최대 개념강도 인지기법을 이용한 데이터베이스 자동선택 방법에 관한 연구*

A Study on Automatic Database Selection Technique Using the Maximal Concept Strength Recognition Method

정도헌(Do-Heon Jeong)**

초 록

본 연구에서 제안하는 기법은 최대 개념강도 인지기법(Maximal Concept-Strength Recognition Method: MCR)이다. 신규 데이터베이스가 입수되어 자동분류가 필요한 경우에, 기 구축된 여러 데이터베이스 중에서 최적의 데이터베이스가 어떤 것인지 알 수 없는 상태에서 MCR 기법은 가장 유사한 데이터베이스를 선택할 수 있는 방법을 제공한다. 실험을 위해 서로 다른 4개의 학술 데이터베이스 환경을 구성하고 MCR 기법을 이용하여 최고의 성능값을 측정하였다. 실험 결과, MCR을 이용하여 최적의 데이터베이스를 정확히 선택할 수 있었으며 MCR을 이용한 자동분류 정확률도 최고치에 근접하는 결과를 보여주었다.

ABSTRACT

The proposed method in this study is the Maximal Concept-Strength Recognition Method(MCR). In case that we don't know which database is the most suitable for automatic-classification when new database is imported, MCR method can support to select the most similar database among many databases in the legacy system. For experiments, we constructed four heterogeneous scholarly databases and measured the best performance with MCR method. In result, we retrieved the exact database expected and the precision value of MCR based automatic-classification was close to the best performance.

키워드: 자동분류, 자동범주화, 최대 개념강도 인지기법, 자동 데이터베이스 선택, 텍스트마이닝
automatic classification, automatic categorization, maximal concept-strength recognition, automatic database selection, text mining

---

* 관련특허: 최대 개념강도 인지기법을 이용한 최적의 데이터베이스 선택장치 및 그 방법(10-2009-0118944).
** 한국과학기술정보연구원 정보유통본부 정보기술연구실 선임연구원(heon@kisti.re.kr)
■ 논문접수일자 : 2010년 8월 15일 ■ 최초심사일자 : 2010년 8월 21일 ■ 게재확정일자 : 2010년 8월 27일
■ 정보관리학회지, 27(3): 265-281, 2010. 〔DOI:10.3743/KOSIM.2010.27.3.265〕

1. 서 론

다양한 학술정보 데이터베이스를 구축하고 서비스하는 경우, 보다 일관성 있는 서비스를 위해 개별문서를 분류하고 이를 기반으로 통합 정보검색 환경을 구축하고자 하는 경우가 많다. 그러나, 개별문서의 자동분류 문제는 대용량의 데이터베이스 환경 하에서 학습모델을 대용량 기반으로 처리해야 하는 어려움 때문에, 기존의 여러 연구에서 제안된 최적의 알고리즘을 실제 서비스에 적용하기 어려운 점이 있다. 통합 정보서비스의 경우에는 통합 서비스의 표준 주제분류 체계에 맞도록 여러 자원을 일관성 있게 자동분류해야 하지만, 이기종 데이터베이스는 상호간의 자동분류 성능이 현저히 떨어지기 때문에 현실적인 시스템에서의 실효성이 매우 떨어진다. 논문과 특허, 또는 논문과 연구보고서의 경우와 같이 서로 상이한 분류와 용어속성을 갖는 연구정보 간의 문제 뿐 아니라, 비교적 유사한 영역이라 생각되는 학술논문의 범위 내에서도 사용되는 용어개념이 매우 상이하여 분류학습 모델을 일반화하는 것이 매우 어렵다.

본 연구에서는 다양한 분야와 속성을 가진 정보자원이 구축된 데이터베이스 환경에 새로운 자원이 입수되어 자동으로 범주를 할당해야 하는 환경에서, 신규로 입수되어 자동으로 분류코드를 할당받아야 하는 자원이 기존의 여러 데이터베이스 중 가장 유사한 데이터베이스 영역으로 자동 할당되는 모델을 구현하고자 한다. 기존에 구축된 데이터베이스의 환경을 변화시키지 않은 상태에서 신규 데이터베이스가 입수되었을 때 가장 유사한 데이터베이스를 선택하여 최적의 자동분류 결과를 도출할 수 있는가를 측정하고자 한다.

이를 위해, 특정 개념이 가장 잘 표현되는 학습모델은 신규 문헌의 요청에 대해 가장 강한 출력 값을 나타낼 것이며 이를 측정하여 유사 데이터베이스를 찾아낼 수 있다는 가정 하에 새로운 기법인 최대 개념강도 인지기법(Maximal Concept-Strength Recognition Method: MCR)을 제안하고 다양한 데이터베이스 간 교차실험을 하여 이를 증명하고자 한다.

2. 관련연구 및 새로운 기법의 제안

2.1 자질벡터의 투표를 이용한 FV 분류기

자질값 투표형 분류기(Feature Value Voting Classifier: FVC)는 문서를 구성하고 있는 자질값에 대한 주제범주, 가중치값의 벡터정보를 이용하여 투표방식을 통해 분류를 수행한다. 이러한 방식은 자질값의 주제분야 가중치 값을 다수결 투표 방식으로 출력하는 분류방법으로서, Ko과 Seo(2004)는 자질값 투영기법(feature projection technique)을 이용한 투표형 분류기인 TCFP(Text Categorization using Feature Projections)를 제안하고 k-NN, Rocchio, Naive Bayes 분류기 등과 성능을 비교하여 속도와 성능면에서 좋은 결과를 얻어냈다. 이재윤(2005)은 자질선정 방식과 자질값의 가중치 부여방식을 변경하며 성능변화를 측정하여 로그승산비를 이용한 투표기반 분류기를 SVM(Support Vector Machine) 분류기와 비교하였다. 실험

결과, 적절한 자질값 선정방식을 통해 성능이 좋은 분류기로 알려져 있는 SVM와 비교하여 우위의 성능을 보임을 확인하였다.

FV 분류기는 성능이 우수하면서도 계산상의 복잡성이 상대적으로 낮고 처리속도가 빠르다는 장점을 가지고 있어 향후 대용량 학습모델 기반의 분류기 개발을 위한 좋은 방안이 될 수 있을 것으로 본다. 또한 이 방식은 색인어를 중심으로 한 주제가중치 벡터를 생성하기 때문에 이를 지식베이스로 구축할 경우, 언어자원의 연구 재활용성이 매우 높은 구조적인 장점도 가지게 된다.

$$S(f_i, c_j) = \frac{TP}{\sqrt{(TP+TN)(TP+FP)}} \quad (1)$$

자질 가중치값 S는 자질과 범주의 연관도를 의미하며 저자키워드(자질)의 주제분야(범주)간 유사도를 측정하기 위하여, 고빈도어 선호경향을 갖는 연관성 척도인 오치아이 유사계수를 사용하였다. TP는 자질 f가 범주 c에 출현한 빈도, TN는 자질 f가 범주 c에 출현하지 않은 빈도, FP는 자질 f가 나타나지 않은 범주 c의 빈도를 각각 의미한다(공식(1)).

$$\vec{d} = \{vs(f_1, c_j), vs(f_2, c_j), \ldots, vs(f_n, c_j)\} \quad (2)$$

데이터베이스는 문헌 d로 구성되는데, n개의 용어를 갖는 문헌 d는 계산된 자질별 가중치 값을 이용해 아래 공식(2)와 같은 자질값 벡터로 표현할 수 있다.

$$vs(f, c_j) = (1+\log tf) \times \log(N/df) \times S(f, c_j) \quad (3)$$

이때, 문헌벡터 $\vec{d}$를 구성하는 $vs(f, c_j)$는 문헌 d안의 자질값 f의 가중치 값을 의미하며, TF*IDF을 이용해 위의 공식과 같이 계산할 수 있다(Salton & Buckley 1998). IDF를 의미하는 logN/df은 전체문헌의 수 N을 특정용어가 발생한 문헌의 수 df(문헌빈도)로 나눈 값이다. 용어의 가중치 값을 구하기 위해 유사도 $S(f, c_j)$를 측정하여 자질별 가중치 값인 $vs(f, c_j)$를 측정할 수 있다. 문서를 구성하는 모든 자질 f에 대해(주제분류, 가중치값)의 쌍으로 구성되는 용어벡터를 구성하고 이를 FV 분류기에 적용한다.

$$decision[c_j] = \arg max_{c_j \in C} \sum_i vs(f_i, c_j) \quad (4)$$

공식(4)는 최종적으로 분류결정을 위한 공식이며, 실험 문서인 $d = \{f_1, f_2, \ldots, f_n\}$, 주제범주 $C = \{c_1, c_2, \ldots, c_m\}$ 라고 할 때, 자질 fi가 범주 cj에 대해서 가지는 자질값을 $vs(f_i, c_j)$라고 하면 자질값 투표 분류기는 공식(4)을 만족하는 범주 cj를 문서에 할당한다.

2.2 최대 개념강도 인지 기법(Maximal Concept-Strength Recognition Method)

본 연구에서 새롭게 제안하는 최적의 데이터베이스 선택기법은 "최대 개념강도 인지 기법(Maximal Concept-Strength Recognition: MCR)"이다. 특정 문헌이 입력되었을 때 기 구축되어 학습된 분류기 중에서 신규 문헌을 가장 잘 분류할 수 있는 분류기를 선택하는 것이 이 기법의 목적이다. 이것은 새로운 주제범주

를 가장 잘 할당할 수 있는 학습모델은 문헌을 구성하는 개별단위의 의미(자질값)에 대해 반응하는 개념의 출력 강도가 가장 클 것이라는 가정에 근거한다.

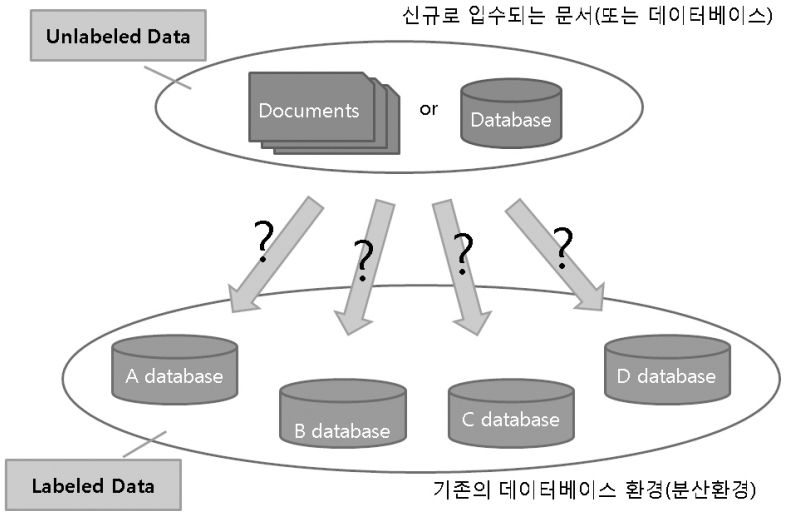
이것은 정보검색 이론에서 언급되는 데이터 결합, 그 중에서도 컬렉션 결합과 개념적으로 유사한 면이 있다. 분산 검색환경에서 다양한 정보원으로부터 검색결과를 결합하는 컬렉션 결합(collection fusion)은 여러 개의 데이터베이스를 검색한 결과가 이 데이터베이스들을 통합한 데이터베이스를 검색한 결과와 유사하도록 만드는 것에 중점을 둔다(Voorhees, Gupta, & Johnson-Laird 1995). 그러나, 컬렉션 결합은 어떻게 각 데이터베이스로부터 적절한 개수의 문헌을 검색하고 순위화하여 검색결과 리스트를 재구성하는 가에 목적이 있는 반면(Nuray & Can 2005), MCR 기법은 데이터베이스 단위로 신규정보와 기존정보의 유사성을 측정하는 데 목적이 있다.

〈그림 1〉과 같이 범주가 부여되지 않은 문헌 집단이 들어왔을 때 범주가 부여되어 있는 여러 데이터베이스 중 의미적으로 가장 유사한 (범주를 정확히 부여할 수 있는) 데이터베이스를 선택할 수 있으면 그 데이터베이스로부터 학습된 분류기를 통해 가장 좋은 자동분류의 성능을 이끌어 낼 수 있게 된다.

입력문서에 대해 각 분류기가 발산하는 최대의 개념강도를 측정하기 위한 MCR 기법의 알고리즘을 설명하면 다음과 같다.

$$d^{pred} = decision[c_j, v_{c_j}^{\max}] = \arg max_{c_j \in C} \sum_i vs(f_i, c_j) \qquad (5)$$

앞서 설명한 공식(4)에서는 개별 문서에 범주를 할당하기 위해 최종 주제범주 c를 결정하였다. 이때의 최종 가중치 값을 측정하여 개념의 최대 출력값으로 사용할 수 있다. 공식(5)와 같이 문헌 d의 최대개념 강도 예측값을 d^{pred}로



〈그림 1〉 범주를 부여할 신규문서와 목표 데이터베이스 환경

표현하면 최종 할당되는 주제범주 c와 그때의 최대 가중치의 합 v의 쌍(c,v)를 $decision[c,v]$와 같이 리스트로 표현할 수 있다. 이와 같이 MCR 기법에 적용하기 위해서는 FV 분류기에서 사용했던 주제범주와 가중치 값이 모두 필요하다. d^{pred}은 모든 자질값의 범주별 가중치 합을 투표방식으로 계산하고, 최종적으로 선택된 한 개의 주제범주와 최대값의 쌍을 의미한다.

$$MCR_{single}(d_j) = decision_{d_j}[D_k, c] = \arg max_{D_k \in domains}(d_j^{pred}[2], D_k) \quad (6)$$

공식(6)은 개별문서의 최대개념강도를 측정한 결과인 MCR_{single}을 설명하고 있다. MCR_{single}은 개별 문서가 어떤 데이터베이스에 가장 유사한가와 함께 해당 분류를 동시에 측정하는 과정이다. 기구축된 데이터베이스 중 k번째인 D_k에 대해 신규 입력데이터의 j번째 문헌인 d_j가 갖는 최대의 개념 출력값을 측정하여 여러 데이터베이스 중 선택된 D와 그때의 주제범주 C를 $decision[D,c]$의 리스트로 출력한다. $d^{pred}[2]$는 결과리스트의 값 중 두 번째 요소인 최대 개념 출력값 v를 이용하였다는 의미이다.

$$MCR_{multi}(set\{d\}) = desicion[D_k] = \arg max_{D_k \in domains} \sum_i binary(MCR_{single}(d_i)[1]) \quad (7)$$

MCR_{multi}는 MCR_{single}의 모든 데이터베이스 선택 결과를 누적하여 최종득표가 많은 데이터베이스를 최종 선정하는 과정이다. 공식(7)은 신규문헌의 전체집합인 set{d}에 대해 개별문헌의 데이터베이스 선택결과인 MCR_{single}(d)를 산출하여 데이터베이스 별로 최종 누적득표가 가장 많은 데이터베이스를 선택($decision[D]$)하는 것이다. $MCR_{single}(d)[1]$은 결과리스트의 첫 번째 요소 값인 특정 데이터베이스 D_k를 투표에 이용하였음을 의미한다.

본 기법은 알고리즘 설명을 통해 기술한 바와 같이 최적의 데이터베이스를 선정함과 동시에 중간단계에서 개별문서의 자동범주화가 이루어지는 과정을 거치므로, 최적의 데이터베이스를 찾아내는 동시에 자동분류가 수행될 수 있는 추가적인 장점을 가지고 있다.

3. 실험설계

3.1 실험 데이터

실험을 위한 데이터로 4가지 유형의 데이터를 이용하였다. KISTI의 글로벌동향정보브리핑(과학기술 동향정보 기사) 4만여 건을 실험을 위한 데이터로 선정하였고, 다양한 과학기술분야의 국내외 학술논문 데이터를 이용하였다. 주제범주화 실험을 위해 주제분류 코드는 국가과학기술표준분류체계(A-S 19개 과학기술 대분류)를 이용하였다.[1] 실험에 이용한 데이터 속성과 문서의 수는 〈표 1〉에서 보는 바와 같다.

1) 과학기술기본법 제27조(국가과학기술표준분류체계의 확립) 및 동법 시행령 제41조에 따른 과학기술표준분류체계임. 이 연구에서는 2005년판을 기준으로 사용하였음.

〈표 1〉 실험을 위한 Data Collection

실험문서 집단	GTB[2]	SOC[3]	NDS[4]	GNS(통합환경)	BIST[5]
문서내용	과학기술동향 (기사브리핑)	국내학술논문	해외학술논문	GTB+SOC +NDS	해외학술논문 (한글키워드가공)
분류체계 통일방법	국가과학기술 표준분류(기준)	DDC-과기표준 매핑	DDC-과기표준 매핑	국가과학기술 표준분류(기준)	DDC-과기표준 매핑
추출키워드 언어	한/영	한/영	영어	한/영	한/영
범주의 수	국가과학기술표준분류 18개(S 범주인 과학기술정책만 제외)				
학습문서의 수 (복수분류고려)	20,217	24,613	21,510	66,340	없음 (실험용으로만 사용)
실험문서의 수 (복수분류고려)	20,080	24,781	21,896	없음 (학습용으로만 사용)	41,807

3.2 실험방법

실험은 크게 세단계로 진행을 하였다. 1단계는 전처리 단계에서의 다양한 실험을 통해 분류결과가 전처리에 의해 왜곡되지 않도록 하였으며, 2단계는 교차분류 실험을 통해 이기종의 데이터베이스 분류실험의 결과를 비교하였다. 이 과정에서 통합된 학습모델에 대한 실험도 실시하였다. 마지막 3단계에서 MCR 기법을 적용하고 그 성능을 측정하였다. 분류기를 비롯하여 새로운 기법의 적용 실험을 위한 소프트웨어는 모두 직접 개발하여 사용하였다.

3.2.1 1단계: 전처리과정의 비교실험

논문의 자질은 저자의 키워드를 이용하였고, 데이터 희소성에 의해 자동분류 성능측정이 왜곡되지 않도록 총 8가지 유형으로 구분하여 실험하였다. 스테밍과 형태소분석을 위해 porter stemmer와 KLT 2.1.0을 사용하였다. 언급한 8가지 유형은 다음과 같고, 이로부터 생성된 자질의 구축건수는 〈표 2〉와 같다.

① 저자 키워드의 원형을 그대로 사용
② 키워드의 원형에서 공백문자(space)만 제거(①+스페이스 제거)
③ 키워드 중 3어절 이상은 2어절 바이그램(Bigram)을 추가 생성
④ 키워드 바이그램(Bigram)을 생성 후 공백문자를 제거(③+스페이스 제거)
⑤ 스테밍(영어)과 형태소분석(한글)을 함
⑥ 스테밍, 형태소분석을 하고 공백문자를

2) GTB(http://http://radar.ndsl.kr/tre_index.do): 한국과학기술정보연구원에서 제공하는 글로벌 동향 브리핑(Global Trends Briefing) 정보. 해외 과학기술동향 정보를 일일단위로 제공.
3) SOC(http://society.kisti.re.kr/): 과학기술학회마을. 한국과학기술정보연구원에서 제공하는 국내 과학기술 학회정보 제공 사이트.
4) NDS(http://scholar.ndsl.kr/): 한국과학기술정보연구원에서 제공하는 NDSL 과학기술정보 통합서비스의 논문서비스(본 실험을 위해서는 해외학술논문 정보만을 사용하였음).
5) BIST: 한국과학기술정보연구원에서 구축한 해외 과학기술전문정보 데이터베이스. 주제별 전문가들이 부여한 한글 타이틀, 키워드, 주제분류 정보들이 가공되어 있음(현재는 온라인 서비스하지 않음).

〈표 2〉 자질값 전처리 방식에 따른 자질 생성건수 비교

전처리 형식 구분	GTB	SOC	NDS	GNS(통합환경)
① full(원형)	51,929	130,617	73,050	237,320
② full_space(공백)	50,448	126,364	72,870	231,156
③ full_bigram(바이그램)	63,563	177,207	97,840	314,629
④ full_space_bigram	61,907	172,005	97,601	307,150
⑤ stem(스테밍)	69,061	178,533	88,317	446,751
⑥ stem_space	67,520	173,849	88,139	440,713
⑦ stem_bigram	143,482	223,418	111,939	540,310
⑧ stem_space_bigram	141,407	216,630	111,633	531,781

제거(⑤+스페이스 제거)

⑦ 스테밍, 형태소분석을 하고 바이그램(Bigram) 추가생성

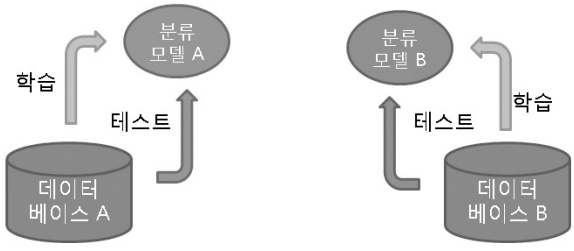
⑧ 스테밍, 형태소분석을 하고 바이그램(Bigram) 생성 후 공백문자 제거(⑦+스페이스 제거)

3.2.2 2단계: 기본 분류실험

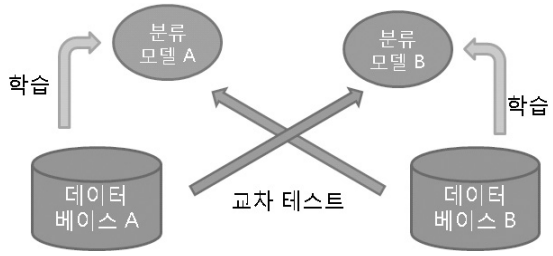
1단계의 8가지 자질값 전처리 방식에 대해 기초실험 3가지를 실시하였다. 2단계의 교차실험에 대한 실험개요도는 아래와 같다. 우선, 일반적인 학습 환경으로써 개별 데이터베이스 환경에서 분류기를 생성하고 이를 이용해 실험결과를 측정한 후(그림 2 참조), 같은 환경에서 실험 문서만을 교차로 적용하여 정확률을 측정하였다(그림 3 참조). 마지막으로 전체 데이터베이스를 통합하여 하나의 분류기를 생성한 후 개별 데이터베이스의 실험 셋을 각각 적용하여 데이터베이스 통합환경 하에서의 정확률을 측정하였다(그림 4 참조). 모든 실험에서 제시하는 정확률 값은 다양한 데이터베이스가 주제별 데이터 분포에 민감하지 않도록 마이크로 정확률로 산출하였다.

3.2.3 3단계: 최대 개념강도 인지(MCR) 기법의 적용

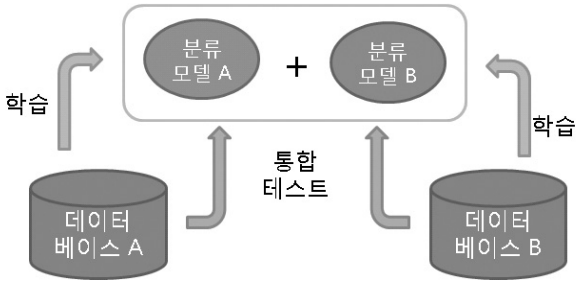
3단계 실험은 본 연구에서 제안한 새로운 기법인 MCR 기법을 적용한 데이터베이스의 자동선정 결과가 실제 데이터베이스의 구성과 정확히 일치하는가를 확인하는 실험이다. 성능을



〈그림 2〉 실험 1: 개별 데이터베이스별 분류실험



〈그림 3〉 실험 2: 데이터베이스 교차적용 테스트



〈그림 4〉 실험 3: 통합된 학습 환경에 대해 데이터베이스별 분류 적용실험

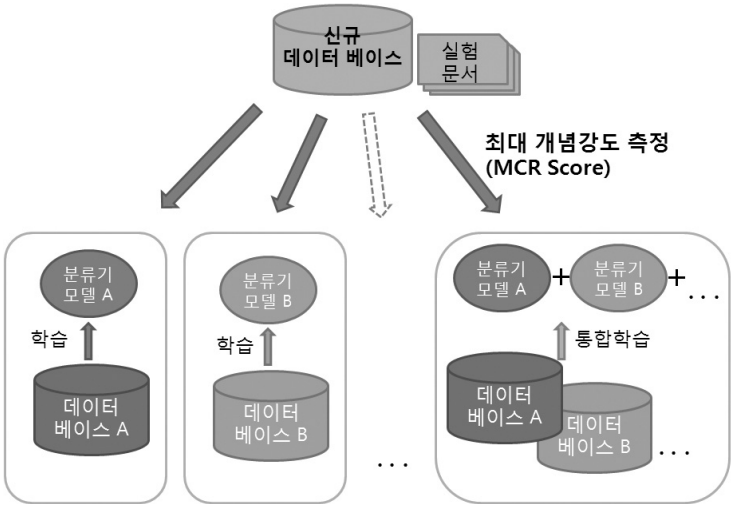
측정하기 위해 세가지 종류의 데이터베이스로 다양한 테스트 환경을 구축하고 MCR을 구현하기 위해 직접 개발한 소프트웨어를 적용하여 실험을 하였다.

마지막으로 새로운 데이터가 입수되었을 경우에도 MCR 기법이 기존의 데이터베이스 중에서 가장 유사한 데이터베이스를 찾아낼 수 있는가를 알아보기 위해, 신규 데이터(BIST)를 'GTB', 'SOC', 'NDS', 'GNS 통합'의 4가지 학습모델에 적용하여 MCR 스코어와 분류성능을 비교하는 마지막 실험을 수행하였다(그림 5 참조).

4. 실험 결과

4.1 전처리 유형에 따른 자동분류 성능변화

〈표 3〉에서 보는 바와 같이 8가지 전처리 유형에 따라 각 데이터베이스별 마이크로 정확률을 측정하였다. 키워드의 공백문자를 제거하거나 스테밍, 형태소 분석 등을 통해 어형의 통제효과를 다소 가져올 수 있으며, 바이그램을 생성하거나 형태소 분석 방법을 통해 데이터의 희소성 문제를 다소 완화할 수 있다. 실험결과, 〈그림 6〉과 같이 방법의 변화에 따라 성능의 차이가 나타나고 있다. 키워드의 원형을 이용하는 경우에는, 공백을 제거하는 방



〈그림 5〉 신규 문서의 데이터베이스 별 MCR 측정을 위한 분류기모델 구성환경

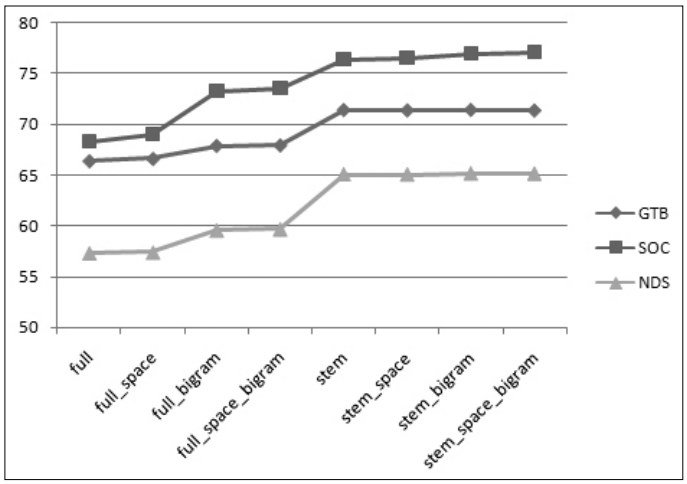
법에 비해 3어절 이상의 색인어에 바이그램을 추가 생성하여 자질수를 늘려주는 방법이 효과적이었으며, 스테밍과 형태소 분석 결과를 이용하는 방법이 원형을 이용하는 다양한 방법에 비해 성능이 좋게 나타났다. 그러나, 스테밍과 형태소 분석을 이용하는 경우에는 바이그램을 추가로 생성하여 자질 수를 크게 늘려도 성능향상에는 큰 차이가 없음을 확인할 수 있다(표 3 참조).

4.2 학술정보 데이터베이스 교차적용 결과

〈표 4〉~〈표 6〉의 결과표에서 나타나는 것처럼, 개별 학습모델에 실험문서 셋을 교차 적용하였을 때, 학습문서와 실험문서가 동일한

〈표 3〉 8가지 전처리 유형에 따른 자동분류 정확률의 변화

실험문서	GTB	SOC	NDS
학습모델	GTB	SOC	NDS
full(원형)	66.399	68.306	57.344
full_space(공백)	66.628	68.968	57.421
full_bigram(바이그램)	67.908	73.209	59.591
full_space_bigram	67.958	73.548	59.705
stem(스테밍)	71.444	76.389	65.103
stem_space	71.404	76.555	65.08
stem_bigram	71.454	76.966	65.19
stem_space_bigram	71.394	77.063	65.167



〈그림 6〉 전처리에 따른 정확률 향상 그래프

〈표 4〉 GTB 실험문서 셋에 대한 이기종 데이터베이스 교차실험

실험문서	GTB 실험문서 셋			
학습모델	GTB(%)	SOC(%)	NDS(%)	GSN통합(%)
full(원형)	66.399	25.558	30.603	65.742
full_space(공백)	66.628	26.076	30.852	65.812
full_bigram(바이그램)	67.908	28.217	32.46	67.012
full_space_bigram	67.958	28.625	32.639	67.047
stem(스테밍)	71.444	34.98	42.869	66.534
stem_space	71.404	34.925	42.903	66.614
stem_bigram	71.454	35.458	42.918	66.678
stem_space_bigram	71.394	35.418	42.968	66.688

〈표 5〉 SOC 실험문서 셋에 대한 이기종 데이터베이스 교차실험

실험문서	SOC 실험문서 셋			
학습모델	GTB(%)	SOC(%)	NDS(%)	GSN통합(%)
full(원형)	16.065	68.306	23.292	68.371
full_space(공백)	16.589	68.968	23.421	69.009
full_bigram(바이그램)	19.523	73.209	29.099	72.947
full_space_bigram	20.064	73.548	29.123	73.129
stem(스테밍)	34.292	76.389	45.022	73.69
stem_space	34.305	76.555	45.083	73.79
stem_bigram	34.43	76.966	45.301	74.367
stem_space_bigram	34.51	77.063	45.349	74.493

〈표 6〉 NDS 실험문서 셋에 대한 이기종 데이터베이스 교차실험

실험문서	NDS 실험문서 셋			
학습모델	GTB(%)	SOC(%)	NDS(%)	GSN 통합(%)
full(원형)	27.85	26.014	57.344	57.504
full_space(공백)	27.868	26.078	57.421	57.545
full_bigram(바이그램)	30.225	28.768	59.591	59.623
full_space_bigram	30.234	28.809	59.705	59.682
stem(스테밍)	40.939	33.581	65.103	60.614
stem_space	40.939	33.581	65.08	60.586
stem_bigram	40.848	33.897	65.19	61.194
stem_space_bigram	40.848	33.887	65.167	61.221

데이터베이스가 아닌 경우에는 정확률이 현저하게 떨어짐을 확인할 수 있다. 3종의 데이터베이스의 학습문서를 모두 통합하여 학습한 모델(GNS 통합)의 경우에는 용어학습의 평탄화(Smoothing)가 이루어져서 비교적 최고치에 근접하는 성능을 보이고 있음을 알 수 있다. 그러나, 실제 환경에서는 대용량 학습에 제한점이 존재하므로, 본 연구에서는 MCR의 성능 비교실험 용으로만 구축하여 사용하였다. 이와 관련하여 실제 대용량 환경을 위한 자동분류기 구축에 대한 연구는 추후 별도로 진행할 예정이다.

4.3 MCR 기법을 적용한 최적의 데이터베이스 선택

MCR 기법을 이용해 최적의 데이터베이스를 선택하는 실험을 수행하여 다음과 같은 결과를 얻을 수 있었다. 우선, GTB 실험문서 셋을 이용해 세 가지 데이터베이스에 대해 MCR 스코어를 측정하였고 GTB 데이터베이스를 정확히 선택함을 확인하였다(표 7 참조). 자동분류 성능을 측정한 결과 〈표 8〉에서 보는 바와 같이 GTB 데이터베이스에 직접 분류를 수행하는 경우, 통합된 분류환경을 이용하는 경우와 MCR 스코어를 측정함과 동시에 수행한 자동분류 성능이 다른 이기종 데이터베이스인 SOC과 NDS에 대한 분류성능에 비해 높게 나타나고 있다. 〈표 9〉~〈표 12〉는 다른 실험문서 셋인 SOC, NDS의 경우에도 일관성 있는 결과가 나타나고 있음을 보여주고 있다

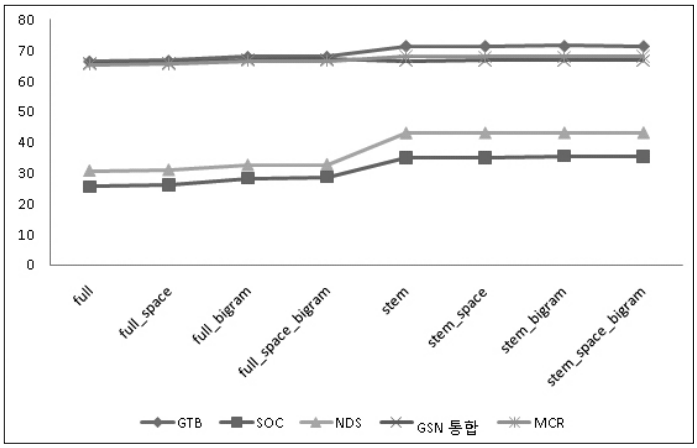
이로써 MCR 기법을 이용하여 다양한 데이터베이스 구축환경에서 자동분류의 성능을 높이기 위해 가장 적합한 데이터베이스를 선택할 수 있음을 확인하였다. 본 실험에서는 자동분류의 정확률은 학습문서와 실험문서를 같은 데이터베이스로 일치시키는 경우 대체로 가장 좋은 성능을 보이며, 통합 환경과 MCR을 이용하는 자동분류 성능은 우열을 가리기 힘들었다.

〈표 7〉 GTB 실험문서 셋의 MCR 최적화 적용 및 데이터베이스 별 선택비율 측정

실험문서	GTB 실험문서 셋		
학습모델	MCR 기법에 의한 데이터베이스 선택 비율		
	GTB(%)	SOC(%)	NDS(%)
full(원형)	87.68	8.055	4.265
full_space(공백)	88.013	8.247	3.74
full_bigram(바이그램)	86.728	8.392	4.88
full_space_bigram	87.166	8.588	4.246
stem(스테밍)	83.013	16.265	0.722
stem_space	83.127	16.165	0.707
stem_bigram	82.928	16.121	0.951
stem_space_bigram	83.108	15.976	0.916

〈표 8〉 GTB 문서 셋으로 MCR 기법을 이용한 자동범주화 결과비교

실험문서	GTB 실험문서 셋				
학습모델	GTB(%)	SOC(%)	NDS(%)	GSN통합(%)	MCR(%)
full(원형)	66.399	25.558	30.603	65.742	65.359
full_space(공백)	66.628	26.076	30.852	65.812	65.613
full_bigram(바이그램)	67.908	28.217	32.46	67.012	66.529
full_space_bigram	67.958	28.625	32.639	67.047	66.549
stem(스테밍)	71.444	34.98	42.869	66.534	67.988
stem_space	71.404	34.925	42.903	66.614	67.913
stem_bigram	71.454	35.458	42.918	66.678	67.943
stem_space_bigram	71.394	35.418	42.968	66.688	68.003



(GTB ≒ GSN 통합 ≒ MCR 〉 SOC ≒ NDS)

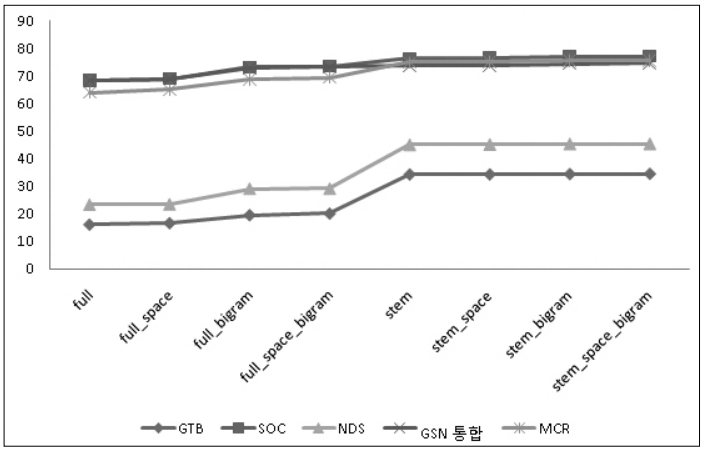
〈그림 7〉 GTB 셋의 자동분류 결과

〈표 9〉 SOC 실험문서 셋의 MCR 최적화 적용 및 데이터베이스별 선택비율 측정

실험문서	SOC 실험문서 셋		
학습모델	MCR 기법에 의한 데이터베이스 선택 비율		
	GTB(%)	SOC(%)	NDS(%)
full(원형)	10.223	82.505	7.271
full_space(공백)	10.108	83.46	6.432
full_bigram(바이그램)	8.348	83.642	8.01
full_space_bigram	8.268	84.382	7.351
stem(스테밍)	4.952	94.241	0.807
stem_space	4.968	94.249	0.783
stem_bigram	4.774	94.281	0.944
stem_space_bigram	4.77	94.318	0.912

〈표 10〉 SOC 문서 셋으로 MCR 기법을 이용한 자동범주화 결과비교

실험문서	SOC 실험문서 셋				
학습모델	GTB(%)	SOC(%)	NDS(%)	GSN통합(%)	MCR(%)
full(원형)	16.065	68.306	23.292	68.371	64.045
full_space(공백)	16.589	68.968	23.421	69.009	65.09
full_bigram(바이그램)	19.523	73.209	29.099	72.947	68.819
full_space_bigram	20.064	73.548	29.123	73.129	69.396
stem(스테밍)	34.292	76.389	45.022	73.69	75.029
stem_space	34.305	76.555	45.083	73.79	75.211
stem_bigram	34.43	76.966	45.301	74.367	75.614
stem_space_bigram	34.51	77.063	45.349	74.493	75.659



(SOC ≒ GSN 통합 ≒ MCR 〉 GTB ≒ NDS)

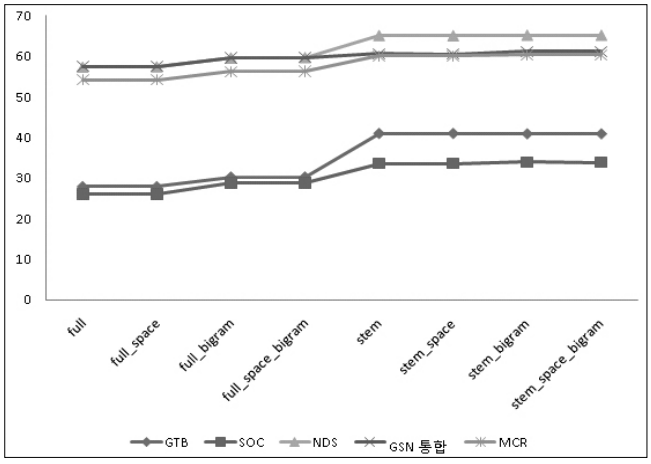
〈그림 8〉 SOC 셋의 자동분류 결과

〈표 11〉 NDS 셋의 MCR 최적화 적용 및 데이터베이스 별 선택비율 측정

실험문서	NDS 실험문서 셋		
학습모델	MCR 기법에 의한 데이터베이스 선택 비율		
	GTB(%)	SOC(%)	NDS(%)
full(원형)	14.844	16.431	68.725
full_space(공백)	14.894	16.461	68.645
full_bigram(바이그램)	13.327	17.203	69.47
full_space_bigram	13.383	17.262	69.355
stem(스테밍)	9.714	22.454	67.832
stem_space	9.709	22.436	67.855
stem_bigram	9.696	21.919	68.385
stem_space_bigram	9.686	21.91	68.404

〈표 12〉 NDS 셋으로 MCR 기법을 이용한 자동범주화 결과비교

실험문서	NDS 실험문서 셋				
학습모델	GTB(%)	SOC(%)	NDS(%)	GSN통합(%)	MCR(%)
full(원형)	27.85	26.014	57.344	57.504	54.188
full_space(공백)	27.868	26.078	57.421	57.545	54.225
full_bigram(바이그램)	30.225	28.768	59.591	59.623	56.22
full_space_bigram	30.234	28.809	59.705	59.682	56.289
stem(스테밍)	40.939	33.581	65.103	60.614	60.084
stem_space	40.939	33.581	65.08	60.586	60.057
stem_bigram	40.848	33.897	65.19	61.194	60.372
stem_space_bigram	40.848	33.887	65.167	61.221	60.34



(NDS ≒ GSN 통합 ≒ MCR 〉 GTB ≒ SOC)

〈그림 9〉 NDS 셋의 자동분류 결과

마지막 실험은 실험에 사용했던 데이터베이스인 'GTB, SOC, NDS, GNS통합' 4가지 학습모델에 해당되지 않는 외부의 새로운 문서정보셋을 적용하는 것이다. 이를 위해 KISTI의 해외과학기술서지정보(BIST) 데이터 4만여 건을 실험에 이용하였다. 〈표 13〉에서 보는 바와 같이 신규 BIST 데이터는 MCR 스코어를 측정한 결과, 해외학술논문 데이터인 NDS 학습모델에 가장 유사한 것으로 나타났다. 이 결과는 본 연구의 가설과 실험결과에 따르면, BIST 데이터가 기 구축된 여러 데이터베이스 중에서 NDS 데이터베이스와 가장 개념적으로 유사하다는 의미를 가지는 것이다.

최종 분류실험을 통해 측정한 결과, 이전의 모든 실험과 동일하게 MCR 기법이 선택한 데이터베이스인 NDS에 분류하는 경우, 통합된 분류환경을 이용하는 경우와 MCR 스코어를 측정함과 동시에 수행한 자동분류 성능이 다른 이기종 데이터베이스인 GTB나 SOC에 대한 분류성능에 비해 높게 나타나고 있다(표 14 참조).

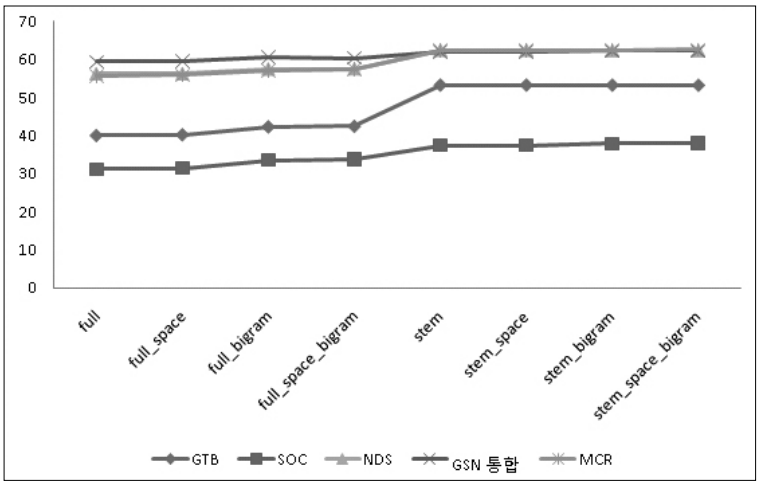
〈표 13〉 새로운 실험문서(BIST)를 적용하여 최적의 학습모델 자동선택

실험문서	신규 BIST 실험문서 셋		
학습모델	MCR 기법에 의한 데이터베이스 선택 비율		
	GTB(%)	SOC(%)	NDS(%)
full(원형)	19.59	14.101	66.309
full_space(공백)	19.33	14.491	66.179
full_bigram	17.851	13.936	68.213
full_space_bigram	17.554	14.321	68.125
stem(스테밍)	13.067	14.713	77.22
stem_space	12.959	14.84	72.237
stem_bigram	12.895	14.498	72.608
stem_space_bigram	12.663	14.699	72.639

〈표 14〉 새로운 실험문서(BIST) 셋으로 MCR 기법을 이용한 자동범주화 결과비교

실험문서	신규 BIST 실험 문서셋				
학습모델	GTB(%)	SOC(%)	NDS(%)[6] (영문 37,926건)	GSN통합(%)	MCR(%)
full(원형)	40.02	31.201	56.187(59.657)	59.406	55.62
full_space(공백)	40.23	31.447	56.215(59.7)	59.54	55.912
full_bigram	42.292	33.499	57.454(62.709)	60.619	57.108
full_space_bigram	42.428	33.808	57.51(62.751)	60.332	57.364
stem(스테밍)	53.178	37.434	62.327(67.479)	62.006	62.317
stem_space	53.178	37.439	62.322(67.474)	62.014	62.327
stem_bigram	53.156	37.91	62.248(67.587)	62.243	62.511
stem_space_bigram	53.139	37.998	62.241(67.582)	62.214	62.53

6) NDS 데이터는 모두 영문정보이므로 신규정보인 BIST 데이터 중 분류할 수 없는 한글정보를 제외한 성능을 ( )안에 별도표기함. GNS와 MCR은 언어에 따른 제한이 없이 수행되었음.



(NDS ≒ GSN 통합 ≒ MCR 〉 GTB ≒ SOC)

〈그림 10〉 신규 정보(BIST)의 자동분류 결과

5. 결 론

본 연구에서는 최대 개념강도 인지기법(MCR)을 제안하여, 자동분류를 수행할 때 가장 큰 제약으로 존재하는 일반화된 대용량 학습환경 구축의 어려움을 해결하기 위한 대안으로써 여러 데이터베이스 중에서 가장 분류성능을 높게 향상시킬 수 있는 최적의 유사 데이터베이스를 찾아내는 방법을 제시하고자 하였다. MCR 기법의 특징과 다양한 실험을 통해 얻어진 결과를 요약하면 다음과 같다.

① MCR 기법을 이용하면 질의문서(자동분류 대상문서 집합)와 가장 유사한 데이터베이스를 찾을 수 있다.

② 즉, 대용량 학습모델 구축을 통한 범주화가 어려운 환경에서는 분산된 개별 학습모델 중에 최적의 모델을 찾을 수 있다.

③ 제3의 새로운 정보원에 대해서도 자동범주 할당을 위한 최선의 선택을 할 수 있다.

④ MCR을 이용해 최적의 데이터베이스를 찾을 뿐만 아니라 동시에 자동분류를 수행할 수 있으며 대용량 학습모델 환경과 비교하였을 때 대체로 비슷한 성능을 보인다.

향후 데이터베이스 수와 학술정보의 규모를 다양화하여 실험의 결과를 일반화하는 연구를 수행해야 한다. 또한 본 연구에서 제시한 분류모델을 다른 모델들과 비교하여 성능을 검증하는 실험이 필요할 것이다.

본 기법은 다양한 데이터로 구성된 대용량 분산 환경에서 더욱 효과적인 식별방법을 제공하기 때문에, 향후 추가연구를 통해 분산 환경에서의 시맨틱 에이전트 간 의사소통 기술로 발전시킬 수 있을 것으로 기대한다.

참 고 문 헌

국가과학기술표준분류체계. [online]. [cited 2010. 7.10]. 〈http://www.kistep.re.kr/major/duty/plan_02_05.jsp〉.

이재윤. 2005. 문서측 자질선정을 이용한 고속 문서분류기의 성능향상에 관한 연구. 『정보관리연구』, 36(4):51-69.

정영미. 2005. 『정보검색연구』. 서울: 구미무역출판부.

Deng, Z. H., S. W. Tang, D. Q. Yang, M. Zhang, X. B. Wu, and M. Yang. 2002. "Two odds-radio-based text classification algorithms." *Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops)*, 223-231.

Ko, Y. and J. Seo. 2004. "Using the feature projection technique based on a normalized voting method for text classification." *Information Processing and Management*, 40(2): 191-208.

Nuray, R. and F. Can. 2005. "Automatic ranking of information retrieval systems using data fusion." *Information Processing and Management*, 42(3): 595-614.

Salton, G. and C. Buckley. 1988. "Weighting approaches in automatic text retrieval." *Information Processing and Management*. 24(5): 513-523.

Voorhees, E. M., N. K. Gupta, and B. Johnson-Laird. 1995. "Learning collection fusion strategies." *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 172-179.

Witten, I. H. and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann.