한국일보-20000/한국일보-40075 문서범주화 실험문서집합
HKIB-20000/HKIB-40075 Korean Text Categorization Test Collections

설명서 (버전 1.0) / 2009년 5월 31일

Group for Intelligent Information Systems / 한국과학기술정보연구원 / 한국일보
www.kristalinfo.com / www.kisti.re.kr / news.hankooki.com

1. 개요

이 설명서는 '한국일보-20000'(HKIB-20000)과 '한국일보-40075'(HKIB-40075) 실험문서집합(test collection)에 대한 설명을 담고 있습니다. 이 두 가지의 실험문서집합은 정보검색(information retrieval), 자동문서범주화(automatic text categorization), 기계학습(machine learning) 등과 같이 코퍼스 기반의 연구에 활용될 수 있습니다.

한국일보-40075 실험문서집합은 한국일보가 제공한 1998~1999년의 2년간 신문기사를 바탕으로 40,075개의 각 문서별로 3단계 분류체계의 말단 범주를 부여하여 구축하였습니다. 이 문서집합은 충남대학교 이석훈 교수 연구실과 한국과학기술정보연구원이 공동으로 제작하였습니다. 한국일보-20000 실험문서집합은 한국일보-40075 집합의 기사 중 20,000건을 별도로 추출하여 분류체계를 보다 현실적으로 수정하였으며, 3단계 분류체계의 모든 노드에 기사를 할당하여 구축한 계층적 분류체계의 문서범주화용 실험문서집합입니다. 한국일보-20000 실험문서집합은 한국과학기술정보연구원과 연세대학교 나동열 교수 연구실이 공동으로 구축하였습니다.

한국일보-40075는 비계층적 분류체계를 가지는 단일범주 실험문서집합인 반면, 한국일보-20000은 계층적 분류체계를 가지는 다중범주 실험문서집합으로 요약할 수 있습니다 (한국일보-40075의 분류체계 자체는 3단계 계층형이지만 문서에 대한 범주 부여 방식이 소분류만을 대상으로 하기 때문에 비계층적입니다).

2. 저작권

한국일보-40075 또는 한국일보-20000 실험문서집합을 사용하여 연구 결과를 발표할 경우에는 해당 실험문서집합 이름과 배포 주소를 명기하여 주시기 바랍니다. (배포 주소는 3장을 참조하십시오.) 영문으로 연구결과를 발표하실 경우에는 한국일보-40075 실험문서집합의 경우 `HKIB-40075 Test Collection`으로 한국일보-20000의 경우에는 `HKIB-20000 Test Collection`으로 표기할 수 있습니다.

3. 실험문서집합의 배포

한국일보-20000과 한국일보-40075 실험문서집합은 하나의 파일뭉치(gzipped tar archive)로 배포됩니다. 배포 주소는 다음과 같습니다.

* URL: `http://www.kristalinfo.com/TestCollections/#hkib`

배포판 파일뭉치는 gzip으로 압축된 tar 파일(`hkib-20000-40075.tar.gz`, 42.9MB)이며 그 내용물은 다음과 같습니다. (각 파일의 내용에 대해서는 5장과 6장을 참조하십시오.)

```
readme.txt
HKIB-20000/
        HKIB-20000_001.txt                  (한국일보-20000 문서 파일 1)
        HKIB-20000_002.txt                  (한국일보-20000 문서 파일 2)
        HKIB-20000_003.txt                  (한국일보-20000 문서 파일 3)
        HKIB-20000_004.txt                  (한국일보-20000 문서 파일 4)
        HKIB-20000_005.txt                  (한국일보-20000 문서 파일 5)
        hkib20000-cat03-all.categories
        hkib20000-cat03-file1.categories
        hkib20000-cat03-file2.categories
        hkib20000-cat03-file3.categories
        hkib20000-cat03-file4.categories
        hkib20000-cat03-file5.categories
        hkib20000-cat07-all.categories
        hkib20000-cat07-file1.categories
        hkib20000-cat07-file2.categories
        hkib20000-cat07-file3.categories
        hkib20000-cat07-file4.categories
        hkib20000-cat07-file5.categories
HKIB-40075/
        HKIB-40075_001.txt                  (한국일보-40075 문서 파일 1)
        HKIB-40075_002.txt                  (한국일보-40075 문서 파일 2)
        HKIB-40075_003.txt                  (한국일보-40075 문서 파일 3)
        HKIB-40075_004.txt                  (한국일보-40075 문서 파일 4)
        HKIB-40075_005.txt                  (한국일보-40075 문서 파일 5)
        hkib40075-cat03-all.categories
        hkib40075-cat03-file1.categories
        hkib40075-cat03-file2.categories
        hkib40075-cat03-file3.categories
        hkib40075-cat03-file4.categories
        hkib40075-cat03-file5.categories
```

배포파일에 포함된 전체 내용물은 본 설명서 파일(`readme.txt`)과 `HKIB-20000`, `HKIB-40075`의 2개 디렉토리로 구성되어 있습니다. `readme.txt` 파일에는 본 설명서 파일에 대한 링크를 담고 있습니다.

배포파일에 포함된 두 개의 디렉토리에는 각각 한국일보-20000 실험문서집합(`HKIB-20000/` 디렉토리)과 한국일보-40075 실험문서집합(`HKIB-40075/` 디렉토리)이 포함되어 있습니다. 각 디렉토리에는 `.txt` 확장자를 가지는 실제 문서/범주 파일이 5개씩 포함되어 있습니다. 각 데이터 파일에 대해서 범주당 문서수를 `*-file?.categories` 파일들에 수록하였습니다. 실험문서집합 전체의 범주당 문서수는 `*-all.categories` 파일에 포함되어 있습니다. 보다 자세한 표기 양식은 5장과 6장을 참조하십시오.

4. 실험문서집합 구축이력

한국일보 실험문서집합은 한국일보의 1998년 1월 1일부터 1999년 12월 31일까지 만2년간의 뉴스기사를 포함하고 있습니다. 한국과학기술정보연구원(KISTI)은 인터넷한국일보와의 협약을 통해 이 신문기사들을 연구용 목적으로 제공받아, 충남대학교 이석훈 교수 연구실과 협동으로 각 기사에 대한 범주 할당 작업을 수행하였습니다. 이 결과로 2003년에 한국일보-40075 실험문서집합이 구축되었습니다.

한국일보-40075 실험문서집합의 특징은 대분류, 중분류, 소분류의 3단계 계층을 가지는 총 120개의 말단 범주를 구성하고 40,075개의 문서에 대해 각각 1개의 말단 범주를 부여하였다는 것입니다. 이 실험집합에 사용된 분류체계(2003분류체계)는 계층형이긴 하나 모든 문서는 소분류의 말단 범주 1개씩을 가지고 있습니다.

반면, 실질적으로 문서나 웹 사이트의 분류가 이루어지고 있는 현장의 경우에는 분류체계의 말단 노드뿐만 아니라 중간 노드의 범주에도 문서들이 할당되고 있습니다. 또한 하나의 문서는 1개 이상의 범주를 가지는 경우도 다수 발생하고 있습니다. 이러한 현장의 특성을 반영할 수 있는 실험문서집합의 구축이 요구되어 2007년에는 한국일보-40075에 수록된 신문기사 중에서 20,000개 문서를 대상으로 계층적 분류체계를 갱신하고 모든 범주 노드에 대해서 문서가 할당되면서, 각 문서는 1개 이상의 다중범주를 가질 수 있는 한국일보-20000 실험문서집합이 구축되었습니다. 이 작업은 한국과학기술정보연구원과 연세대학교 나동열 교수 연구실이 공동으로 수행하였습니다.

5. 문서양식

한국일보-20000과 한국일보-40075 실험문서집합은 각각 5개의 텍스트 파일로 구성되어 있습니다(`.txt` 확장자를 가지는 파일들). 실험문서집합을 5개의 파일로 분할하여 구성한 것은 5겹 교차검증(5-fold cross validation)의 편이성을 위한 것입니다. 표1과 표2에 각 파일에 담긴 문서 수와 크기를 수록하였습니다.

[표1] 한국일보-20000(HKIB-20000)의 신문기사 파일

구분	문서수	파일명	크기(kb)
파일1	3988	HKIB-20000/HKIB-20000_001.txt	8,864
파일2	4048	HKIB-20000/HKIB-20000_002.txt	8,880
파일3	4029	HKIB-20000/HKIB-20000_003.txt	8,956
파일4	3913	HKIB-20000/HKIB-20000_004.txt	8,708
파일5	4022	HKIB-20000/HKIB-20000_005.txt	8,920
계	20000		44,332

[표2] 한국일보-40075(HKIB-40075)의 신문기사 파일

구분	문서수	파일명	크기(kb)
파일1	8123	HKIB-40075/HKIB-40075_001.txt	17,916
파일2	8012	HKIB-40075/HKIB-40075_002.txt	17,404
파일3	7922	HKIB-40075/HKIB-40075_003.txt	17,420
파일4	8040	HKIB-40075/HKIB-40075_004.txt	17,708
파일5	7978	HKIB-40075/HKIB-40075_005.txt	17,672
계	40075		88,124

표1,2에서 볼 수 있듯이 각 파일에 담긴 문서의 수는 균일하지 않은데, 이는 각 파일을 시험집합(test set)으로 사용할 때 어느 정도 범주 및 문서 양에서 편차를 주려는 의도에서 구성된 것입니다. 각 텍스트 파일별로 학습결과를 시험할 경우 약간씩 다른 성향을 보여줄 것으로 예상됩니다.

각 문서 파일은 일반 텍스트 문서이며 UTF-8로 인코딩되어 있습니다. 그리고 각 문서의 내용은 문서 구분자, 문서ID, 2003범주(2003년에 할당된 범주), 2007범주(2007년에 할당된 범주), 제목, 본문으로 구성되어있습니다. 하나의 파일 내에서 각각의 문서는 "@DOCUMENT"(9바이트)으로 구분되며, 문서의 ID는 "#DOCID : "(9바이트)로 시작하는 줄에, 2003범주는 "#CAT'03: "(9바이트)로 시작하는 줄에, 2007범주는 "#CAT'07: "(9바이트)로 시작하는 줄에, 기사 제목은 "#TITLE : "(9바이트)로 시작하는 줄에 표시됩니다. 기사 본문은 "#TEXT : "(9바이트)로 표시하는 다음 줄 부터 다음 문서구분자(@DOCUMENT)가 출현하기 직전까지입니다. 줄별로 각 구분자는 모두 순서대로 출현합니다. 다만 한국일보-40075 실험문서집합의 경우에는 2007범주("#CAT'07: "로 시작하는 줄)가 존재하지 않습니다.

한국일보 실험문서집합의 문서 표현형식

구분자	길이(byte)	내용	비고
@DOCUMENT	9	문서구분자	파일내에서의 문서간 구분자 역할
#DOCID :	9	문서 식별자	document identifier
#CAT'03:	9	2003범주	2003분류체계에 따라 부여된 범주
#CAT'07:	9	2007범주	2007분류체계에 따라 부여된 범주
#TITLE :	9	제목	기사 제목
#TEXT :	10	본문	기사본문. 구분자 끝에 '\n' 포함

다음은 한국일보-20000 실험문서집합에 수록된 신문기사의 예제입니다. 아래 예제에서 볼 수 있듯이 2007범주("#CAT'07: "로 표시된 줄)는 1개 이상의 범주를 가지며 각 범주는 세미콜론(;)으로 구분됩니다. 2003범주("#CAT'03: "으로 구분되는 줄)는 각 문서당 1개의 범주를 가집니다. (한국일보-40075 실험문서집합은 2003년에 범주부여 작업이 완료되었고, 한국일보-20000의 경우에는 2007년에 수정된 범주체제를 채택하였습니다. 이 문서에서는 한국일보-40075에 부여된 기존 범주를 2003범주, 한국일보-20000에서 새로 부여된 범주를 2007범주로 표기합니다.)

```
@DOCUMENT
#DOCID : 3789
#CAT'03: /여가생활/실외/여행관광
```

```
#CAT'07: /경제/수입;/경제/가계 물가
#TITLE :  상공부, 가격표시제대상 추가
#TEXT  :
   정부는 피라미드 판매의 주요 대상품목으로 꼽히고 있는 자석요를 가격표
   시 대상품목에 새로 넣어 공장도 가격이나 수입가격을 의무적으로 표시토
   록했다. (이하 본문 생략)

@DOCUMENT
#DocID : HKIB99-43141
#CAT'03: /사회/사회질서/사건사고(화재)
#CAT'07: /사회/사건사고/화재
#TITLE :  청소년 수련원서 불, 유치원생 23명 사망
#TEXT  :
```

위의 예에서 2003범주는 3단계로 표기되어 있는 반면 2007범주는 2단계 또는 3단계로 표기되어 있음을 알 수 있습니다. (실제로는 1단계, 즉 대분류 범주에도 문서가 할당되어 있습니다.) 이 점이 2003범주와 2007범주의 가장 큰 차이점중의 하나입니다. 즉 2003범주체계는 소분류가 항상 말단 범주이지만, 2007분류체계에서는 대분류, 중분류, 소분류 모두가 말단 범주가 될 수 있습니다. 보다 자세한 내용은 6장에서 설명합니다.

다음은 한국일보-40075 실험문서집합에 수록된 신문기사의 예제입니다. 2007범주가 없다는 점외에는 한국일보-20000 실험문서집합과 동일한 양식입니다.

```
@DOCUMENT
#DocID : 52054
#CAT'03: /건강과 의학/건강/체력단련
#TITLE : [골프단신] 94남자 프로테스트 용인 프라자CC에서 열려
#TEXT  :
   94남자프로골프 제1차테스트가 오는 25~28일 용인프라자CC에서
   치러진다. (이하 본문 생략)

@DOCUMENT
#DocID : 1193
```

한국일보-40075 집합 내의 모든 문서는 2003분류체계중 소분류 범주 1개가 부여되어 있습니다. 이러한 점 때문에 한국일보-40075는 단일범주 실험문서집합으로 한국일보-20000은 다중범주 실험문서집합이라고 할 수 있습니다.

6. 범주 구성

자동문서분류 실험문서집합(text categorization test collection)은 {문서와 이 문서에 대해 사람이 부여한 범주}의 집합으로 구성됩니다. 한국일보 실험문서집합의 경우 /대분류/중분류/소분류의 3단계 계층형 분류체계를 채택하고 있습니다. 다만 두 실험문서집합은 범주 부여 방식에 있어서 한국일보-40075는 비계층적으로 범주를 부여하였으며, 한국일보-20000은 계층적으로 범주를 부여한 점이 다릅니다.

4장의 이력에서 설명드린 바와 같이 한국일보 실험문서집합은 2가지의 분류체계를 가지고 있습니다. 2003년에 구축된 한국일보-40075 실험문서집합의 분류체계는 2003분류체계라고 명명하였으며, 2007년에 정제된 한국일보-20000 실험문서집합의 분류체계는 2007분류체계라 명명합니다. 한국일

보-20000 실험문서집합은 한국일보-40075 실험문서집합을 기반으로 구축되었기 때문에 2003분류체계에 따라 부여된 범주를 기본적으로 상속받았습니다. 따라서 한국일보-20000 실험문서집합은 2003분류체계("#CAT'03 : " 줄에 표시)와 2007분류체계("#CAT'07 : " 줄에 표시)를 동시에 수용하고 있습니다.

배포판 파일의 각 디렉토리에는 다음과 같이 말단 범주별 문서수를 표현한 파일들이 포함되어 있습니다. 2007 및 2003 분류체계 전체를 확인하고자 할 경우에는 HKIB-20000/hkib20000-cat07-all.categories과 HKIB-40075/hkib40075-cat03-all.categories를 참조하십시오. 2003분류체계에서는 소분류가 말단 범주이며, 2007분류체계에서는 대중소분류 어느 것이나 말단 범주가 될 수 있습니다.

```
HKIB-20000/
     hkib20000-cat07-all.categories          전체 컬렉션의 2007범주당 문서수
     hkib20000-cat07-file1.categories        파일1의 2007범주당 문서수
     hkib20000-cat07-file2.categories        파일2의 2007범주당 문서수
     hkib20000-cat07-file3.categories        파일3의 2007범주당 문서수
     hkib20000-cat07-file4.categories        파일4의 2007범주당 문서수
     hkib20000-cat07-file5.categories        파일5의 2007범주당 문서수
     hkib20000-cat03-all.categories          전체 컬렉션의 2003범주당 문서수
     hkib20000-cat03-file1.categories        파일1의 2003범주당 문서수
     hkib20000-cat03-file2.categories        파일2의 2003범주당 문서수
     hkib20000-cat03-file3.categories        파일3의 2003범주당 문서수
     hkib20000-cat03-file4.categories        파일4의 2003범주당 문서수
     hkib20000-cat03-file5.categories        파일5의 2003범주당 문서수
HKIB-40075/
     hkib40075-cat03-all.categories          전체 컬렉션의 2003범주당 문서수
     hkib40075-cat03-file1.categories        파일1의 2003범주당 문서수
     hkib40075-cat03-file2.categories        파일2의 2003범주당 문서수
     hkib40075-cat03-file3.categories        파일3의 2003범주당 문서수
     hkib40075-cat03-file4.categories        파일4의 2003범주당 문서수
     hkib40075-cat03-file5.categories        파일5의 2003범주당 문서수
```

6.1. 2003분류체계

2003범주체계는 한국일보-40075 테스트 컬렉션 구축시에 충남대학교 이석훈 교수 연구실에서 만들어진 것으로 대분류, 중분류, 소분류의 3단계 범주체계를 가집니다. HKIB-40075/hkib40075-cat03-all.categories 파일에 모든 소분류 범주당 문서수가 수록되어 있습니다. 이 파일에는 2003분류체계에서 사용되는 모든 범주가 포함되어 있으니 참조하십시오. 아래는 2003분류체계의 각 소분류 범주에 할당된 문서수를 일부 보여주고 있습니다. 전체 목록은 해당 파일을 참조하십시오.

```
한국일보-40075 실험문서집합의 2003범주당 문서수 (일부)
----------+----------------------------------------
   범주당 |
   문서수 |  /대분류/중분류/소분류
----------+----------------------------------------
     61   |  /건강과 의학/건강/영양 식품 식사
     35   |  /건강과 의학/건강/체력단련
     41   |  /건강과 의학/의약학/성인병
     12   |  /건강과 의학/의약학/수의학
     60   |  /건강과 의학/의약학/질병(암)
    228   |  /건강과 의학/의약학/질병(암외의질병)
```

```
    46   |  /건강과 의학/의약학/치의학
    40   |  /건강과 의학/의약학/한의학 전통의학
   144   |  /경제/가계 물가/가계 물가
   227   |  /경제/국가/수입
   548   |  /경제/국가/수출
   348   |  /경제/국가/재정 경기전망
    95   |  /경제/금융/보험(생명)
   101   |  /경제/금융/보험(손해)
     ...
----------+----------------------------------------
```

2003분류체계의 첫번째 특징은 모든 문서에는 소분류 단위에서 범주가 부여되어 있다는 것입니다. 위의 예에서 볼 수 있듯이 모든 범주는 3단계로 표현됩니다. 한국일보-40075 실험문서집합에서 대분류의 수는 9개, 중분류는 32개, 소분류의 수는 120개입니다. 두번째 특징은 모든 문서가 단일 범주를 가진다는 것입니다. 한국일보-40075 실험문서집합은 40,075개의 문서에 1:1로 범주가 부여되어 있어서 총 부여된 범주의 수는 문서의 수와 동일한 40,075개 입니다.

6.2. 2007분류체계

2007분류체계는 2003분류체계와는 두가지 측면에서 다릅니다. 첫번째 차이점은 2007분류체계의 경우에는 대분류와 중분류도 소분류와 마찬가지로 말단 범주의 역할을 한다는 것 입니다. 즉 대분류나 중분류 단계의 범주로 할당된 문서들이 다수 존재합니다. 두번째 특징은 각 문서가 복수의 범주를 가질 수 있다는 것입니다. 한국일보-20000 집합의 2007범주는 20,000개의 문서에 총 23,434개의 범주가 부여되어 있습니다. 이는 문서당 평균 1.17개의 범주가 부여되었음을 의미합니다.

```
한국일보-20000 실험문서집합의 2007범주당 문서수 (일부)
----------+----------------------------------------
  범주당 |
  문서수 |  /대분류[/중분류[/소분류]]
----------+----------------------------------------
   160   |  /
    25   |  /건강과 의학
     7   |  /건강과 의학/성인병
    10   |  /건강과 의학/수의학
    68   |  /건강과 의학/영양 식품 식사
    75   |  /건강과 의학/의약품
    32   |  /건강과 의학/질병/암
   168   |  /건강과 의학/질병/암외의질병
     3   |  /건강과 의학/체력단련
     5   |  /건강과 의학/치의학
    19   |  /건강과 의학/한의학 전통의학
    57   |  /경제
   282   |  /경제/가계 물가
   313   |  /경제/국가/국제
   255   |  /경제/국가/한국
     ...
----------+----------------------------------------
```

2007분류체계로 구축된 한국일보-20000 실험문서집합에는 미분류 데이타가 존재합니다. 9개 대분류 어디에도 속하지 않는 것으로 판정되는 160개의 문서가 "/"로 태깅이 되어 있습니다. 이 문서들은 학습집합에는 수용을 하되 성능평가에서는 제외할 것을 권고합니다.

또한 2007분류체계에서는 대분류 9개, 중분류 90개, 소분류 61개가 독자적인 문서를 가지는 말단 범주로서의 역할을 합니다. 이는 소분류만이 말단 범주로서의 역할을 하는 2003분류체계와의 가장 두드러진 특징입니다.

2003분류체계에서는 모든 말단 범주가 소분류이기 때문에 대분류, 중분류, 소분류에 따라 성능을 평가하는 것이 명확합니다. 그러나 2007분류체계에서는 대,중,소분류 모두가 말단 범주로서의 역할을 하기 때문에 대분류, 중분류, 소분류의 구분에 의한 성능평가 기준이 모호할 수 있습니다. 따라서 성능평가 시 다음 예와 같이 대,중,소분류를 구분할 것을 권장합니다.

한국일보-20000 집합을 2007분류체계로 성능평가할 경우의 대,중,소분류 분별 예

범주 예제	대분류	중분류	소분류
/건강과 의학	/건강과 의학	/건강과 의학	/건강과 의학
/건강과.../성인병	/건강과 의학	/건강과 의학/성인병	/건강과 의학/성인병
/건강과.../질병/암	/건강과 의학	/건강과 의학/질병	/건강과 의학/질병/암
/경제	/경제	/경제	/경제
/경제/가계 물가	/경제	/경제/가계 물가	/경제/가계 물가
/경제/국가/한국	/경제	/경제/국가	/경제/국가/한국

6.3. 실험문서집합/분류체계별 문서-범주 통계

한국일보-20000과 한국일보-40075의 각 실험문서집합은 대중소 분류별로 다음과 같은 통계를 가집니다.

[표3] 한국일보-20000/2007분류체계의 대/소분류별 통계

대분류	문서수	소분류 범주수*	소분류 범주별 평균문서수
/건강과 의학	344	10	34.4
/경제	6725	34	197.8
/과학	554	10	55.4
/교육	589	5	117.8
/문화와 종교	3144	26	120.9
/사회	4179	18	230.5
/산업	3162	21	150.6
/여가생활	271	4	67.6
/정치	872	32	27.3
/	160	-	-
계	20000	160	124.0**

(참고1) 분류 "/"는 9개 대분류에 포함되기 어려운 문서를 의미함.
따라서 이 범주에 속하는 미분류로 간주되며, 성능평가시 시험집합에서는 제외할 수 있음.
(*) 대분류 9개를 제외한 151개 소분류 범주중 90개는 중분류로서의 말단노드이며, 나머지 61개는 소분류로서의 말단노드.
(**) 미분류 "/" 범주를 제외한 평균값.

[표4] 한국일보-20000/2003분류체계의 대/중/소분류별 통계

대분류	문서수	중분류 범주수	중분류 범주별 평균문서수	소분류 범주수	소분류 범주별 평균문서수
/건강과 의학	449	2	224.5	8	56.1
/경제	6258	5	1251.6	19	329.4
/과학	680	2	340.0	8	85.0
/교육	583	4	145.8	4	145.8
/문화와 종교	2963	4	740.8	20	148.2
/사회	4520	2	2260.0	8	565.0
/산업	4191	5	838.2	18	232.8
/여가생활	356	2	178.0	4	89.0
계	20000	26	769.2	89	224.7

한국일보-20000 실험문서집합의 2003분류체계에서는 "/정치" 대분류에 속하는 범주를 부여받은 문서가 존재하지 않아서, 이 실험문서집합내에서는 8개의 대분류만이 나타납니다.

[표5] 한국일보-40075/2003분류체계의 대/중/소분류별 통계

대분류	문서수	중분류 범주수	중분류 범주별 평균문서수	소분류 범주수	소분류 범주별 평균문서수
/건강과 의학	523	2	261.5	8	65.4
/경제	7300	5	1460.0	19	384.2
/과학	794	2	397.0	8	99.3
/교육	680	4	170.0	4	170.0
/문화와 종교	3457	4	864.3	20	172.9
/사회	5273	2	2636.5	8	659.1
/산업	4890	5	978.0	18	271.7
/여가생활	614	2	307.0	5	122.8
/정치	16544	6	2757.3	30	551.5
계	40075	32	1252.3	120	334.0

한국일보-40075 실험문서집합의 경우 "/정치" 대분류에 속하는 범주를 가진 문서가 전체 문서집합의 41.3%에 달할 정도로 매우 많다는 것이 특징적입니다.

7. 문서범주화 실험

7.1. 시험집합과 학습집합

일반적으로 자동문서분류 실험에서는 실험문서집합을 학습집합(training set)과 시험집합(test set)의 두 개의 부분집합으로 나누어서 사용합니다. 문서분류기의 학습에는 학습집합만을 사용하고, 학습이 끝난 후에는 시험집합을 사용하여 문서분류기의 성능을 평가하게 됩니다. 시험집합의 각 문서당 문서분류기가 제시하는 범주와 시험집합의 해당 문서에 대해 사람이 할당한 범주를 비교함으로써

문서분류기의 성능을 평가하게 됩니다.

한국일보-20000과 한국일보-40075 실험문서집합의 경우 각각 5가지의 조합으로 시험집합과 학습집합을 구성하여 5겹 교차검증(5-fold cross validation)을 수행할 수 있도록 하였습니다. 즉, 5개의 데이터 파일 중 하나를 시험집합으로 사용하고 나머지 4개 파일을 학습집합으로 사용하는 것을 5번 반복할 수 있습니다. 표6과 표7에 한국일보-20000과 한국일보-40075 실험문서집합의 시험집합/학습집합 분할의 5가지 경우를 각각 실었습니다.

[표6] 한국일보-20000 실험문서집합의 시험/학습집합 구분

구분	시험집합	학습집합
Set1	HKIB-20000_001.txt	HKIB-20000_00[2-5].txt
Set2	HKIB-20000_002.txt	HKIB-20000_001.txt HKIB-20000_00[3-5].txt
Set3	HKIB-20000_002.txt	HKIB-20000_00[1-2].txt HKIB-20000_00[4-5].txt
Set4	HKIB-20000_002.txt	HKIB-20000_00[1-3].txt HKIB-20000_005.txt
Set5	HKIB-20000_001.txt	HKIB-20000_00[1-4].txt

[표7] 한국일보-40075 실험문서집합의 시험/학습집합 구분

구분	시험집합	학습집합
Set1	HKIB-40075_001.txt	HKIB-40075_00[2-5].txt
Set2	HKIB-40075_002.txt	HKIB-40075_001.txt HKIB-40075_00[3-5].txt
Set3	HKIB-40075_002.txt	HKIB-40075_00[1-2].txt HKIB-40075_00[4-5].txt
Set4	HKIB-40075_002.txt	HKIB-40075_00[1-3].txt HKIB-40075_005.txt
Set5	HKIB-40075_001.txt	HKIB-40075_00[1-4].txt

한국일보-20000 실험문서집합의 경우 한국일보-40075 실험문서집합에서 할당된 범주가 "#CAT'03 : "(2003범주)에 수록되어 있으므로 문서범주화 실험을 2가지 분류체계에 대해서 추가로 실행할 수 있습니다. 따라서 한국일보 실험문서집합 2개를 모두 이용한 실험에서는 다음과 같이 15가지 경우에 대한 성능평가를 수행할 수 있습니다.

시험/학습집합 분할 및 분류체계별 성능평가 경우의 수

순서	시험/학습 분할	시험대상 분류체계
1	한국일보-20000 Set 1	2007분류체계
2	한국일보-20000 Set 2	2007분류체계
3	한국일보-20000 Set 3	2007분류체계
4	한국일보-20000 Set 4	2007분류체계
5	한국일보-20000 Set 5	2007분류체계
6	한국일보-20000 Set 1	2003분류체계
7	한국일보-20000 Set 2	2003분류체계
8	한국일보-20000 Set 3	2003분류체계
9	한국일보-20000 Set 4	2003분류체계
10	한국일보-20000 Set 5	2003분류체계
11	한국일보-40075 Set 1	2003분류체계
12	한국일보-40075 Set 2	2003분류체계

13	한국일보-40075 Set 3	2003분류체계
14	한국일보-40075 Set 4	2003분류체계
15	한국일보-40075 Set 5	2003분류체계

7.2. 계층형 분류체계를 활용한 성능평가

6장에서 설명한 바와 같이 한국일보 실험문서집합은 대분류, 중분류, 소분류의 3단계 계층형 분류체계를 가지고 있습니다. 따라서 성능평가 결과는 대분류, 중분류, 소분류에 대해서 각각의 성능 평가 결과를 제시할 수 있습니다. 그러므로 한국일보 실험문서집합은 7.1절에서 보여준 15가지 시험집합/학습집합 분할에 대해서 대중소 3가지 분류에 따라 45개의 성능평가 결과를 도출할 수 있습니다.

시험/학습집합 분할 및 분류체계, 대중소 범주별 성능평가 경우의 수

순서	시험/학습 분할	시험대상 분류체계	대상 범주 단계
1~ 3	한국일보-20000 Set 1	2007분류체계	대, 중, 소분류
4~ 6	한국일보-20000 Set 2	2007분류체계	대, 중, 소분류
7~ 9	한국일보-20000 Set 3	2007분류체계	대, 중, 소분류
10~12	한국일보-20000 Set 4	2007분류체계	대, 중, 소분류
13~15	한국일보-20000 Set 5	2007분류체계	대, 중, 소분류
16~18	한국일보-20000 Set 1	2003분류체계	대, 중, 소분류
19~21	한국일보-20000 Set 2	2003분류체계	대, 중, 소분류
22~24	한국일보-20000 Set 3	2003분류체계	대, 중, 소분류
25~27	한국일보-20000 Set 4	2003분류체계	대, 중, 소분류
28~30	한국일보-20000 Set 5	2003분류체계	대, 중, 소분류
31~33	한국일보-40075 Set 1	2003분류체계	대, 중, 소분류
34~36	한국일보-40075 Set 2	2003분류체계	대, 중, 소분류
37~39	한국일보-40075 Set 3	2003분류체계	대, 중, 소분류
40~42	한국일보-40075 Set 4	2003분류체계	대, 중, 소분류
43~45	한국일보-40075 Set 5	2003분류체계	대, 중, 소분류

문서범주화 실험은 대부분의 경우 정확율(precison)과 재현율(recall)의 조화평균인 F1 척도(F1 score)로 제시됩니다. 그런데 F1 척도를 계산하는 데 있어서 문서중심의 미시평균 F1(micro-averaged F1)과 범주 중심의 거시평균 F1(macro-averaged F1) 계산 방식이 존재합니다. 따라서 45개의 성능평가에 미시평균법과 거시평균법의 두 가지 방법이 더 부여됨으로써 최종적으로는 90개의 F1값을 실험결과로 제시할 수 있습니다. 7.3절에서는 KRISTAL-IRMS의 자동문서분류기를 사용하여 도출한 90개의 성능평가 결과를 예로써 제시합니다.

7.3. KRISTAL-IRMS의 kNN 분류기를 이용한 성능평가 결과

한국과학기술정보연구원이 개발한 정보검색관리시스템인 KRISTAL-IRMS에는 기본 기능으로 문서분류기가 포함되어 있습니다(http://www.kristalinfo.com 참조). 본 절은 한국일보-40075 실험문서집합과 한국일보-20000 실험문서집합을 대상으로 KRISTAL에 포함된 kNN 문서분류기의 성능을 평가한 것입니다.

모든 결과값은 F1 값 중에서 정확도(precision)와 재현율(recall)이 동일한 값을 가지는 정확도-재현율의 분기점인 BEP(break-even point)를 제시하려고 노력하였습니다. 미시평균 F1(micro-averaged F1)의 경우는 모든 결과값에서 BEP를 제시하였으며, 거시평균 F1(macro-averaged F1)의 경우에는 정확도와 재현율이 99.9% 이상 동일한 값으로 부터 계산된 것으로 BEP에 근접한 값입니다.

[표 8] 한국일보-20000의 2007범주를 대상으로 한 KRISTAL kNN 분류기 성능

구분	Set1	Set2	Set3	Set4	Set5
대분류 miF1	0.788566	0.797560	0.788019	0.795152	0.812399
대분류 maF1	0.717253	0.716409	0.718674	0.742369	0.745079
중분류 miF1	0.692622	0.701345	0.691800	0.707538	0.721281
중분류 maF1	0.474352	0.490875	0.475111	0.478200	0.544887
소분류 miF1	0.637102	0.652837	0.643187	0.651534	0.677645
소분류 maF1	0.469789	0.474778	0.470635	0.484926	0.561387

* 성능평가시 시험집합 내의 "/" 범주는 제외 (6.2절 참조)

[표 9] 한국일보-20000의 2003범주를 대상으로 한 KRISTAL kNN 분류기 성능

구분	Set1	Set2	Set3	Set4	Set5
대분류 miF1	0.803548	0.826373	0.807751	0.813431	0.834471
대분류 maF1	0.653433	0.750162	0.737200	0.742266	0.758183
중분류 miF1	0.758830	0.776563	0.757737	0.759989	0.785438
중분류 maF1	0.609809	0.610755	0.610588	0.604374	0.642539
소분류 miF1	0.641710	0.659873	0.642504	0.641838	0.691028
소분류 maF1	0.519623	0.519659	0.531328	0.528987	0.586023

[표 10] 한국일보-40075의 2003범주를 대상으로 한 KRISTAL kNN 분류기 성능

구분	Set1	Set2	Set3	Set4	Set5
대분류 miF1	0.794306	0.791581	0.795543	0.804657	0.793743
대분류 maF1	0.650182	0.671559	0.667802	0.678257	0.664961
중분류 miF1	0.732087	0.732446	0.730241	0.738674	0.730022
중분류 maF1	0.540164	0.550234	0.544932	0.556471	0.563859
소분류 miF1	0.619404	0.625212	0.614179	0.629913	0.618881
소분류 maF1	0.495485	0.493899	0.489127	0.515153	0.511786

표 8, 9, 10에서 miF1은 미시평균 F1(micro-averaged F1) 척도, maF1은 거시평균 F1(macro-averaged F1) 척도를 의미합니다. 모든 실험 결과값은 KRISTAL 문서분류기의 *k*값을 10으로 설정하고, 자질선택(feature selection)은 문서빈도(DF) 2이상 800(한국일보-20000의 경우) 또는 1600(한국일보-40075)이하로 한 환경에서 얻은 것입니다. 여기서 800은 한국일보-20000의 각 학습집합의 약 5.0%, 1600은 한국일보-40075의 각 학습집합의 약 5.0%에 해당합니다. KRISTAL의 kNN 문서분류기는 자질을 선택기준을 색인어가 출현하는 문서의 빈도(DF; document frequency)로 설정합니다.

8. 실험집합내 문서의 중복에 대한 고찰

한국일보 실험문서집합에는 2회 이상 출현하는 신문기사들이 발견됩니다. 이러한 현상은 자동문서분류 실험집합을 구축하는 과정에서 다수의 작업자가 동시에 범주 부여 작업을 수행하는 중에 발생하는 절차상 오류에 기인한 것으로 보입니다. 한국일보-40075 실험문서집합을 DocID 별로 통계를 내보면, 고유한 DocID는 모두 35,543 건이었습니다. 한국일보-20000 실험문서집합에서는 고유한 문서식별자의 수가 18,094개였습니다. 아래 표11에 각 실험문서집합별로 중복된 문서ID 통계를 요약하여 실었습니다.

표 11에서 보듯이 한국일보-40075 실험문서집합에는 3697개의 문서가 1회이상 중복되어 수록되어 있으며, 한국일보-20000에서는 1708개의 문서가 1회 이상 중복되고 있다. 이는 전체 문서집합에서 한국일보-40075 실험문서집합은 11.3%, 한국일보-20000 실험문서집합은 9.5%의 문서가 중복되었다는 것을 의미합니다.

[표11] 실험문서집합내 문서식별자(DocID) 중복 통계

출현횟수	한국일보-40075	한국일보-20000
1회	31,846	16,386
소계	31,846	16,386
2회	3,037	1,530
3회	530	160
4회	99	16
5회	19	2
6회	10	0
7회	2	0
소계	3,697	1,708
계	35,543	18,094

중복된 문서 1개의 예를 들어보겠습니다. 문서ID 30736 문서의 기사제목은 "한국소비자보호원, '잘못알고있는 식생활상식 19가지'펴내"이며 민간요법이나 음식에 관한 건강상식을 다루는 출판물을 다루고 있습니다. 아래에는 HKIB-20000/HKIB-20000_001.txt에 수록된 위 문서의 일부를 보여주고 있습니다.

```
@DOCUMENT
#DocID : 30736
#CAT'03: /산업/농축산수산/쌀
#CAT'07: /문화와 종교/도서 출판;/문화와 종교/음식 요리
#TITLE : 한국소비자보호원, '잘못알고있는 식생활상식 19가지'펴내
#TEXT  :

 (서울=연합) 우리 주변에 통념으로 자리잡은 여러 식생활 상식중에는
임상효과가 증명된 민간요법도 있지만 검증되지 않은채 의학상식처럼
통용되는 경우도 흔히있다.

 한국소비자보호원이 발행하는 소비자시대 1월호는 "잘못 알고 있는
```

식생활 상식19가지"를 싣고,일반적인 식생활 상식의 허와 실을 규명
했다.

"채식주의자가 장수한다"는 상식의 경우 식물성 식품에는 비타민
B12가 없어 악성 빈혈이나 뇌장애를 가져올 수 있다는 것도 염두해야
하며 "우유에는 소금을 타먹는 것이 좋다"는 말을 따를 경우 우유에
이미 적당량의 염분이 들어 있어 염분을 과다하게 섭취하게 될 우려가
있다. (이하 본문 생략)

한국일보-20000 실험문서집합에서 문서ID 30736은 5회 반복하여 출현하며 각각 아래와 같이 범주가 부여되어 있습니다.

한국일보-20000에서 문서 30736의 중복출현 위치와 부여된 범주

위치	2003범주	2007범주
파일1	/산업/농축산수산/쌀	/문화와 종교/도서 출판 /문화와 종교/음식 요리
파일3	/건강과 의학/의약학/성인병	/건강과 의학/영양 식품 식사
파일3	/산업/음식료품/곡물 가공 면류	/사회/시설단체 /문화와 종교/도서 출판
파일5	/건강과 의학/건강/영양 식품 식사	/건강과 의학/영양 식품 식사
파일5	/과학/자연과학/생물	/건강과 의학/영양 식품 식사

위의 예에서 보듯이 중복된 문서라고 하더라도 작업자에 따라 다른 범주들이 부여된 것으로 보입니다. 따라서 단순하게 중복문서를 제거하기는 어려운 것으로 판단되어 현 배포판에서는 중복문서를 제거하지 않은 상태를 유지하였습니다. 중복문서가 문서범주화의 전체적인 성능에 미치는 영향에 관한 연구와 정제된 차기 버전의 실험문서집합 구축은 향후 과제로 남겨져 있습니다.

9. 감사의 글

연구용으로 2년간의 신문기사를 흔쾌히 제공해주신 한국일보 신문사에 감사드립니다. 한국일보 실험문서집합 구축사업에 참여해주신 충남대학교 이석훈 교수님, 연세대학교 나동열 교수님과 고단한 수작업을 감내해주신 범주부여 작업자들에게 깊은 감사를 드립니다.

문서작성: 김진숙 (jinsuk@kisti.re.kr) - 오류 및 수정사항은 메일로 알려주시면 감사하겠습니다.