 계산된 벡터의 데이터량이 상대적으로 많지 않아 메모리 상주용량이 적기 때문에 자질 종수의 제한이 없으며 각 가중치의 선형결합을 실시하므로 자질 종수의 증가에 따른 속도저하도 거의 없는 고속의 분류기이며, 이에 대한 실험도 병행하여 수행하였다[[10].

2. 분류체계 간 자동 매핑

2.1 분류체계 파일의 자동 파싱

일반적으로 콘텐츠를 관리・서비스하는 기관에서는 여러 입수처(기관)로부터 다양한 형식의 콘텐츠를 입수하게 된다. 이 때 입수되는 콘텐츠의 형식(Text, XML, MARC, MARC XML, Excel, Tagged Text, SGML, HTML, WORD, HWP 등)은 다양하며, 마찬가지로 분류체계의 입수 형식 또한 너무나 다양하고, 향후에는 더욱더 형식의 다양성이 증대될 것이다.

이에, 다양한 형식의 분류체계 입수 형태로부터 자동으로 분류체계 구조를 파싱하여 입수 분류체계 트리(tree)를 생성하여 주는 자동 파서(AP, Automatic Parser)에 대한 개발이 필요하고, 이는 기관에서 채택한 관리 표준 분류체계의 입수 시에도 반드시 필요하며, 이와 관련 입수/관리 분류체계 파일의 자동 파싱 프로세스는 다음과 같다.

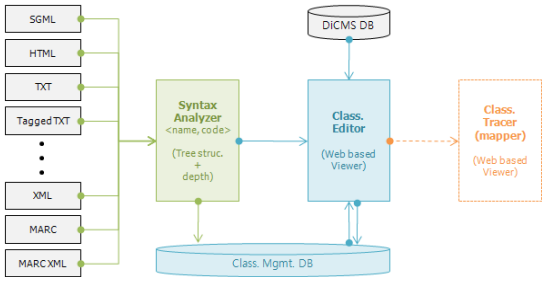


그림 5. 분류체계 파일의 자동 파싱 프로세스
Fig. 5 Automatic parsing process of CS files

2.2 분류체계 간 자동 매핑

본 절에서는 이질적인 분류체계를 사용하는 학술정보 간의 관계를 확률적인 강도로 표현하여 그 관계를 추론하고, 분류체계 간 자동 매핑하는 방안에 대한 기술하고자 한다. 이와 유사하게 메타데이터에 기반하여 정보시스템간의 의미 유사도를 측정하려는 시도가 있었으며[11], 또한 단일 분류체계 내의 각 분류 간에 의미적인 유사성을 산출하여 유사주제분류의 상호 의미관계를 확률강도로 표현하려는 확률적 온톨로지 기법에 관한 연구도 최근 수행되었다[12].

분류체계명(자질)의 주제분야(범주)간 유사도를 측정하기 위하여, 고빈도어 선호경향을 갖는 연관성 척도인 코사인 유사계수를 사용하였다(표 1). 유사계수 결과값은 모두 가중치 부여방식으로 산출된 것으로 0과 1사이의 값을 갖는다.

$$\text{Cosine 계수} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

표 3. 자질(분류체계명)와 범주(주제분야)간 2x2 분할표
Table. 3 2x2 Div. between feature-category

	범주 cj 소속	범주 cj 미소속
자질 fi 출현	a	b
자질 fi 미출현	c	d

표 3의 자질 f는 분류체계명에 해당하며, 범주 c는 분류체계명이 속한 주제분류를 의미한다. 자동분류 시, 자질값(자질과 범주의 연관도) 투표방식을 사용하는데 분류대상 문서에 나타난 n개의 단어 자질집합과 후보범주 m개의 집합을 각각 F={f1, f2, …, fn}와 C={c1, c2, …, cm}로 표현하고, 자질 fi가 범주 cj에 대해서 가지는 자질값을 V(fi, cj)라고 하면 자질값 투표 분류기는 다음 공식을 만족하는 범주 cj를 문서에 할당한다[9].

$$\arg\max_{c_j \in C} \sum_i V(f_i, c_j)$$

이러한 투표형 퍼셉트론(VPT: Voted Perceptron) 방식은 기본적인 신경망 모형 중 하나인 퍼셉트론의 결과를 다수결 투표 방식으로 출력하는 분류방법으로서,

성능이 좋은 분류기로 알려져 있는 SVM와 비교하여 거의 대등하거나 약간 떨어지는 성능을 보이면서도 계산상의 복잡성이 상대적으로 낮고 처리속도가 빠르다는 장점을 가지고 있다[1]. 제안한 방식에 대한 증명을 위해 실제 대용량 데이터 처리를 위한 VPT 방식의 분류기를 직접 개발하였고, 이에 대한 실험을 실시하였다[12].

2.3 실험 및 평가

본 절이기종 분류체계 간의 유사강도를 확률적으로 추출하여 두 범주간 매칭테이블을 자동으로 작성하기 위해 다음과 같이 실험하였다.

① 최초에 구축된 정보서비스 자원은, 두 학술정보간 분류체계가 상이하므로 상호 주제해석이 불가능한 상태임

② 자질로서 영문저자 키워드, 범주로서 논문의 개별 분류정보를 추출

③ 코사인계수를 이용하여 자질(키워드)-범주(분류코드)간 유사값을 산출하고 VPT 방식으로 키워드별 후보주제분야를 유사가중치의 합으로 표현함

④ 이기종 분류체계인 정보원(중국 학술정보)로부터 키워드와 분류정보를 추출함

⑤ 학습 시스템에 이기종 환경에서 추출한 자질을 매칭하여 두 이질적인 분류체계간 유사가중치 값을 추출함

⑥ 두 분류체계간 유사가중치를 일련의 벡터값으로 표현하여 최종 추천 분류코드를 확률적 온톨로지 형태로 제시함

실험결과 중국학술정보의 Q969.42를 분류코드가 1867896, 1866925, 1905787번 논문에서 각각 발생했다

표 4. 이기종 학술정보 분류체계간 매핑 테이블 결과
Table 4. The results of mapping between heterogeneous contents

순번	중국학술 분류체계	대/중분류 내용 (1자리 대분류, 2자리 중분류임)	발생 빈도	KISTI 표준분류에 대한 매칭률(%)	매칭된 분류 내용	일치여부 (순위)
1	TP391	Automation , Computer Engineering	5998	NA:34.59, ET:9.08, LA:8.45	전산	1
2	O4	Physics	5224	CA:17.21, ET:13.77, EC:9.23	X	X
3	R	Medicine and Health Sciences	4684	BM:44.52, BD:12.38, BF:6.34	의학	1
4	TP3	Automation , Computer Engineering	4408	NA:35.53, EJ:9.61, ET:9.39	전산	1
5	O6	Chemistry	4293	CA:21.8, BD:11.79, BF:11.59	화학	1
6	TP393	Automation , Computer Engineering	4065	NA:35.67, NB:21.38, ET:10.43	전산	1
7	TP311	Automation , Computer Engineering	3276	NA:36.68, EJ:9.92, ET:7.73	전산	1
8	N	Natural Science	2613	LA:12.21, NA:5.88, AA:4.6	수학,전산,건설(?)	△
9	R6	Surgery	2530	BM:41.54, BD:2.96, BC:2.93	의학	1
10	TP391.9	Automation , Computer Engineering	2491	NA:14.29, ET:5.0, ND:4.25	전산	1
11	TP273	Automation , Computer Engineering	2279	EC:11.32, NA:9.28, MA:8.58	전기공학	1
12	TP18	Automation , Computer Engineering	2115	NA:23.53, LA:8.11, ET:3.45	전산	1
13	O1	Mathematics	1971	LA:47.24, EJ:8.61, ET:3.8	수학	1
14	R81	Radiology, Sport medicine, Diving medicine, Aerospace medicine	1933	BM:39.56, PA:4.35, BA:1.61	의학	1
15	TP391.41	Automation , Computer Engineering	1902	NA:20.67, LA:6.14, TB:4.79	전산	1
16	R318	Human anatomy, Physiology, Pathology, Microbiology, Parasitology	1618	BM:11.09, BB:7.8, BD:5.69	의학	1
17	O41	Physics	1432	CA:11.91, LA:8.88, ET:5.84	X	X
18	P208	Geodesy	1431	NA:12.16, AC:8.16, EJ:4.28	전산	2
19	X703.1	Waste Management and Recycling	1429	AE:19.17, BB:11.11, CB:10.58	환경공학	1
20	P4	Meteorology	1421	RB:17.04, AE:8.99, AA:5.47	환경공학	2

면, 각각의 키워드로부터 학습시스템에 매칭한 결과, 분류값과 유사계수값은 가중치값으로 나타낼 수 있다. 즉, 1867896번 논문은 Q969.42 분류코드를 가지는데, 확률적 온톨로지로 Q969.42 = BH08:0.57, BA05:0.16 ... 와 같이 표현할 수 있다. 최종적으로 모든 자질값을 투표한 결과, Q969.42 분류코드는 BH08 코드와 75% 수준으로 매칭되며, 2순위이하 BA05와 BA01에는 각각 17%, 8% 수준으로 매칭될 확률을 보였다.

이러한 확률적 온톨로지 방법론은 기존의 정보검색이나 데이터마이닝 분야에서 개발된 통계적 연관성 측정방식을 이용하여 대상 범주간 연관성을 통계적, 확률적으로 파악하여 도출하는 방법이지만 아직 분명한 정의가 제시되지는 못하고 있다. 단 기존의 온톨로지에서 개념간의 관계가 확정적인 것과는 달리 확률적으로 연결강도가 표현되는 점이 다르다.

이기종 학술정보 분류체계간 매핑한 실험의 최종결과 중국분류코드는 1자리 숫자는 대분류, 2자리 숫자로 표현된 것은 중분류레벨이다. 그 이하로 세분류까지 표현되어 있다. 이에 대해 확률적인 형태로 KISTI 표준분류체계(해외학술정보의 분류체계)를 표현하고 상호 일치여부를 확인하였다. 상위 20개의 분류에서 2개는 불일치하였으며, 명확하지 않은 1개 항목이 있었다. 불일치한 항목은 모두 물리학이었으며, 중국학술 분류의 자연과학의 범주가 KISTI 표준체계상에 존재하지 않는 관계로 매칭관계가 정확하게 이루어지지 않았음을 알 수 있다. 그러나 전체 중국 분류코드의 여러 레벨에서 모두 일관성있게 KISTI 주제분야가 할당되었으므로 상호운용의 일관성이 나타나고 있어 이기종 분류간 자동매핑 가능성을 확인할 수 있었다.

현재 발생빈도 상위 20개의 결과를 매칭하였는데, 전체 중국학술논문수가 492,136개였고 이중 11.6%인 57,113개의 논문분류가 상위 20개 매칭만으로 해결되었다. 상위 50위까지 확대할경우에는 18.9%인 93,345개, 상위 100위인 경우 27.7%인 136,485개 논문의 분류를 자동매칭할 수 있다. 만약 중국분류체계의 매핑수준을 중분류로 할 경우에는 코드체계가 매우 단순해지므로, 자동매칭성능과 효율성이 매우 증대될 수 있을 것으로 기대한다. 향후 추가실험을 통해 이질적인 두 시스템간 최적의 매칭 레벨을 찾아내는 과정을 수행해야 할 것이다.

Ⅳ. 결론 및 향후 연구

지금까지 새로운 분류체계 관리를 위한 여러 측면에서의 부분별 설계를 시도하였다. 우선, 입수/관리 분류체계의 입수에 있어서 기존에는 웹 상에서 view하거나, 문서 또는 데이터베이스 형태로 전달받아 이를 콘텐츠 관리/서비스 기관에서 문서 또는 데이터베이스 형태로 관리하고 있는 실정이다. 이는 인력/예산 낭비뿐만 아니라 관리적인 측면에서 많은 어려움을 초래하였다. 이에, 기존의 수동적인 입수절차를 탈피하여 파서(parser)에 의한 자동 파싱(parsing) 기능 및 이를 위한 수정/삽입/삭제 기능을 설계하였다. 이를 통해 향후에 입수・적용되는 분류체계에 대해 손쉽게 자동으로 입수・관리할 수 있는 기반을 제공할 수 있었다.

두 번째는 매핑에 있어, 기존의 자동범주화 모델에서 제한점으로 존재하였던 대용량 문서의 처리와 학습결과의 재사용 문제를 해결함으로써 학습대상 문서의 추가 변경 시 전체 데이터를 반복적으로 처리해야 하는 텍스트 마이닝의 취약점을 보완한 대용량 기반의 동적 분류기 생성방안을 제시하였다. 또한 성능이 우수한 것으로 알려진 SVM과의 베이스라인 성능 비교를 통하여 새롭게 제안한 모델의 가능성을 검증하였다.

끝으로 새로운 분류체계의 적용에 있어서 기존의 관리・서비스 체계에서는 새로이 입수되는 콘텐츠에 대해서만 새로운 분류체계가 적용가능하거나, 기존의 분류체계에 대해 일괄적인 갱신(update)를 통해서만 가능하다. 하지만, 제안된 방식을 통해서는 입수-관리체계와의 매핑 또는 관리체계 버전-업(version up) 시에, 새로운 콘텐츠에 자동으로 새로운 관리체계 코드가 적용되고, 기존의 콘텐츠에 대해서는 별도의 갱신이 없이 매핑 테이블 관리를 통해서 만으로도 관리 및 서비스가 가능해질 수 있다.

그리고 이질적인 분류코드 간의 의미관계를 확률적인 강도로 표현하는 자동기법을 제안하였다. 이기종 도메인간 의미관계를 파악하기 위해 자동기법으로 매핑 테이블을 생산한 후 전문가의 검증과정을 거치는 프로세스를 적용한다면 매우 비용효과적인 상호 매핑 테이블을 작성할 수 있을 것이다. 본 연구를 통해 이질적인 학술정보 분류체계 간에 유사성을 확률적으로 산출하여 매핑을 자동화할 수 있는 방안을 마련하였으나 전체적인 성능평가 부분은 수행하지 않았다.

향후, 성능을 향상시킬 수 있는 자질 가중치 요소를 모델에 추가할 예정이며, 다양한 유형의 데이터를 기반으로 기존의 분류모델과의 비교검증을 추가로 수행하여 일반화된 성능검증 결과를 도출할 예정이다. 또한, 형태소분석기를 통해 색인어를 선정하는 자질선정 방법과 학습환경을 확대하거나 유사어 확장 등을 통해 저자키워드의 매칭률을 높이기 위한 다양한 방법론을 적용하여 매칭성능을 높이기 위한 연구도 의미가 있을 것이다.

참고문헌

[1] 홍성화, 서태설, "분류체계 일치를 통한 과학기술 정보 상호 교환 방법에 관한 기초 연구", 정보관리연구, 제35, 제2호, pp.109-123, 2004.

[2] 오용선, 한정수, "디지털콘텐츠 기술 분류체계 및 분과기술 소개", 한국콘텐츠학회논문지, 제7권 제4호, pp36-71, 2009.

[3] 황다영, 김영인, 이병민, "기술융합 특성에 따른 새로운 분류체계의 제안", 기술혁신학회지 제11권, 제4호, pp.592-612, 2008.12.

[4] 김비연 "DDC의 학제적 주제 분류에 관한 연구", 한국문헌정보학회지, 제45권, 제1호, pp.333-351, 2011.

[5] 백지원, "주제어 기반 분류에 관한 연구: 미국 공공도서관의 사례를 중심으로", 한국문헌정보학회지, 제44권, 제4호, pp179-201, 2010.

[6] 백지원, "이용자 중심의 주제어 기반 분류를 위한 주제명 개발에 관한 연구: 지식조직체계 분석을 바탕으로", 정보관리학회지, 제28권 제1호, pp.171-192, 2011.

[7] 정도헌, "최대 개념강도 인지기법을 이용한 데이터베이스 자동선택 방법에 관한 연구", 정보관리학회지, 제27권, 제3호, pp.265-281

[8] Ko. Y, J. Seo, "Using the feature projection technique based on a normalized voting method for text classification", *Information Processing and Management,* vol.40, no.2, pp.191-208, 2004

[9] 이재윤, "문서측 자질선정을 이용한 고속 문서분류기의 성능향상에 관한 연구", 정보관리연구, 제36권, 제4호, pp.51-69, 2005

[10] 정도헌, 황명권, 성원경, "대용량 문서학습을 위한 분류기 생성 및 결합방법", 한국정보처리학회 춘계학술대회 논문집, 제18권, 제1호, pp.1551-1554, 2011.5

[11] 임정은, 최오훈, 나홍석, 백두권, "메타데이터 기반 정보 시스템 간 의미 유사도 측정 방법", 2006 한국컴퓨터종합학술대회 논문집, 제33권, 제1호, pp.85-87, 2006

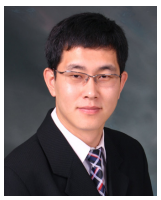
[12] 이정연, 이재윤, 정한민, 강인수, 신숙경, "확률적 온톨로지와 연구자 네트워크를 이용한 심사자 자동추천에 관한 연구", 정보관리학회지, 제24권, 제3호, pp.43-65, 2007

저자소개

이원구(Won-Goo Lee)

2000년 한남대학교 대학원(석사)
2005년 한남대학교 대학원(박사)
2005년~현재 KISTI 선임연구원

※관심분야: 데이터베이스 지식관리, 과학데이터



황명권(Myung-Gwon Hwang)

2006년 조선대학교 대학원(석사)
2011년 조선대학교 대학원(박사)
2011년~현재 KISTI 선임연구원

※관심분야: 텍스트마이닝, 지식베이스, 의미식별

이민호(Min-Ho Lee)

2000년 충남대학교 대학원(석사)
2006년 충남대 대학원(박사수료)
2001년~현재 KISTI 선임연구원

※관심분야: 시맨틱 웹, 빅데이터, 지식관리

신성호(Sung-Ho Shin)

2000년 경북대학교(학사)
2002년 경북대학교 대학원(석사)
2002년~현재 KISTI 선임연구원

※관심분야 : 데이터통합, 데이터품질, IS평가

김광영(Kwang-Young Kim)

2001년 부산대학교 대학원(석사)
2011년 충남대학교 대학원(박사)
2001년~현재 KISTI 선임연구원

※관심분야 : 정보검색, 아카이빙, 개인화 검색

윤화묵(Hwa-Mook Yoon)

1997년 공주대학교 대학원(석사)
2008년 배재대학교 대학원(박사)
1977년~현재 KISTI 책임연구원

※관심분야 : 데이터베이스 정보검색, 온톨로지

성원경(Won-Kyung Sung)

1989년 연세대학교 대학원(석사)
1996년 파리대학교 대학원(박사)
2004년~현재 KISTI 책임연구원

※관심분야 : 데이터베이스 지식관리, 과학데이터

정도헌(Do-Heon Jeong)

2003년 연세대학교 대학원(석사)
2011년 연세대 대학원(박사수료)
2003년~현재 KISTI 선임연구원

※관심분야 : 텍스트마이닝, 시맨틱 웹, 정보검색