# 이기종 학술정보 분류체계간 상호운용에 관한 연구

## A Study of Interoperability between Heterogeneous Scholarly Classification Code Structures

정도헌, 이상환, 신기정
한국과학기술정보연구원

Do-Heon Jeong, Sang-Hwan Lee, Ki-Jeong Shin
Korea Institute of Science and Technology Information(KISTI)

**요약**

이기종 도메인간 상호운용성 확보는 정보표준화, 정보서비스 분야와 같이 복잡하고 다양하게 구성된 시스템과 콘텐츠를 운영하는 영역에서 매우 중요한 사항이다. 대용량의 정보자원을 구축하고 서비스하는 정보시스템의 경우, 내부 자원간의 상호운용성의 문제가 결국 전체 서비스의 품질에 큰 영향을 주게 된다. 서로 다른 표준에 따라 구축된 자원을 통합, 연계하기 위해 자동화된 기법을 사용한다면 매우 효율적인 시스템을 구축할 수 있을 것이다. 본 연구에서는 두개의 상이한 학술정보 자원의 주제분류간에 자동화된 매칭기법을 적용하여 상호운용을 가능케 하는 방법을 제시하였다. 의미표현의 수준이 매우 상이한 두가지 분류 체계간에 자동생산된 연계 정보를 통해 보다 효율적인 정보서비스가 가능할 것으로 기대한다.

**Abstract**

Interoperability between heterogeneous domains is a very important point considered in the field of scholarly information service as well information standardization. In case of the large information system, interoperability between internal information resources becomes to affect the performance of the whole system. The automatic method for understanding heterogeneous system environment will be very helpful to solve the problems like this. This paper shows that automatic method for interoperability between heterogeneous scholarly classification code structures will be effective in enhancing the information service system.

## I. 서 론

### 1. 연구배경 및 목적

여러 정보시스템간 상호운용성 확보의 문제는 정보서비스, 데이터베이스, MDR(XMDR), 온톨로지와 시맨틱 웹 관련분야와 같은 정보표준화, 정보서비스 영역 및 차세대 응용 연구분야 등에서 매우 중요한 이슈가 되어왔다. 이기종 정보시스템간, 또는 이기종 도메인간 정보공유는 시스템간의 이질성, 구조적인 이질성, 의미체계의 이질성 등으로 인해, 서로 다른 표준에 따라 구축된 자원을 통합, 연계하는데 어려움이 따르게 된다[4].

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.

본 연구에서는 두개의 이질적인 분류체계로 구축된 학술논문 정보를 이용해 분류 체계간 매칭테이블을 자동적으로 작성하는 방법을 통해 이기종 도메인간의 상호운용을 위한 자동화 방안을 제시하였다.

## II. 관련 연구 및 방법론

### 1. 관련 연구

본 연구는 이질적인 분류체계를 사용하는 학술논문 정보간의 관계를 확률적인 강도로 표현하여 그 관계를 추론하는데 목적이 있다. 이와 유사하게 메타데이터에 기반하여 정보시스템간의 의미 유사도를 측정하려는 시도가 있었으며[4], 또한 단일 분류체계 내의 각 분류 간에 의미적인 유사성을 산출하여 유사주제분류의 상호 의미관계를 확률강도로 표현하려는 확률적 온톨로지 기법에 관한 연구도 최근 수행되었다[3].

## 2. 유사도 측정과 VPT 기법

저자키워드(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.

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

[표 1] 자질(키워드)와 범주(주제분야)간 2x2 분할표

| | 범주 $c_j$ 소속 | 범주 $c_j$ 미소속 |
|---|---|---|
| 자질 $f_i$ 출현 | a | b |
| 자질 $f_i$ 미출현 | c | d |

표 1의 자질 $f$는 키워드에 해당하며, 범주 $c$는 키워드가 속한 주제분야를 의미한다. 자동분류시 자질값(자질과 범주의 연관도) 투표방식을 사용하는데 분류대상 문서에 나타난 $n$개의 단어 자질집합과 후보범주 $m$개의 집합을 각각 $F=\{f_1, f_2, \cdots, f_n\}$와 $C=\{c_1, c_2, \cdots, c_m\}$로 표현하고, 자질 $f_i$가 범주 $c_j$에 대해서 가지는 자질값을 $V(f_i, c_j)$라고 하면 자질값 투표 분류기는 다음 공식을 만족하는 범주 $c_j$를 문서에 할당한다[2].

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

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]. 본 연구를 위해 실제 대용량 데이터 처리를 위한 VPT 방식의 분류기를 직접 개발하였다.

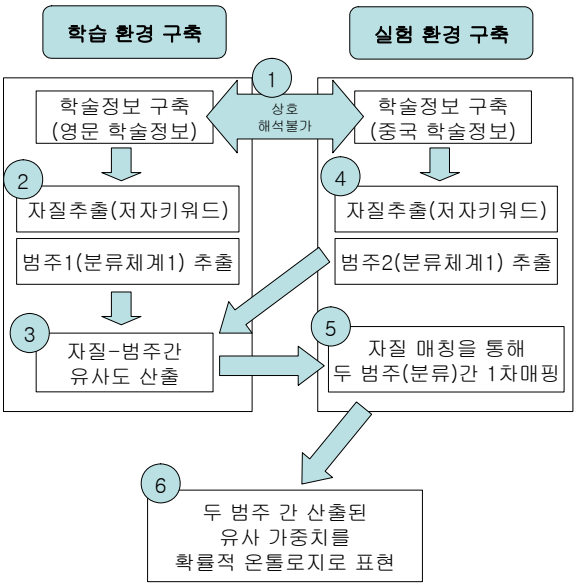
# III. 실험 및 결과 분석

## 1. 실험환경 구축

학습용 시스템 환경을 구축하기 위한 원천 데이터베이스는 KISTI 해외학술정보 데이터베이스에서 추출한 1,096,950건의 영문 학술논문이다. 자질선정을 위해 이로부터 추출한 5,551,841개의 영문키워드와 분류코드를 이용하였으며, 논문당 평균 저자키워드는 약 5.06개였다. 실험용 데이터베이스는 고유한 분류체계를 별도로 갖는 KISTI 중국학술정보 데이터베이스로 선정하고, 492,136건의 논문정보와 1,962,291개의 영문 저자 키워드(평균 3.99개)를 추출하여 구축하였다.

## 2. 실험과정

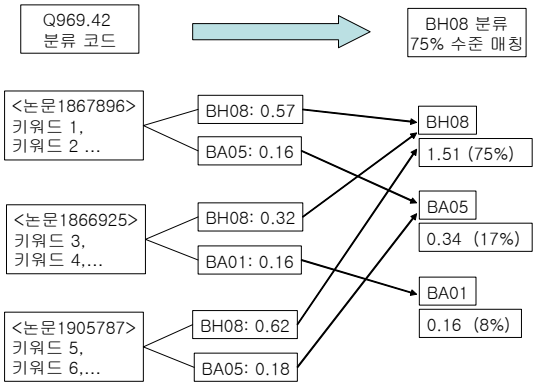
이기종 분류체계 간의 유사강도를 확률적으로 추출하여 두 범주간 매칭테이블을 자동으로 작성하기 위해 다음과 같이 실험하였다(그림 1).




▶▶ 그림 1. 실험 개요도

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

VPT 분류를 이용한 최종결과를 확률적 온톨로지 형태로 나타내기 위한 일련의 추론과정을 간략하게 그림으로 나타내면 다음과 같다.




▶▶ 그림 2. VPT를 통한 주제분야의 자질값 투표

중국학술정보의 Q969.42를 분류코드가 1867896, 1866925, 1905787번 논문에서 각각 발생했다면, 각각의 키워드로부터 학습시스템에 매칭한 결과, 분류값과 유사계수값은 그림 2와 같이 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 보이고 있다(그림 2).

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[3].

## 3. 실험 결과

실험검증은 중국학술지에서 발생한 상위빈도 20위 이상의 분류코드에 대해 1순위-3순위까지의 매칭성공 여부을 확인하였다. 중국학술정보 분류체계의 대・중・소(세)분류 코드를 과학기술표준분류코드의 대분류 코드체계에 확률적 온톨로지 형태로 매핑하였다(표 4). 표 2는 실험대상 도메인인 중국데이터의 학술정보 분류체계의 일부이다.

중국학술 분류체계(Chinese Libarary Classification)에 대한 의미매칭을 위해 사용한 KISTI의 과학기술표준분류체계는 대분류 44개, 중분류 약 250여개와 세분류까지 모두 3레벨 940여개의 분류로 구성되어 있다(표 3). 본 실험에서는 250여개의 중분류레벨에서 매칭률을 계산한 후, 자료부족(data sparseness)현상을 없애기 위해 최종적으로 대분류 수준에서 결과를 산출하였다.

[표 2] Chinese Library Classification의 분류체계(22개 대분류)

| 분류 코드 | 내 용 |
|---|---|
| A | Marxism, Leninism, Maoism & Deng Xiaoping Theory |
| B | Philosophy and Religion |
| C | Social Sciences |
| D | Politics and Law |
| E | Military Science |
| F | Economics |
| G | Culture, Science, Education and Sports |
| H | Languages and Linguistics |
| I | Literature |
| J | Art |
| K | History and Geography |
| N | Natural Science |
| O | Mathematics, Physics and Chemistry |
| P | Astronomy and Geoscience |
| Q | Life Sciences |
| R | Medicine and Health Sciences |
| S | Agricultural Science |
| T | **Industrial Technology** |
| U | Transportation |
| V | Aviation and Aerospace |
| X | Environmental Science |
| Z | General, Miscellaneous, Auxiliary and Others |

T 중분류

| 중분류 | 내 용 |
|---|---|
| TB | General Industrial Technology |
| TD | Mining Engineering |
| TE | Petroleum, Natural Gas |
| TF | Extractive metallurgy, Smelting |
| TG | Metallurgy, Metalworking |
| TH | Machinery, Instrumentation |
| TJ | Military Technology |
| TK | Power Plant |
| TL | Nuclear technology |
| TM | Electrical Engineering |
| TN | Electronic Engineering, Teleommunication Engineering |
| TP | **Automation , Computer Engineering** |
| TQ | Chemical Engineering |
| TS | Light Industry, Handicraft |
| TU | Construction Engineering |
| TV | Water Resources, Hydraulic Engineering |

TP 세분류

| 세분류 | 내 용 |
|---|---|
| TP3 | 計算技術, 計算机技術 |
| TP30 | 一般性問題 |
| TP301 | 理論、方法 |
| TP301.1 | 自動机理論 |
| TP301.2 | 形式語言理論 |
| ... | ... 이하 생략 |

[표 3] KISTI 과학기술표준분류의 대분류 코드체계

| 분류코드 | 내용 | 분류코드 | 내용 |
|---|---|---|---|
| AA | 건설공학 | EJ | 정보과학 |
| AB | 건축공학 | ET | 전자공학 |
| AC | 토목공학 | LA | 수학 |
| AD | 도시공학 | MA | 기계공학 |
| AE | 환경공학 | MB | 열유체공학 |
| BA | 생물학 | MC | 기계제작기술 및 산업기계 |
| BB | 생물공학 | MD | 수송공학 |
| BC | 생화학 | NA | 전산학 |
| BD | 약학 | NB | 정보통신 |
| BE | 수산학 | NC | 정보학 |
| BF | 식품 | ND | 정보가공 |
| BG | 농화학 | PA | 물리학 |
| BH | 농림업 | PB | 진동학 |
| BL | 축산 | RA | 금속공학 |
| BM | 의학 | RB | 자원공학 |
| CA | 화학 | RC | 에너지공학 |
| CB | 화공 | SA | 과학기술일반 |
| CC | 고분자 | SB | 경영경제 |
| CD | 요업 | SC | 사회과학 |
| CE | 섬유 | SD | 인문과학 |
| EC | 전기공학 | SE | 예술 |

표 4는 이기종 학술정보 분류체계간 매핑한 실험의 최종결과 중 일부이다. 표 4에서 보는 바와 같이 중국분류코드는 1자리 숫자는 대분류, 2자리 숫자로 표현된 것은 중분류레벨이다. 그 이하로 세분류까지 표현되어 있다. 이에대해 확률적인 형태로 KISTI 표준분류체계(해외학술정보의 분류체계)를 표현하고 상호 일치여부를 확인하였다.

상위 20개의 분류에서 2개는 불일치 하였으며, 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. 결 론

본 연구에서는 서로 다른 언어권의 학술논문정보를 통합 구축하면서 발생한 테이블 간 이질적인 분류체계 문제를 해결하기 위해, 이질적인 분류코드 간의 의미관계를 확률적인

[표 4] 이기종 학술정보 분류체계간 매핑 테이블 결과(전체 매칭결과 중 발생빈도 상위 20위만 표시)

| 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 |

강도로 표현하는 자동기법을 제안하였다. 이기종 도메인간 의미관계를 파악하기 위해 자동기법으로 매칭테이블을 생산한 후 전문가의 검증과정을 거치는 프로세스를 적용한다면 매우 비용효과적인 상호매칭 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, 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.

또한 표 4에서 음영으로 표시한 7개 항목은 모두 35% 이상의 비교적 높은 확률수준으로 상호 매핑을 한 경우인데, 의학· 전산·수학의 3개 분야였다. 각 학문영역별로 매칭률을 비교하여 주제분야별 특성을 분석하는 연구 또한 의미가 있을 것으로 보인다.

## ▌참 고 문 헌▐

[1] 김판준, "기계학습을 통한 디스크립터 자동부여에 관한 연구", 정보관리학회지 제23권 제1호 pp.279-299, 2006

[2] 이재윤, "문서측 자질선정을 이용한 고속 문서분류기의 성능향상에 관한 연구", 정보관리연구, 제36권 제4호 pp51-69, 2005

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

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

[5] 정도헌, 김환민, 김혜선, 신기정, "과학기술 전문용어의 주제 분야별 전문성과 자동분류 성공률 간의 연관성 비교", 제14회 한국정보관리학회 학술대회 논문집, pp.31-36. 2007

[6] Deng, Zhi-Hong, Shi-Wei Tang, Dong-Qing Yang, Ming Zhang, Xiao-Bin Wu, and Meng Yang. "Two odds-radio-based text classification algorithms." Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops), 223-231.2002.

[7] Witten, Ian H., and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. (2nd ed.). San Francisco: Morgan Kaufmann. 2005.

[8] "Chinese Library Classification", Wikipedia, http://en.wikipedia.org/wiki/Chinese_Library_Classification (cited 2007.10.15)