커뮤니티 주제 어휘의 상호운용에 관한 연구*

Interoperability of Community-Oriented Subject Vocabulary

이 원 숙(Won-Sook Lee)**

〈 목 차 〉

초 록

본 연구에서는 최근 활발히 이용되고 있는 커뮤니티 주제 어휘의 특징을 선행연구인 ULIS-DL과 디지털 오카야마 대백과(デジタル岡山大百科)를 중심으로 알아보고, 이들 어휘들의 상호운용 모델을 일본의 동경도(東京都), 홋카이도(北海道), 한국의 충청남도의 각 도청 홈페이지의 디렉토리 용어들을 이용하여 제안하였다. NDLSH(National Diet Library Subject Heading)와 NDC(Nippon Decimal Classification)를 스위칭 언어를 사용하여 다섯개의 어휘 매핑 모델을 제안하였다. 마지막으로, 각각의 모델들에 대한 평가 및 한계점에 관하여 매핑의 정확율과 재현율을 이용하여 논하였다.

키워드: 커뮤니티 주제 어휘, 상호운용, 스위칭 언어

ABSTRACT

In this research, the first the characteristics of community-oriented vocabulary are investigated with preceding researches which are ULIS-DL and Digital Okayama Dai-Hyakka(DODH). The second this paper proposes a few mapping schemes to connect community directories and compares them by applying them to the resource directories of three local governments Tokyo and Hokkaido in Japan and Chungcheongnam-do in Korea. The mapping schemes use National Diet Library Subject Heading(NDLSH) and/or Nippon Decimal Classification(NDC) as a switching language. Evaluation of the proposed schemes shows their advantages and limitations.

Keywords: Community-Oriented Subject Vocabulary, Interoperability, Switching Language

* 이 논문은 연구자가 2008년 3월 일본 쯔꾸바대학(筑波大学)에 제출한 박사학위논문(コミュニティ指向のサブジェクトゲートウェイとその主題語彙に関する研究) 중 일부분을 수정하고 보완한 것임.

** 공주대학교 사범대학 문헌정보교육과 강사(wonsook.lee@gmail.com)

• 접수일: 2009년 2월 20일 • 최초심사일: 2009년 2월 25일 • 최종심사일: 2009년 3월 21일

Ⅰ. 서 론

1. 연구의 배경 및 필요성

인터넷의 폭넓은 보급은 웹 자원의 무제한적인 증가를 허용하였고, 이는 웹 전반의 품질을 떨어뜨리는 요인이 되었다. 이용자들은 많은 웹 자원들 속에서 자신이 원하는 양질의 정보자원을 신속하게 선별하기 어려워졌고, 이러한 경험의 반복이 결국 인터넷에 대한 신뢰도를 저하시키는 결과를 초래하였다.

이러한 문제들에 대한 대안의 하나로 인터넷 상에 많은 서브젝트 게이트웨이가 출현하게 되었는데, 이들은 도서관에서 오랜 기간 제공해오던 서비스, 즉 사람의 수작업을 통한 자료의 선정, 메타데이터의 부여 등, 일반 검색 엔진과 차별화된 서비스를 통하여 기존의 검색 엔진이 가지고 있던 많은 문제점들에 대한 해결책을 제시하고 있다.

일반적으로 서브젝트 게이트웨이에서는 디렉토리 인터페이스, 즉 분류체계를 이용하여 이용자들에게 정보를 제공한다. 온라인 환경에서 분류체계를 이용하는 것은 여러 가지 장점이 있는데, 그 중 Svenonius는 적합율과 재현율을 높이고, 이용자의 시간을 절약하며, 탐색어에 대한 문맥 및 브라우징 기능을 제공하고 상이한 언어간의 변환을 위한 메커니즘을 제공하는데 있어서 분류체계의 이용이 유용하다고 보았다. 그리고, 분류체계가 용어간의 상관관계를 계층적으로 보여주어서 의미론적 브라우징에 도움이 된다고 지적하고 있다.[1)]

BUBL[2)] 등의 규모가 크고 복합주제를 다루고 있는 서브젝트 게이트웨이에서는 주제 분야의 광범위성, 글로벌성(global use) 및 기계 가독형식의 제공[3)] 등의 이점 때문에 UDC(Universal Decimal Classification), DDC(Dewey Decimal Classification) 등의 도서관 분류 체계를 이용하고 있으나, 기존의 분류체계를 수정하거나 자신의 커뮤니티에 맞는 분류체계를 새롭게 개발하여 사용하는 커뮤니티 또한 증가하고 있다. 이는 도서관의 분류체계만으로는 각각의 커뮤니티가 가지고 있는 정보자원의 특징을 풍부하게 표현하기 어렵고, 그로 인하여 다양한 이용자의 요구에 적합한 맞춤 서비스가 어려워질 수 있기 때문이다.

---

1) E. Svenonius, "Use of Classification in Online Retrieval," *Library Resources & Technical Services*, Vol.27, No.1(Jan/Mar 1983), pp.76-80.

2) BUBL Home page, [cited 2009. 2. 19], 〈http://bubl.ac.uk/〉.

3) Michael Day, Cross-browsing subject gateways with Dewey Decimal Classification in the Renardus Service, 2004, 〈http://www.ukoln.ac.uk/metadata/presentations/jiscterm-2004/demo.html〉 [cited 2009. 2. 15].

〈그림 1〉 커뮤니티 주제 어휘의 예: 서울특별시청 홈페이지

대표적인 커뮤니티 주제 어휘로는 〈그림 1〉[4]과 같은 시·도 단위의 지방정부의 홈페이지에서 사용하고 있는 디렉토리 용어를 들 수 있는데, 이들 용어들은 각 지역의 생활방식 및 문화적 특징에 따라 다양하게 표현되고 있음을 알 수 있다.

이러한 커뮤니티 주제 어휘는 커뮤니티의 특성 및 이용자의 요구를 충분히 반영할 수 있기 때문에 앞에서 지적했던 전통적인 도서관 분류체계의 단점을 충분히 보완할 수 있다. 하지만, 그 반면 다양한 주제 어휘의 사용은 커뮤니티 간의 상호 정보 이용을 방해하는 큰 요인이 될 수 있으므로, 이들의 상호 이용 모델도 함께 연구되어야 한다.

이러한 배경으로 본 연구에서는 커뮤니티 주제 어휘와 관련된 선행연구를 간략하게 소개하고, 이어서 일본의 동경도(東京都), 홋카이도(北海道)와 한국의 충청남도의 홈페이지의 주제 어휘를 이용하여 커뮤니티 주제 어휘들 간의 상호 운용 모델을 제안하였다.

2. 관련 연구

스위칭 언어를 이용한 주제 어휘의 상호 이용에 관한 연구는 Renardus의 프로젝트가 가장 대표적이다. 이것은 유럽 5개국이 참여하는 프로젝트로 다양한 언어로 구성되어 있는 서브젝트 게이트웨이

4) 서울특별시 홈페이지, 〈http://www.seoul.go.kr/〉 [인용 2009. 2. 16].

를 통합·제공하는 서비스이며, 이때 사용된 스위칭 언어가 DDC(Dewey Decimal Classification)이다. Jens-Erik 역시 DDC를 스위칭 언어로 사용하여 복수의 주제 어휘를 연계할 것을 제안하였다.[5)]

독일의 ETH(Eidgenössische Technische Hochschule)의 ETHICS(ETH library Information Control System)에서는 영어, 프랑스어, 독일어의 시소러스 용어에 각각 UDC 번호를 부여하여 시소러스 간의 연계를 도모하였으며, 이창수는 UDC의 적용분야에 관한 연구에서 스위칭 언어로써의 UDC의 가능성에 대하여 언급하였다.[6)]

이 외에 서로 다른 색인언어를 사용하는 정보센터들 사이에서 정보교환을 위해 중재적인 목적으로 사용할 스위칭 언어를 개발하였는데, 이것이 BSO(Broad System of Ordering)이다.[7)]

Ⅱ. 커뮤니티 주제 어휘: 선행연구

커뮤니티 주제 어휘란 특정 커뮤니티의 목적과 의도에 따라 개발된 용어의 집합으로, 주로 커뮤니티의 정보자원을 분류하여 제공하는 디렉토리 서비스에 이용된다. 어휘(vocabulary)는 '어떤 일정한 범위 안에서 쓰이는 낱말의 수효, 또는 낱말 전체'[8)]를 의미하는데, 본 연구에서는 브라우징 서비스에 사용되는 '디렉토리 용어들의 집합' 및 'NDLSH와 NDC 용어의 집합'의 의미로 사용되어 개개의 디렉토리 용어를 뜻하는 주제어와는 차별화된 집합명사로써의 의미를 갖는다.

본 장에서는 커뮤니티 주제 어휘에 대한 이해를 돕기 위하여 선행연구로 진행한 두 연구를 소개한다. 이하의 두 개의 연구는 특정 커뮤니티에서 사용될 어휘는 모든 주제를 망라적으로 다루고 있는 도서관의 분류체계와는 차별화되어 개발될 필요가 있다는 사실을 명확히 보여주고 있다.

1. ULIS-DL

ULIS-DL(University of Library and Information Science - Digital Library)은 일본의 도서관 및 문헌정보학 교육기관에서 제작한 정보자원을 수집하여 제공하는 문헌정보학에 관한 서브젝트 게이트웨이로, 더블린코어의 15 요소 세트를 중심으로 개발된 메타데이터 스키마를 바탕으로 메타데이터를 작성·축적하고 있다.[9)] 메타데이터의 기술대상은 문헌정보학·정보미디어 연구에

5) Jens-Erik Mai, "The Future of General Classification." *Cataloging & Classification Quarterly*, Vol.37, No.1/2(2003), pp.3-12.
6) 이창수, "UDC의 적용분야에 관한 연구," 한국도서관·정보학회지, 제35권, 제4호(2004), pp.1-21.
7) 이경호, "스위칭언어로서의 BSO(Broad System of Ordering)," 한국도서관·정보학회지, 제14권(1987), pp.149-179.
8) Daum 국어사전, 〈http://krdic.daum.net/dickr/view_top.do〉 [인용 2009. 2. 7].
9) Lee, W. and Sugimoto, S., "Toward core subject vocabularies for community-oriented subject gateways," *Int. J. Metadata, Semantics and Ontologies*, Vol.1, No.3(2006), p.170.

관한 인터넷 정보자원 및 잡지기사 등으로, 약 4만 건의 메타데이터가 축적되어 있다.[10]

ULIS-DL은 1999년 시작된 이래로 처음 몇 년간은 디렉토리 서비스 없이 제공되었으나, 서브젝트 게이트웨이로서의 유용성을 보다 높이기 위하여 ULIS-DL에 축적되어 있는 메타데이터의 특징을 잘 반영한 주제 디렉토리 인터페이스를 구축하게 되었다. 이를 위하여 NDC의 이용을 고려하였지만, 다루고 있는 정보자원이 문헌정보학에 집중되어 있었기 때문에 NDC만으로는 전문적인 문헌정보학 디렉토리 구축이 어려웠다. 이러한 이유로 이미 작성된 메타데이터 레코더의 주제어들를 사용하여 디렉토리를 구축하였다. 대상이 된 메타데이터 레코더의 수는 26,358건이며, 이들 레코더에 부여되어있는 주제어는 모두 29,394개였다. 그러나 이 주제어를 조사해 본 결과, 도서관명, 인명 등의 고유명사와 메타데이터 작성자의 실수로 잘못 작성된 주제어 등의 출현빈도가 한번뿐인 주제어가 총 주제어의 4분의 3을 차지하고 있다는 사실이 밝혀졌으며, 이들을 그대로 주제 디렉토리로 사용하기에는 무리가 있었다. 이에, 메타데이터 레코더의 사용된 주제어들의 출현 빈도를 조사하여, 주제 디렉토리에 사용될 '핵심 어휘(core vocabulary)'를 개발하였다.

〈표 1〉 주제어의 출현빈도와 레코드 커버율

	주제어의 수	주제어를 포함하지 않는 레코드 수	비커버율(%)
출현빈도-2	3,979	1,519	6
출현빈도-3	2,045	2,083	8
출현빈도-4	1,366	2,590	10
출현빈도-5	*1,025*	*2,801*	*11*

위의 〈표 1〉[11]은 출현빈도에 따른 주제어 수와 그 주제어가 포함되지 않는 메타데이터의 레코드 수를 나타내고 있다.

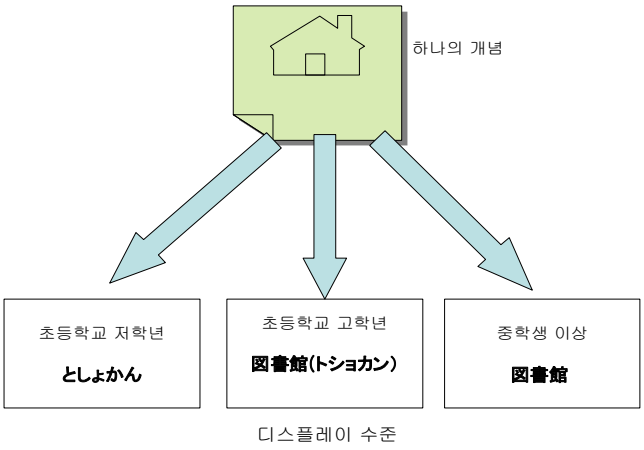
본 연구에서는 사람의 손으로 관리할 수 있는 합리적인 주제어 수를 천개 정도로 가정하였는데, 출현빈도가 5회 이상인 주제어가 이에 가장 근접한 1,025개 인 것으로 나타났다. 이 주제어들이 하나도 포함되어있지 않은 메타데이터의 레코드 수는 2,801개고, 대상 레코드의 10% 정도를 차지했다. 즉, 1,025개의 주제어로 90%정도의 레코드를 커버할 수 있다는 사실이 증명되었기 때문에, 이것을 '핵심 어휘'로 사용하여 주제 디렉토리를 개발하였다.[12]

10) ULIS-DL의 메타데이터를 대상으로 2007년도에 필자의 연구실에서 조사한 내용임.
11) Lee, W. and Sugimoto, S., 전게논문.
12) 상게논문.

2. 디지털 오카야마(岡山) 대백과

디지털 오카야마(岡山) 대백과[13]는 오카야마현립(岡山県立) 도서관이 현내의 공공기관과 협력하여 개발한 지역 포털 사이트로 오카야마의 모습을 백과사전과 같이 조사할 수 있도록 한 전자도서관 시스템이다. Z39.50을 기반으로 현내의 공공도서관을 통합적으로 검색할 수 있으며, 레퍼런스 데이터베이스 및 지역 정보자원을 다루고 있는 향토정보 네트워크를 제공하고 있다.[14]

향토정보 네트워크에서는 어린이들을 위한 주제 어휘(Kids Vocabulary: KV, 이하 KV)와 행정 자료의 제공을 위한 일반 주제 어휘(Prefecture Vocabulary: PV, 이하 PV)를 개발하였다. 또한 이상의 두 개의 주제 어휘와 함께 NDC를 함께 이용하여 이용자에게 디렉토리 서비스를 제공하고 있다.[15]



〈그림 2〉 이용자 수준에 맞춘 디스플레이

KV는 〈그림 2〉와 같이 하나의 개념에 대하여 이용자의 수준의 맞추어 다양한 디스플레이를 제공하고 있는데, 이것은 커뮤니티 주제 어휘의 전형적의 특징인 이용자 맞춤 서비스의 하나라고 할 수 있다.

PV는 현(県)의 주요 행정과 관련된 정보자원을 제공하기 위한 어휘로, 일반 이용자 및 어린이가 모두가 이용 대상자가 된다. 이 때문에 상기의 KV와 마찬가지로 기존 도서관에서 사용해오던 분류체계를 주제어 디렉토리에 사용하지 않고 단순하고, 친밀한 용어들을 사용하여 커뮤니티의 특

13) デジタル岡山大百科, 〈http://www.libnet.pref.okayama.jp/mmhp/index.html〉 [인용 2009. 2. 3].
14) 森山光良, 李沅淑, 杉本重雄, "メタデータのカテゴリ-索に向けたコミュニティ指向のポキャブラリ作成," ディジタル図書館, 제2권(2004), pp.123-134.
15) Lee, W. and Sugimoto, S, 전게논문, p.172.

징 및 요구에 맞추어 개발하였다.

개발한 KV와 PV 용어들을 〈표 2〉[16]에서 NDC 용어들과의 비교·분석하였다. KV와 PV는 각각 293, 287개의 용어로 구성되어 있으나, 하나의 카테고리 용어가 복수의 NDC 카테고리에 속할 수 있기 때문에 합계의 수가 443, 349로 되는 결과가 나왔다.

〈표 2〉 KV, PV 용어와 NDC 용어의 분포 비교

	NDC 카테고리	000	100	200	300	400	500	600	700	800	900	합계
KV 293	# KV 용어	17	8	8	196	58	54	28	62	6	6	443
	비율(%)	3.8	1.81	1.81	44.2	13	12	6.3	14	1.35	1.35	100
	#NDC 용어	7	7	3	44	27	27	20	26	4	4	169
PV 287	#PV 용어	15	2	12	171	30	56	44	17	1	1	349
	비율(%)	4.3	0.6	3.4	49	8.6	16	13	4.9	0.3	0.3	100
	#NDC 용어	4	2	5	34	11	18	25	15	1	1	116

내용을 분석해보면, KV:NDC와 PV:NDC의 용어에서 NDC 용어가 각각 169, 116개만이 사용되었고, 이들 중에서도 특히 사회과학(300) 분야가 양쪽 모두의 매핑에서 가장 높은 매핑율을 나타내고 있음을 알 수 있다.[17] 이 결과로 모든 학문 분야를 망라적으로 포함하고 있는 도서관의 분류체계가 특정한 커뮤니티의 주제어로 사용되기에는 부피가 너무 크며, 특성화 되어있지 않음을 알 수 있다. 즉, 도서관 분류체계를 그대로 서브젝트 게이트웨이의 주제어 디렉토리에 사용하는 것은 바람직하지 않으며, 커뮤니티에서 제공하는 정보자원과 이용자의 특성 및 요구사항을 잘 반영한 주제어의 개발이 요구된다는 사실을 보여주고 있다.

Ⅲ. 주제 어휘의 상호 운용 기본 모델

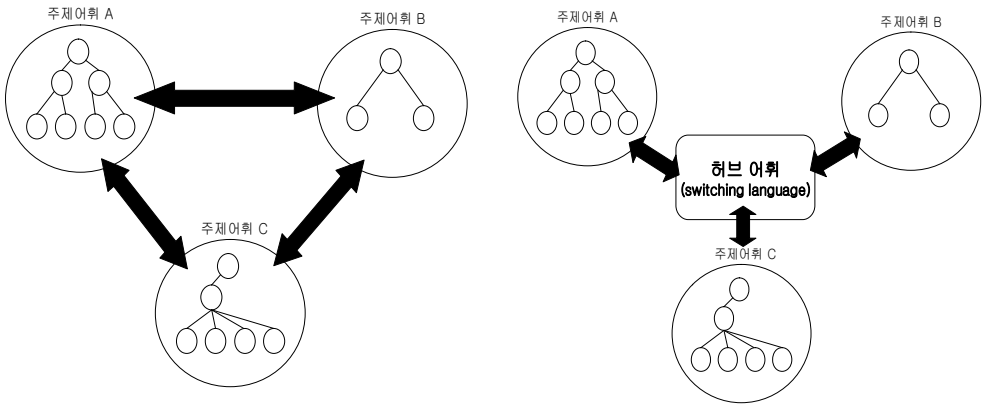
1. 매핑 모델

Doerr는 매핑을 대체적으로 동등한 용어, 개념 및 계층관계를 식별하는 과정이라고 정의하고 있다.[18]

주제 어휘의 상호이용에 관한 연구는 수년에 걸쳐 다각도에서 이루어져 왔다. 그 중 주제 어휘의 매핑에 관한 연구를 살펴보면 크게, (1) 복수의 어휘를 직접 매핑하는 방법과, (2) 스위칭 언어라

16) 상게논문.
17) 상게논문.
18) Martin Doerr, "Semantic problem of thesaurus mapping." *Journal of Digital Information*, Vol.1, No.8(2001), 〈http://jodi.tamu.edu/Articles/v01/i08/Doerr/〉.

불리는 매개체를 사용하여 매핑하는 방법으로 나눌 수 있다.

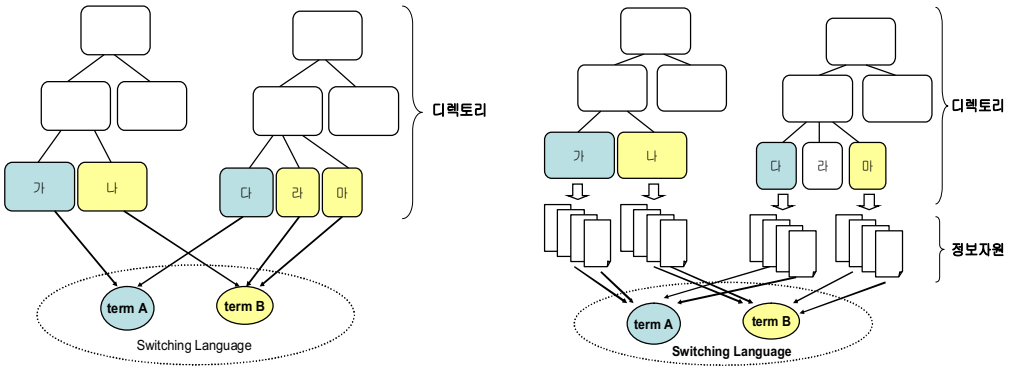


〈그림 3〉 매핑의 기본 모델

(1)은 〈그림 3〉의 좌측과 같이 각 주제 어휘들 사이를 중개물 없이 직접 연결하는 방법으로, 각 어휘 사이의 구조 및 어휘의 양의 차이에 따라 매핑이 복잡해질 우려가 있지만, 포함되어 있는 주제어의 수가 적고, 유동적이지 않을 경우에는 유용하다. 유럽의 공동 프로젝트인 MACS, OCLC에서 실시한 LSCH와 DDC의 매핑, Northwestern University의 LCSH와 MeSH와의 매핑 프로젝트[19] 등이 이 매핑을 사용한 대표적인 연구들이다.

(2)의 방법은 〈그림 3〉의 우측과 같이 각각의 주제 어휘를 스위칭 언어를 사용하여 연결하는 방법으로, 복수의 어휘를 일일이 매핑하는 수고와 (1)의 단점이었던 복잡함을 피할 수 있는 방법이다. 간접적 매핑으로 인하여 (1)에 비해 정확율이 떨어질 우려가 있지만, 그러한 단점에도 불구하고, (1)에 비해 비교적 많이 이용되고 있는 매핑 방법이며, 본 연구에서도 이와 같이 스위칭 언어를 이용한 매핑 방법을 이용하여 주제 어휘의 상호운용 모델을 제시하였다.

19) Marcia Lei Zeng, Lois Mai Chan, "Trends and Issues in Establishing Interoperability Among Knowl- edge Organization Systems." *Journal of the American Society for Information Science and Technology*, Vol.55, No.5(2004), pp.377-395.



〈그림 4〉 직·간접 매핑 비교

또한 주제 어휘의 매핑에 정보자원을 사용하였으며, 〈그림 4〉와 같이 정보자원의 사용 여하에 따라 직·간접 매핑으로 나누었다. 관리자의 관점에 따라 정보자원의 소속 카테고리가 결정되고, 이로 인해 관련이 높은 정보자원임에도 불구하고, 카테고리 용어의 차이에 의하여 매핑 되지 않는 경우가 종종 발생하였기 때문에 실제로 카테고리에 소속되어 있는 정보자원에 스위칭 언어를 부여하여 카테고리들을 매핑 하는 간접 방법도 사용하였다.

2. 매핑 대상

일본의 동경도(東京都),[20] 홋카이도(北海道)[21] 그리고 한국의 충청남도[22]의 각 도청 홈페이지의 디렉토리 용어들을 대상으로 매핑을 실시하였다.

한국과 일본은 전자정부의 구축에 대한 노력으로 모든 자치 단체가 독자의 홈페이지를 가지게 되었으며, 각각의 홈페이지를 통하여 활발하게 온라인 행정 서비스를 제공하고 있다. 이러한 홈페이지는 각 자치단체에서 제공하는 정보자원을 주제별로 분류하여, 최종 이용자에게 제공하고 있다. 분류된 주제항목은 '생활', '복지', '세금' 등 생활에 직접 관련이 있는 항목이 대부분이며, 모든 이용자층에게 알기 쉬운 용어들로 구성되어 있다. 기존의 도서관 분류 체계를 사용하지 않고, 각 커뮤니티의 특성과 의도에 맞추어 독자적으로 개발된 것으로, 본 연구에서 사용된 세 개의 주제 어휘 모두 계층구조를 가지고 있다. 이하의 〈표 3〉에 동경도의 주제 어휘의 일부를 소개한다.

20) 東京都 홈페이지, 〈http://www.metro.tokyo.jp/〉 [인용 2009. 2. 10].
21) 北海道 홈페이지, 〈http://www.pref.hokkaido.lg.jp/〉 [인용 2009. 2. 3].
22) 충청남도 홈페이지, 〈http://www.chungnam.net〉 [인용 2009. 2. 9].

〈표 3〉 동경도의 주제 어휘 일부

최상위 항목	하위의 항목	최하위의 항목
도민과 생활	교육·문화	교육
		평생교육
		도서관
		문화
		스포츠
		청소년·아동
	복지·인권	복지일반
		아동·가정
		고령자
		장애인
		인권

3. 스위칭 언어

스위칭 언어로는 일본 국립 국회도서관 주제표목표(이하 NDLSH)와 일본 십진분류법(이하 NDC)를 사용하였다.

NDLSH(National Diet Library Subject Heading)는 국립 국회도서관의 국내?외 서적의 목록에 사용되었던 주제표목을 수록해 놓은 주제표목표로, 2006년도 판은 1949년(昭和24年)부터 2007년(平成19年)까지의 주제어가 수록범위이며, 수록 주제어는 참조어를 포함하여 총 14,748건이다.[23]

NDC(Nippon Decimal Classification)은 모리키요시(もり・きよし)가 듀이의 십진분류법과 카터의 전개분류법에 준거하여 편찬하였으며, 1929년 초판이 발행되었다. 제5판까지는 모리키요시 개인이 편찬하였지만, 신정(新訂) 6판부터는 일본 도서관협회 분류위원회로 넘겨졌다. 1995년에 발행된 9판부터 2책으로 발간되었다. 본표 이외에의 보조표로는 형식구분, 지리구분, 해양구분, 언어구분, 언어공통구분, 문학형식구분, 문학공통구분, 상관색인이 있다.[24]

복수의 주제 어휘를 연계하기 위해 사용된 스위칭 언어는 앞장의 선행연구에서 언급한 것 같이 UDC와 같은 분류체계가 주류를 이루며, 주제표목표의 사용은 그 예를 찾아보기 힘들다.

그러나 다양한 커뮤니티 서브젝트 게이트웨이의 출현에 따른 주제 어휘의 다양화와 이용자의 높은 기대 수준은 더 많은 스위칭 언어 및 주제 어휘의 통합 방법에 관한 연구를 요하게 되었다.

이에 본 연구에서는 분류체계인 NDC와 함께 주제표목표인 NDLSH를 스위칭 언어로 제안함으로 다양한 스위칭 언어 및 주제 어휘의 매핑 모델(Mapping Scheme: 이하 MS)의 개발을 도모하였다.

---

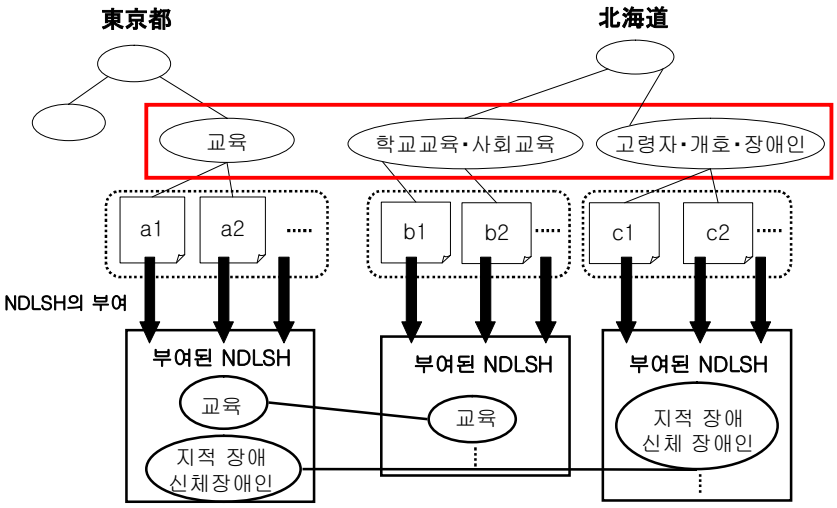
23) 國立國會図書館件名表目標2006年版序說, 〈http://www.ndl.go.jp/jp/library/data/josetsu/josetsu.pdf〉 [인용 2009. 2. 19].

24) 今まど子, 図書館學基礎資料-第五版(東京 : 樹村房, 2003), p.86.

Ⅳ. 매핑 모델의 제안 및 평가

1. 매핑 모델-1

NDLSH를 스위칭 언어로 사용한 간접 매핑으로, 개념도는 〈그림 5〉와 같다.



東京都의 「교육」카테고리는 NDLSH의 "교육"을 이용하여 **北海道**의「학교교육·사회교육」 카테고리와 관련을 갖으며, NDLSH의 "지적 장애"와" 신체장애인"을 이용하여 「고령자·개호(**介護**)·장애인」카테고리와 관련을 갖게 된다.

〈그림 5〉 MS1의 개념도

가. 매핑 순서

(1) 최하위의 카테고리에 속해 있는 정보자원을 무작위로 5개씩 선택한다. 단, 하나의 카테고리에 속해 있는 정보자원의 수가 5개 이하일 경우에는 모든 정보자원을 그 대상으로 한다.

(2) 선택된 개개의 정보자원에 대하여 NDLSH로부터 적합한 주제어를 최대 5개까지 부여하여, 〈그림 6〉과 같이 한 개의 카테고리별로 NDLSH 단어 집합을 만든다.

(3) 부여된 공통의 NDLSH 주제어를 이용하여 두개의 주제 어휘를 매핑한다.

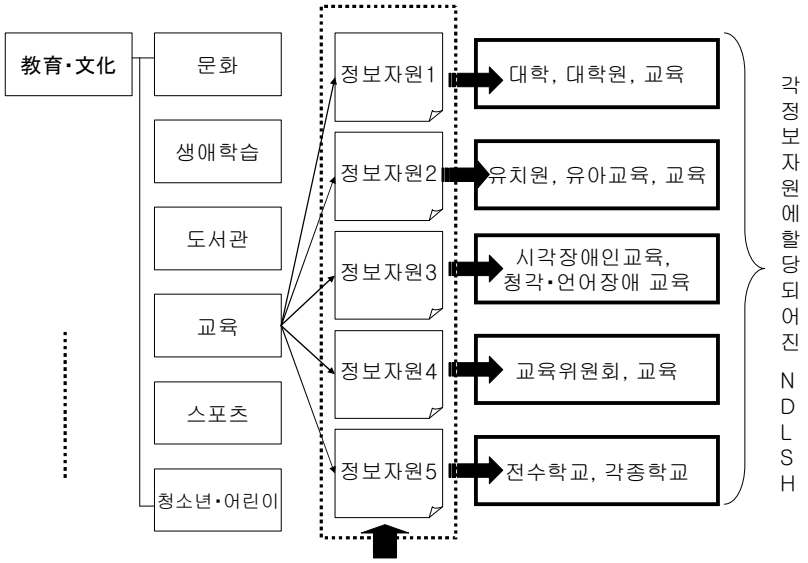
(4) 이 경우, 매핑의 결과가 상당히 광범위하게 나타나기 때문에, 아래의 (a),(b) 두개의 수식을 이용하여 결과의 폭을 줄인다.

$$CVcn = Rcn / Rn \qquad \text{(a)}$$

$$ICVth = \sum CVtn \times CVhn \qquad \text{(b)}$$

CVcn은 카테고리 용어 c와 NDLSH 용어 n과의 관련도를 나타내는 수식이며, ICVth는 NDLSH 용어 n을 경유한 카테고리 t와 h의 관련도를 나타내는 수식이다.

Rcn은 카테고리 c에 속해있으면서, NDLSH 용어 n이 부여되어 있는 정보자원의 수이며, Rn은 카테고리 c에 속해있는 모든 정보자원의 수를 나타낸다.



〈그림 6〉 MS1에서의 스위칭 언어 부여

나. 평가

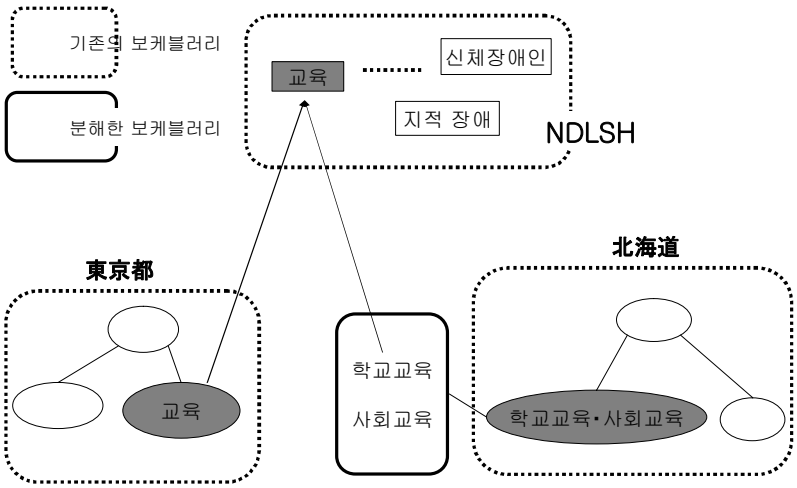
카테고리 용어에 의존하지 않고, 링크되어 있는 정보자원에 스위칭 언어를 부여한 후에 매핑 하는 방법으로, 특정 커뮤니티의 분류 관점에 의존하지 않고, 관련도가 높은 정보 자원이 모여있는 카테고리가 서로 매핑 될 수 있다. 따라서 새로운 분야의 발생 및 용어의 개념 변화 등에도 쉽게 적용할 수 있는 장점이 있다.

그러나, 정보자원 하나하나를 확인하며 스위칭 언어를 부여하여야 하기 때문에 시간과 노력이 많이 들며, 정보자원의 증가 및 변화에 따라 관련도가 변화할 수 있는 단점이 있다.

정보자원의 양이 많지 않고, 양적인 변화가 유동적이지 않는 커뮤니티 사이의 매핑, 예를 들어 조직체계의 변화 등에 의해 복수의 조직의 정보자원을 통합할 경우 유용할 것으로 판단된다.

2. 매핑 모델-2

NDLSH를 스위칭 언어로 사용한 직접 매핑으로, 개념도는 〈그림 7〉과 같다.



東京都의 「교육」카테고리는 NDLSH의 "교육"을 이용하여 **北海道**의「학교교육·사회교육」 카테고리와 관련을 갖게된다.

〈그림 7〉 MS2의 개념도

가. 매핑 순서

(1) 최하위의 카테고리 용어를 추출하여, 복수의 단어로 되어 있는 카테고리 용어를 하나의 단어로 각각 분리한다. 즉, [학교교육・사회교육]을 [학교교육], [사회교육]의 두 개의 카테고리 용어로 분리한다.

(2) 만들어진 각각의 카테고리 용어에 NDLSH로부터 적합한 주제어를 최대 5개씩 부여한다.

(3) 부여된 공통의 NDLSH 주제어를 이용하여 두 개의 주제 어휘를 매핑한다.

나. 평가

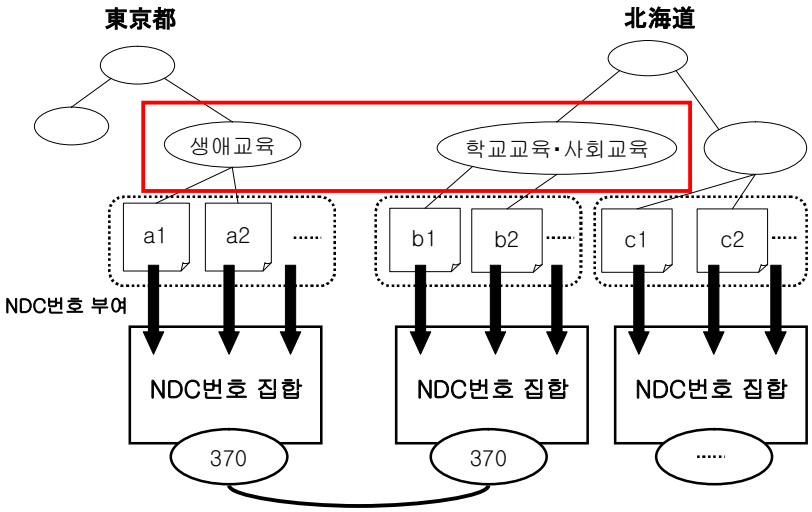
이미 작성되어있는 카테고리 용어에 NDLSH로부터 스위칭 언어를 부여하는 직접 매핑 방법으로, 매핑에 정보자원을 이용하는 간접 매핑 방법에 비해 시간과 노력이 적게 든다는 장점이 있다. 하지만, 앞에서도 지적했듯이 '행정' 등과 같이, 카테고리어의 의미 범위가 넓을 경우에는 NDLSH의 주제어 부여에 상당한 어려움이 따를 수 있다. 그리고, MS1의 장점이 이곳에서는 단점이 될 수 있는데, 즉 특정 커뮤니케이션의 분류 관점에 매핑이 의존되어, 관련이 높은 정보자원들이 서로

매핑 되지 않을 경우가 발생할 수 있다는 것이다.

그러므로, 이 방법은 카테고리의 수가 적고, 카테고리어의 의미가 협의하고 명확한 경우에 활용하는 것이 바람직하다고 판단된다.

3. 매핑 모델-3

NDC를 스위칭 언어로 사용한 간접 매핑으로, 개념도는 〈그림 8〉과 같다.



東京都의 「생애교육」 카테고리는 NDC"370(교육)"을 이용하여**北海道**의「학교교육·사회교육」 카테고리와 관련을 갖게 된다.

〈그림 8〉 MS3의 개념도

가. 매핑 순서

(1) 각 카테고리에 속해있는 정보자원을 무작위로 5개씩 선택한다. 단, 하나의 카테고리에 속해 있는 정보자원의 수가 5개 이하일 경우에는 모든 정보자원을 그 대상으로 한다.(MS1의 (1)과 동일)

(2) 선택된 개개의 정보자원에 대하여 적합한 주제어 및 번호를 [교육(370)]과 같이 최대 5개까지 부여하여, 〈그림 8〉과 같이 한 개의 카테고리별로 NDC 번호 집합을 만든다.

(3) (2)의 집합에서 출현빈도가 가장 높은 NDC번호를 한 개 이상 각 카테고리별로 추출한다.

(4) (3)에서 얻어진 공통의 NDC 번호를 이용하여 두 개의 주제 어휘를 매핑한다.

나. 평가

MS1과 마찬가지로, 간접 매핑으로서의 유사한 장단점을 갖는다.

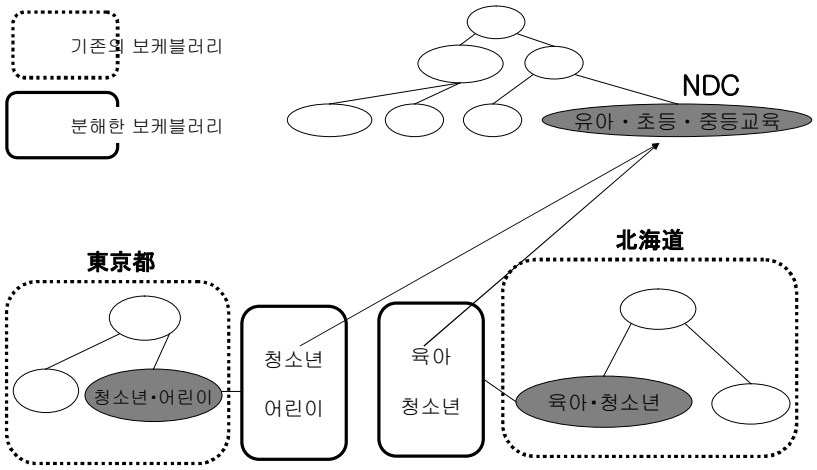
NDC는 도서관에서 사용될 학문의 분류체계로 개발되었기 때문에, 실용적인 이용을 위하여 개발된 카테고리 용어 및 정보자원의 주제어로는 적절하지 못하다는 것이 실제의 경험으로 밝혀졌다. 예를 들어 '치안'이라는 카테고리 용어는 적절히 부여할 NDC 용어가 발견되지 않았기 때문에 '형법', '형사법' 등의 용어가 부여되었다. 이것은 정보자원에 있어서도 동일하게 나타났으며, 이로 인해 주제어 부여에 많은 어려움을 겪었다.

반면, NDC 번호를 언어중립적인 스위칭 언어로써 사용할 수 있으며, NDLSH에 비해 용어의 양이 상대적으로 적기 때문에 주제어 부여에 있어서 스위칭 언어의 검색이 용이하다는 장점도 있다.

언어가 다른 주제 어휘의 국제적 상호 운용에 언어 중립적인 NDC 번호의 사용이 유용할 것으로 추측하지만, 이번 연구에서는 증명할 수 없었다.

4. 매핑 모델-4

NDC를 스위칭 언어로 사용한 직접 매핑으로, 개념도는 〈그림 9〉와 같다.



東京都의 「청소년・어린이」 카테고리는NDC"376(유아・초등・중등교육)"을 이용하여
北海道의 「육아・청소년」 카테고리와 관련을 갖게 된다.

〈그림 9〉 MS4의 개념도

가. 매핑 순서

(1) 최하위의 카테고리 용어를 추출하여, 복수의 단어로 되어 있는 카테고리 용어를 하나의 단

어씩으로 분해한다.(MS2의 (1)과 동일)

(2) 만들어진 각각의 카테고리 용어에 NDC로부터 적합한 주제어 및 번호를 최대 5개씩 할당하여, 카테고리별로 NDC 번호 집합을 만든다.

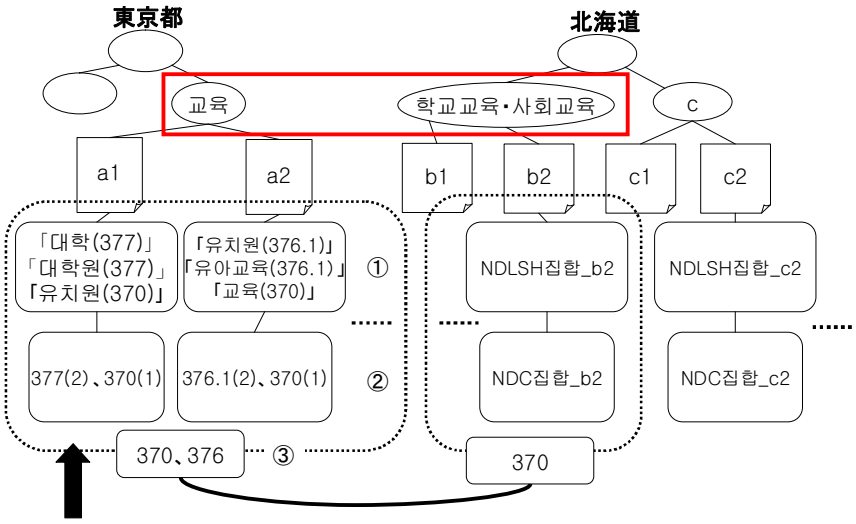
(3) 부여된 공통의 NDC 번호를 이용하여 두 개의 주제 어휘를 매핑한다.

나. 평가

MS2와 마찬가지로, 이미 작성되어 있는 카테고리 용어에 주제어 및 NDC 번호를 부여하는 직접 매핑 방법이기 때문에, 간접 매핑 방법에 비해 코스트가 적게 든다는 장점이 있다. 그 외의 장단점에 관하여는 MS3에서 지적한 내용과 흡사하다.

5. 매핑 모델-5

NDLSH의 각 용어에는 한 개 이상의 NDC 번호가 주어져 있는데, 이것은 해당 NDLSH 용어가 부여되어 있는 정보자원이 NDC의 어떤 카테고리에 분류되어 있는가를 나타내는 번호이다. 예를 들어 '여성복지'라는 NDLSH의 주제어가 주어져 있는 정보자원이 NDC의 369.25에 분류되어 있다면 NDLSH에서 '여성복지-369.25'라는 데이터를 제공하고 있다는 것이다. 본 모델은 이와 같은 NDC 번호를 스위칭 언어로 사용한 일종의 실험적 시도이며, 개념도는 〈그림 10〉과 같다.



「교육」에 대한 ①NDLSH집합, ②NDC번호집합, ③최고의 출현 빈도의 NDC번호
:NDC번호 370을 이용하여 **東京都**의 「교육」은 **北海道**의 「학교교육·사회교육」과 관련을 갖게 된다.

〈그림 10〉 MS5의 개념도

가. 매핑 순서

(1) 각 카테고리에 속해있는 정보자원을 무작위로 5개씩 선택한다. 단, 하나의 카테고리에 속해 있는 정보자원의 수가 5개 이하일 경우에는 모든 정보자원을 그 대상으로 한다.(MS1의 (1)과 동일)

(2) 선택된 정보자원에 NDLSH로부터 적절한 주제어를 최대 5개씩 부여하여, 카테고리별로 NDLSH의 집합을 만든다.

(3) (2)에서 얻어진 NDLSH 집합으로부터 NDC 번호를 유출한 후, 출현 빈도가 가장 높은 NDC 번호를 각각의 카테고리별로 추출한다.

(4) (3)에서 얻어진 공통의 NDC 번호를 이용하여 두개의 주제 어휘를 매핑한다.

나. 평가

NDC 번호는 언어에 의존하지 않는 스위칭 언어로 사용될 수 있다는 강점이 있지만, 앞에서도 지적한 바와 같이 용어상의 특징으로 주제어 부여에 큰 어려움이 있을 수 있다. 본 매핑 방법에서는 NDC 용어를 사용하지 않고, NDLSH의 주제어를 통하여 NDC 번호를 취득하는 방법으로 이러한 문제점을 극복하였다.

그러나 NDLSH에 연결되어있는 NDC 번호가 경우에 따라 상당히 분산되어 있어 예상치 못한 결과로 매핑 되는 경우가 발생하여 순조로운 매핑이 이루어지지 못했다.

Ⅴ. 결과 및 분석

앞장에서는 NDLSH와 NDC를 스위칭 언어로 사용한 주제 어휘의 매핑 모델을 소개하고, 각각의 모델의 장·단점을 매핑의 경험을 토대로 기술하였다.

본 장에서는 재현율과 정확율을 사용한 종합적인 평가를 실시하였다. 단, MS5는 시험적인 시도로 진행된 방법이며, 대부분의 매핑에서 최하위의 정확율을 기록하고 있기 때문에 종합적인 분석에서 제외하였다. 이에 대한 분석은 앞장의 MS5에 대한 평가로 대신한다.

1. 매핑 결과: 재현율과 정확율

재현율은 어떤 검색식에 의하여 데이터베이스 중의 적합 레코드가 어느 정도 검색되었나를 나타내는 비율로, 누락의 적음을 나타낸다. 본 연구에서는 두개의 주제 어휘 사이에서 매핑이 어느 정도 이루어졌는가를 의미하다. 그리고 정확률은 적합률이라고도 하며, 어떤 검색식에 의해 검색된 레코

드 중에 적합 레코드가 어느 정도 포함되어 있는가를 나타내는 비율로 노이즈의 적음을 의미한다. 본 연구에서는 발견된 매핑 중에 어느 정도가 정답집합과 일치하였느냐를 의미하며, 이때 사용된 정답 집합은 사람의 손으로 작성된 매핑 결과로, 두 개의 주제 어휘 사이의 가장 이상적인 매핑을 사람 눈으로 판단하여, 작성하여 놓은 집합이다. 재현율과 정확율의 결과는 아래의 〈표 4〉와 같다.

〈표 4〉 각 매핑 모델의 재현율과 정확율

	MS 1	MS 2	MS 3	MS 4	MS 5
(1) T to H	P=0.603(38/63) R=0.493(38/77)	P=0.734(47/64) R=0.61(47/77)	P=0.475(38/80) R=0.493(38/77)	P=0.402(54/134) R=0.701(54/77)	P=0.333(29/87) R=0.377(29/77)
정확율 순위	2	1	3	4	5
(2) H to T	P=0.4(20/50) R=0.338(20/59)	P=0.708(46/65) R=0.779(46/59)	P=0.345(28/81) R=0.474(28/59)	P=0.285(40/140) R=0.677(40/59)	P=0.275(22/80) R=0.372(22/59)
정확율 순위	2	1	3	4	5
(3) C to T	P=0.573(59/103) R=0.567(59/104)	P=0.570(61/107) R=0.586(61/104)	P=0.295(50/169) R=0.480(50/104)	P=0.239(59/246) R=0.567(59/104)	P=0.177(35/198) R=0.336(35/104)
정확율 순위	1	2	3	4	5
(4) C to H	P=0.284(63/222) R=0.567(63/111)	P=0.503(81/161) R=0.729(81/111)	P=0.161(34/210) R=0.306(34/111)	P=0.221(74/334) R=0.666(74/111)	P=0.234(50/213) R=0.450(50/111)
정확율 순위	2	1	5	4	3

정도(precision) =(검색결과집합중의 정답집합) / (검색결과집합)
재현율(recall) =(검색결과집합중의 정답집합) / (정답집합)
T: 東京都, H: 北海道, C: 충청남도

2. 결과 분석

가. 스위칭 언어의 종류의 차이

NDLSH(National Diet Library of Subject Heading)은 주제표목을 수록한 일본의 주제표목표로, 모든 주제를 커버할 수 있는 포괄적인 주제어휘를 포함하고 있다. 한편, NDC(Nippon Decimal Classification)는 정보자원을 분류하기 위한 계층구조를 갖는 분류체계로 일본의 공공도서관을 비롯한 대부분의 도서관에서 사용되고 있다. 이렇듯 개발 목적 다른 두 개의 주제 어휘를 각각 스위칭 언어로 사용함으로, 어떠한 결과의 차이가 나타났는지를 분석하였다.

MS1과 MS2가 NDLSH를, MS3와 MS4가 NDC를 각각 스위칭 언어로 사용하고 있는데, 정확율의 결과로 분석을 하면,(1)-(4)의 모든 매핑에 있어서 NDLSH를 스위칭 언어로 사용한 쪽이 높은 측정치를 보이고 있는 것을 〈표 4〉에서 확인할 수 있다. 이 결과는 NDLSH의 스위칭 언어로써의 가능성을 보여주고 있다. 그러나 이번 연구에서는 NDLSH와 NDC의 구조적 특징, 즉 NDLSH의 상위어, 하위어, 관련어 등과 NDC의 주류, 강목, 요목의 3단계 계층 구조를 주제 어휘의 매핑에 반영하지 못하였다. 사용 목적과 의도에 따라 생겨난 이러한 구조적 특징들을 주제 어휘

의 매핑에 사용하면 더 다양한 매핑 모델의 개발이 가능할 것이다.

나. 직·간접의 차이

주제 어휘의 매핑에 정보자원을 이용하였는지 여부에 따라 매핑의 종류를 직접과 간접으로 나누고, 각각의 매핑에 대한 정확율을 비교하였다. 매핑에 정보자원을 이용한 이유는, 정보자원의 특성을 충분히 반영하지 못하는 카테고리 용어 및 관리자의 분류 관점 차이 등으로 인해 관련이 깊은 정보자원들이 포함되어 있는 카테고리 사이가 매핑되지 않을 우려가 있기 때문이었는데, 이러한 시도가 매핑 결과에 어떻게 반영되었는지 살펴보았다.

NDC를 스위칭 언어로 사용한 매핑 모델의 경우 간접 매핑의 정확율이 더 높게 산출되었는데, 그 차이가 미미하고, 정확율 자체에 대한 수치가 매우 낮게 나왔다. NDLSH를 사용한 두 매핑 모델의 경우는 〈표 4〉의 (3)충청남도에서 동경도로의 매핑(C to T)을 제외한 세 곳에서 직접 매핑의 정확율이 높게 나타났다.

이러한 결과의 원인으로는 첫째, 직접 매핑의 경우 간접 매핑에 비해 부여된 스위칭 언어가 적고, 이로 인해 사용된 스위칭 언어의 폭이 적기 때문에 매핑이 좀 더 정확히 이루어진 것을 들 수 있다.

두번째 원인으로는 재현율 및 정확율을 계산할 때 사용된 정답 집합과 관련이 있다. 정답 집합은 두 개의 주제 어휘 사이의 가장 이상적인 매핑을 사람의 눈으로 확인하여 작성한 것으로, 정확한 매핑의 판단 기준이 디렉토리 용어들 사이의 의미 일치도였다. 이는 앞에서 설명한 간접 매핑의 의도를 충분히 반영하지 못한 것을 의미하며, 결과적으로 직접 매핑의 정확율이 높게 나오는 원인이 되었다.

다. 매핑 대상의 차이

1960년대에 개발된 BSO(Broad System of Ordering)는 다양한 언어로 쓰여진 각국의 과학기술 공유의 필요성에서 출발하였으며, 현재 언어를 초월한 정보자원 공유 및 이용에 대한 요구는 인터넷을 기반으로 한 정보화 사회에서 더욱더 가속화되고 있다.

필자는 다언어 메타데이터[25]를 통하여 언어장벽을 뛰어넘는 정보자원의 상호운용 모델을 제시한 바 있으며, 본 연구에서는 그 관련 연구의 하나로 스위칭 언어를 이용한 한국과 일본의 주제 어휘에 관한 상호 운용 가능성에 대하여 타진해 보았다.

결과는 〈표 4〉에서 알 수 있듯이 스위칭 언어 및 직·간접의 여부와 상관없이, 동일 언어로 쓰여진 일본의 동경도과 홋카이도 사이의 매핑의 결과에 비교하여 대체적으로 낮은 정확율을 나타내고 있다.

25) Wonsook Lee, Shigeo Sugimoto, Mitsuharu Nagamori, Tetsuo Sakaguchi, Koichi Tabata. *A Subject gateway in Multiple Languages : a Prototype Development and Lessons Learned*. (Proceedings of International DCMI Metadata Conference and Workshop, 2003), pp.59-66.

양국의 주제 어휘는 지방 정부의 홈페이지에서 사용하고 있는 디렉토리 용어들로, 이용자들에게 보다 친숙하고 편리하게 행정 자료를 제공하려는 공통의 목적 하에서 개발되었다. 하지만, 다루고 있는 정보자원 및 문화의 차이로 발생된 주제 어휘의 특성이 매핑에 많은 영향을 미쳤다. 예를 들어, 각각이 제공하는 정보자원을 살펴보면, 일본의 경우 각 도시에서 발행한 안내문 등이 많았으며, 한국의 경우에는 전자 민원에 관련된 자료가 많이 포함되어있었다.

정보자원의 상호이용의 목적은 이용자들에게 보다 좋은 정보자원을 다양하게 제공하기 위해서인데, 이 경우 이용자의 정보 사용 목적과 의도를 파악하여 상호 이용될 가치가 있는 정보자원이 무엇인지 파악하여 그 정보자원들만을 이용하여 상호운용 모델을 개발하는 것이 중요하다. 즉, 본 연구의 예와 같이 행정자료에 관련된 주제 어휘의 상호운용을 고려할 경우, 외국에서 수요가 적을 것으로 예상되는 전자민원에 관련된 디렉토리는 과감하게 제외시킴으로 매핑의 낭비를 없앨 수 있을 것이다. 이러한 각국의 차이와 이용자의 수준을 고려한 매핑 모델을 개발하면 더욱 정확하고 유용한 상호 운용이 가능하게 될 것으로 기대된다.

Ⅵ. 결 론

양질의 정보자원을 원하는 이용자들은 최소의 노력으로 최대의 효과를 기대한다. 즉, 한번의 검색으로 자신의 요구에 맞는 최적의 정보자원을 이용하기 원한다. 이러한 이용자들에게 다양한 커뮤니티를 연계하여 그들의 정보자원을 통합적으로 제공하는 것은 매우 의미 있는 일이다. 하지만, 표준화되지 않은 커뮤니티만의 여러 기준들이 정보자원의 상호운용을 어렵게 하고 있다. 그 대표적인 예가 커뮤니티의 특성에 맞추어 개발된 주제 어휘라고 할 수 있는데, 이러한 외적인 틀의 상호운용에 대한 연구 없이는 실질적인 내용물인 정보자원의 상호활용을 논할 수 없다.

이러한 배경으로 본 연구에서는 NDLSH와 NDC를 스위칭 언어로 사용한 커뮤니티 주제 어휘의 상호 운용 모델을 제안하였으며, NDLSH를 스위칭 언어로 사용한 직접 매핑의 조합이 매핑 모델로써 가장 적당하다는 결론이 나왔다.

정답 집합의 작성 및 매핑 과정에 있어서 구체적인 규칙이 정해져 있지 않아 매핑 및 평가의 객관성이 결여된다는 문제점이 여전히 남아있지만, 커뮤니티 주제 어휘의 상호운용에 관한 매핑 모델을 소개하고, 주제 어휘의 상호운용에 응용할 수 있는 스위칭 언어의 폭을 넓혔다는데 본 연구의 의미가 있다.

〈참고문헌은 각주로 대신함〉