한국어 문서분류 테스트컬렉션 개발

Developing a Test Collection for Korean Text Categorization

나동열, 김윤식, 신현주, 이규희, 김태규, 강현규*,
최호섭**, 윤화묵**
연세대학교, 건국대학교*, 한국과학기술정보연구원**

Dong-Yul Ra, Yunsik Kim, Hyun-Joo Shin,
Kyu-Hee Lee, Tae-Kyu Kim, Hyun-Kyu Kang*,
Ho-Seop Choe**, Hwa-Mook Yoon**
Yonsei Univ., KunKook Univ.*,
Korea Research Institute for Science Technology Information**

요약

문서분류 시스템은 수많은 문서들이 쏟아져 나오는 최근의 인터넷 사회에서 매우 중요한 도구이다. 이러한 이유로 문서분류 기술에 대하여 많은 연구가 있어 왔다. 문서분류 시스템의 개발을 위해서는 보통 교사학습 기법이 이용되는데 이를 위해서 필수적인 것이 테스트컬렉션이다. 영어의 경우에는 여러 가지의 문서분류 테스트 컬렉션이 있어 이 분야의 기술발전에 많은 도움을 주고 있다. 그러나 한국어의 경우에는 공식적으로 공표된 문서분류 테스트컬렉션이 존재하지 않고 있다. 이러한 상황을 개선하기 위해서 우리는 문서분류 테스트컬렉션의 구축을 진행하고 있다. 본 논문에서는 이에 대한 접근 방법 및 구축 상황을 기술하고자 한다.

Abstract

Document categorization system is important in the internet age in which huge number of documents are created and need to be dealt with. By this reason a lot of research has been done in this field. For the development of the system, a supervised learning method is widely used. This approach needs a test collection as a prerequisite. For the case of English, several test collections are available which provide a lot of help for developing systems and doing research. But no public test collections have been reported and are not available in the case of Korean. To improve the situation for Korean we are undergoing the construction of a Korean test collection. In this paper the approaches being used and current stage of the collection will be described.

1. 서 론

수많은 문서가 생성되는 현대 사회에서 모든 문서를 다 읽어 보는 것은 매우 어렵다. 따라서 문서들을 미리 기계가 여러 분야로 분류하여 놓는다면 사람은 자신이 관심이 있는 분야(범주; category)로 분류된 문서만을 읽어 볼 수 있게 되고 이는 많은 시간과 노력을 절감할 수 있도록 한다. 기계에 의한 문서분류 (text categorization)를 위해서는 분야 즉 범주 집합이 미리 주어 져야 한다. 기계는 각 문서를 검토하여 이 문서와 적합한 즉 관련이 있는 범주(들)를 선택하여 범주 레이블을 문서에 붙여 놓는다.

이러한 문서분류 시스템을 개발하기 위해서는 교사학습 (supervised learning) 기반 기계학습 기법을 사용하는 것이 주된 추세이다[3]. 이를 위해서는 상당량의 문서에 범주 레이블을 미리 붙여 놓은 테스트컬렉션이 있어야 한다. 테스트컬렉션의 구축은 많은 시간과 노력이 들어가는 작업으로서 개개의 연구집단이 수행하기에는 너무 부담이 크다. 이러한 이유로 선진국에서는 국가기관이나 대형 연구기관에서 이를 구축하여 연구자들로 하여금 이용할 수 있게 하고 있다[1, 2].

그러나 국내의 경우 아직까지 공식적으로 발표된 테스트컬렉션이 존재하지 않고 있다. 한국과학기술연구원에서 구축한 기본적인 테스트컬렉션이 있으나 품질 개선의 필요성이 제기되어 왔다(이 컬렉션을 KRTC 라 부르자)[4]. 본 그룹에서는 이 기본 테스트컬렉션을 기반으로 하여 보다 만족스런 품질을 가짐으로써, 연구에 많은 도움이 될 수 있는, 문서분류 테스트컬렉션의 구축 작업을 진행하고 있다. 본 논문에서는 우리의 이 구축 작업에 관하여 살펴보고자 한다.

2. 분류체계

2.1 분류체계 원리

문서 분류를 위한 "범주집합" 즉 "문서 분류체계"의 결정은 매우 어려운 작업이다. 최근에는 계층구조를 가진 분류체계를 이용하는 늘어나는 추세이다. 분류체계를 먼저 결정하고 이에 맞게 문서를 수집하는 것이 가장 좋은 방법이지만 수반되는 어려움이 많다. 분류체계의 결정에 관계되는 한 원리는 분류트리 균형의 원리이다. 이는 분류체계를 나타내는 트리가 가능하면 균형트리이어야 한다는 것이다. 트리가 균형적이 되려면 리프노드들의 깊이(depth)에 큰 차이를 보이지 않아야 한다. 예를 들면 리프 노드들 중 최대의 깊이를 가진 것과 최소의 깊이를 가진 것 사이의 차이를 이용할 수 있다.

$$D = D_{max} - D_{min}$$

여기서

D_{max} = 깊이가 최대인 리프노드의 깊이

D_{min} = 깊이가 최소의 리프노드의 깊이

D를 일정 범위 이하로 제한함으로써 균형의 정도를 조정할 수 있다. 예를 들면 D = 0 이면 완전균형을 유지하는 트리로서 예를 들면 B-tree 가 있다. D = 1 이면 상당이 균형적인 트리로 간주되는 데 예를 들면 AVL 트리가 있다. 분류체계가 가능하면 균형적이어야 된다는 것을 분류체계 균형의 원리라 부른다.

분류체계 트리 균형의 원리
분류체계를 나타내는 트리가 가능하면 균형 트리가 되는 것을 선호한다.

이때의 주요한 점은 분류체계에 맞게 문서를 수집하는 일인데, 어떤 범주는 소속 문서수가 매우 많은 어떤 범주는 매우 작은 현상은 바람직한 것이 아니다. 하지만 기사가 많이 생성되는 범주가 그렇지 않은 범주보다는 많은 수를 가지는 것은 당연하다. 그러나 특정범주에 너무 작은 수의 문서가 소속되는 것은 바람직 하지 않다. 모든 범주가 훈련에 적당한 수준의 문서수는 되어야 한다는 것으로 다음 원리가 이를 나타낸다.

분류체계/문서집단 균형의 원리
분류체계의 각 범주에 적당한 수의 문서가 소속되도록 한다.

2.2 분류체계 구축 방법

분류체계의 결정을 위해서는 3 가지 접근법을 생각해 볼 수 있다. 하나는 분류체계를 먼저 결정하고 이것에 맞추어 문서를 수집하여 문서집단을 구축하는 것이다(이를 **C-d** 기법이라 부르자.) 다른 하나는 문서집단을 먼저 구축하고 이 문서 집단에 맞도록 분류체계를 만드는 것이다 (이를 **D-c** 기법이라 부르자.) 세 번째는 이 두 가지를 혼합하여 적절히 양쪽을 왔다 갔다 하면서 구축하는 기법이다 (이를 **M-c-d** 라 부르자.)

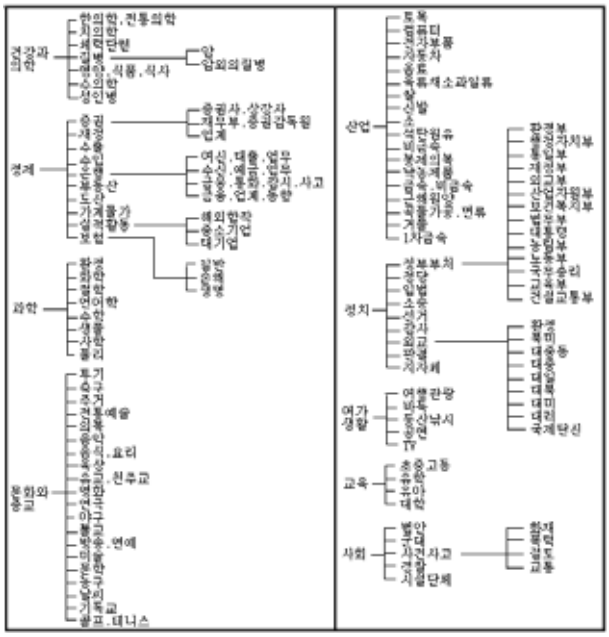
이상적인 기법은 **C-d** 로서 이를 실현할 수 있으면 좋지만 어려운 점은 분류체계의 각 범주에 속하는 문서를 찾아 수집하기가 쉽지 않다는 점이다. 예를 들면 "경제/부동산/단독주택" 이라는 범주가 있다 하자. 이 범주에 속하는 문서를 어느 정도는 수집하여야 하는데 이를 어디에서 어떻게 수집하여야 할지 쉽지 않다. 물론 웹의 포털사이트나 다른 곳에 이미 분류된 문서를 가져 올 수도 있지만 이것은 그리 좋은 방법이 아니다. 그리고 나중에 언급하겠지만 "정보원의 균형성"에 적합하지 않게 될 수 있는 문제의 소지가 있다.

두 번째인 **D-c** 기법에서는 문서집단에 알맞은 범주체계를 만들다 보면 균형트리 원리를 만족하는 분류체계를 구하기가 쉽지 않다는 점이 문제이다. 예를 들면 문서집단이 경제 중에서 증권에 대한 문서만을 가지고 있다고 하면 경제 범주 아래의 다른 분야 즉 보험, 은행 등에 대한 노드를 분류체계에 만들 수 없게 된다.

이러한 관찰을 통해 세 번째인 **M-c-d** 기법이 일반적으로 추천될 만하다. 여기서는 일단 어느 정도 분류체계를 만든 후 이에 대한 문서를 수집한다. 문서 수집 과정에서 어려운 범주는 범주체계에서 제외하고 다른 범주로 대체하는 것을 고려한다. 그러면 분류체계에 넣고자 하는 대안이 되는 새로운 범주가 무엇들이 가능할 가를 고려하고 그 중에서 가장 좋아 보이는 것을 선택한다. 그리고 이에 대한 문서 수집을 검토한다. 이 과정을 반복수행하면서 구축하는 기법이다.

우리는 KRTC 에 포함된 분류체계에서부터 출발하였다. 주어진 KRTC 문서집단에 대한 고려 없이 분류체계를 완전히 새로 시작하면 주어진 문서집단의 많은 문서들을 이용하지 못하게 되는 문제점이 있으며 문서 수집을 완전히 새로 시작해야 하는 문제가 있다. 그렇다면 문서집단을 이용하기 위해 주어진 문서집단을 분석하여 이에 맞는 분류체계를 만드는 것이다(즉 위의 **D-c** 기법을 따르는 것이다.) 그러나 주어진 문서집단의 40,000 개 이상의 문서를 모두 검토하는 것은 매우 시간이 많이 필요한 작업이다. 주어진 우리의 작업 여건에서는 1 년 이상 소요되는 작업이다. 따라서 우리는 KRTC 분류체계를 기반으로 출발하였다(그림 1).

KRTC 분류체계는 KRTC 문서집단에 대한 분류를 위하여 만들어진 것으로서 상당부분 문서집단과 매칭이 된다고 할 수 있다. 우리의 목표는 이 두 데이터를 매칭하여 가면서 분류체계를 수정 보완하여 나가기로 하였다. 이러한 매칭 작업이 완료되면 분류체계와 문서집단에서 부족한 점을 보완하는 2 단계 작업을 거침으로써 보다 좋은 테스트컬렉션을 얻을 수 있다고 생각한다.



▶▶ 그림 1. KRTC 분류체계

기본 분류체계의 모습은 그림1 과 같이 3 단계의 계층구조를 가지고 있다. 루트 바로 아래 레벨을 대분류, 그 아래 레벨을 중분류, 그 아래 레벨을 소분류라고 부르자. 대분류 레벨은 9 개의 노드를 가지고 있다: 사회, 산업, 정치, 여가생활, 건강과의학, 경제, 과학, 교육, 문화와종교. 각 대분류는 중분류로 나누어지는데 대분류에 따라서 나누어지는 중분류의 수에 대한 편차가 상당히 큰 것을 볼 수 있다. 이것은 대분류 자체가 얼마나 넓은 범위의 영역인가에 따라 달라진다고 볼 수 있다. 위에 주어진 분류체계를 바탕으로 새로운 분류체계를 도출하는 것이 목표이다.

3. 범주 태깅

3.1 KRTC 문서집단

우리는 문서집단을 직접 수집하는 대신 KRTC 의 문서집단을 이용하였다[4]. 그 이유는 완전히 새로 시작하는 것보다는 과거의 기본 데이터를 기반으로 출발함으로써 시간과 노력을 절감시킬 수 있고 과거 투자에 대한 낭비를 줄일 수 있기 때문이다. 우리가 작업을 할 문서집단의 특징은 표 1과 같다. 문서집단을 구성하는 문서는 대부분 1990년대 중반의 통신사 기사들이다.

[표 1] KRTC 문서집단

전체 문서수	40,075	
데이터 량	66.91 MB	
문서의 크기 (어절)	최소	4
	최대	2,038
	평균	171

3.2 문서에 대한 범주 태깅

주어진 기본 분류체계와 문서 집단을 바탕으로 각 문서에 범주를 붙이고 (태깅하고), 필요한 경우 분류체계를 수정하는 작업을 진행한다. 주어진 문서의 구성은 다음과 같다.

@DOCUMENT
#FILE: 경제/증권(업계)/20531.txt
#CATEGORY: 경제/증권(업계)
#TITLE: 상장사유상 증자액 2조원돌파
#DocID: 20531
#CONTENT:
올들어 주식시장이 회복세를 보이면서 상장사의 유상증자액이 2조원을 넘어섰다.
20일 증권거래소에 따르면 지난 1월부터 5월말까지 유상증자를 실시한 상장법인수는 75개사이며 유상증자액은 2조1천2백94억원으로 집계됐다. 이는 지난해 같은기간의 51개사 7천7백84억원에 비해 금액상으로 3배 가까이 급증했다.
업종별로는 제조업이 35개사 6천1백4억원이었으며 금융업 21개사 1조1천6백69억원 건설업 14개사 2천6백18억원 도소매업 5개사 9백3억원 등이었다.

▶▶ 그림 2. 범주 태깅될 문서의 예

문서집단은 KRTC.2003 TrainingSet.kst 과 KRTC.2003 TestSet.kst 두 개의 화일에 들어 있으며 각 문서는 문서 시작 태그 @DOCUMENT 로 구분된다. 태그 #CATEGORY 에 원래 붙여 놓은 범주가 있다. 이 값은 항상 #FILE 에 주어진 정보와 동일하다는 사실이 관찰되었다. 따라서 #FILE 필드와 #CATEGORY 필드가 중복된 정보를 포함하고 있는 상태이다. 문서 분류 작업의 편리를 도모하고 에러의 발생을 줄이기 위해 우리는 그림 3과 같은 태깅 도구를 사용한다.

▶▶ 그림 3. 범주 태깅 도구

문서 분류작업의 작업 흐름은 그림 4와 같다.

분류 태깅될 문서 하나에 대하여 작업자는 먼저 TITLE 필드를 검토한다. TITLE 필드는 문서의 내용을 드러내는 가장 중요한 부분으로서 범주 결정에 큰 영향을 준다. 타이틀의 의미가 명확한 경우 CONTENT 의 첫 한 두 문단을 보고 타이

틀에서 제시하는 내용임을 확인한다. 만약 이 확인이 긍정적이지 않은 경우에는 문서의 CONTENT 를 더 읽어 보고 문서가 전달하고자 하는 의미를 정확히 파악한다.

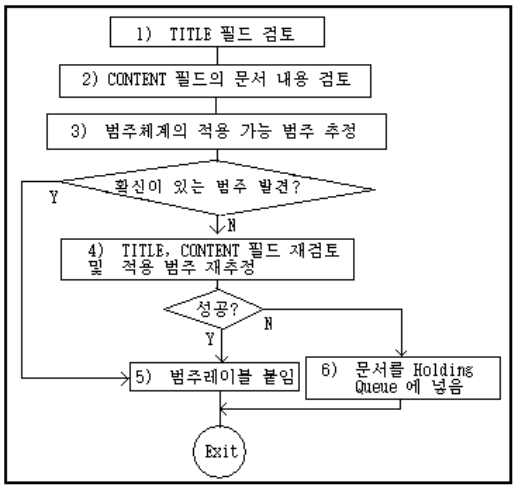
이렇게 하여 문서의 내용이 결정되면 분류체계를 검토하여 적용 가능한 범주들을 선정한다. 이때 선정되는 범주는 다음 원리에서 말한 대로 가능하면 가장 구체적인 범주가 되도록 한다.

최대 깊이의 원리
문서의 범주는 가능하면 깊은 레벨의 노드를 선택한다.

만약 루트로 부터의 경로가 다른 두 범주 이상이 적용 가능하다면 이러한 모든 범주를 선정한다. 즉 다중 범주를 허용한다.

다중 범주의 원리
문서의 내용에 부합되는 모든 범주를 선정한다.

적용에 대한 확신성이 있는 범주가 하나도 발견되지 못할 수 있다. 이 경우는 여러 가지 원인으로 발생한다. 첫째는 문서의 내용에 맞는 범주가 범주체계에 없는 경우이다. 이런 경우에는 문서와 함께 이의 희망 범주를 기록하여 놓는다. 둘째는 여러 범주들이 적용 가능하고 이 중 어느 것을 고르는 것이 좋은지 결정이 애매한 경우이다. 이때 적용 가능한 모든 범주를 선정할 수 있으나 이렇게 하면 범주의 다발성 문제가 생긴다. 우리는 보통 2 개 보다 더 많이 범주를 붙이는 것을 가능한 한 자제시키고 있다. 이 두 가지 경우 중 어느 경우이든 보다 이 문서는 보다 면밀한 검토가 요구되는 문서이다. 이런 상황이 되면 다시 타이틀과 내용을 검토하여 문서 내용에 대하여 보다 숙지를 한 후 범주 선정을 시도한다.



▶▶ 그림 4. 범주 태깅 과정

이런 재 시도에서 성공적이면 선정된 범주 레이블들을 문서에 붙이고 종료하나 그렇지 못하면 이 문서를 임시저장 큐(holding queue) 에 넣는다. 이 큐에 쌓인 문서는 매 주마다 검토회의를 거쳐 결정을 시도한다. 검토회의에서는 희망범주들을 고려하여 새로운 범주를 추가할 지에 대한 사항도 결정한다. 이 과정에서 기존의 범주의 변경이나 삭제와 같은 가능성도 고려하여 분류체계의 향상을 추구한다.

3.3 분류체계의 변경

지금까지의 작업 과정에서 분류체계의 수정이 여러 번 있어왔다. 지면 관계상 변경된 분류체계 전체를 보여 주는 대신 진행된 분류체계 수정 사항을 다음과 같이 언급하는 것으로 대신한다.

1) “경제/증권” 의 세분류를 다음처럼 보다 세분화한다.

증권사, 상장 상장사, 코스닥, 투신사 채권, 재무부 증권감독원, 업계 현황

2) “경제/화폐 통화 외환” “경제/신용카드” “경제/보험/업계동향” 을 추가한다.
3) “경제/기업” 을 만들고 이의 세분류로 “국내”와 “해외”를 둔다. “경제/국가” 를 만들고 이 밑에 “한국” 과 “국제”를 만든다. “경제/세금”을 추가한다.
4) “사회/사건사고” 밑에 다음 세분류들을 추가한다: 성폭행 성희롱, 사기, 살인, 뇌물, 노사분규, 시위.
5) “사회/경찰“ 을 ”사회/경찰 검찰“ 로 변경하고 “사회/전쟁”, “사회/교통”을 추가한다.
6) “문화와 종교/스포츠” 를 두고 모든 스포츠 분야를 이것의 밑에 노드들로 만든다.
7) “문화와 종교/도서출판”, “문화와종교/언론” 을 추가한다.
8) “산업/정보통신”, “산업/해운”을 추가한다.
9) “과학/지구과학”을 추가한다.
10) “여가생활/공연”, “여가생활/TV” 를 없애고 이를 “문화와종교/방송 연예” 에서 담당토록 한다.

[표 2] 기존범주의 변경율

	문서수	변경문서수	변경율(%)
건강과의학	449	238	53.0
경제	5,257	3,535	67.2
사회	3,520	2,003	56.9
문화와종교	2,932	1,287	43.9
산업	1,548	657	42.4
과학	294	43	14.6
전체수	14,000	7763	55.5

3.4 작업 현황

현재까지 14,000 개의 문서에 대하여 태깅 작업을 완료하였다. 범주 선정이 어려워 임시작업 큐로 들어가는 문서의 비율은 약 5% 로서 대부분의 문서에 대하여 큰 어려움 없이 태깅이 가능한 것으로 밝혀졌다. 구축 작업의 당위성을 알아 보기 위해 KRTC 에 원래 붙여 있던 범주 태그가 다른 범주로 변경된 정도를 조사하였다. 범주의 변경은 두 가지 이유로 발생한다. 하나는 원래 태그가 오류이어서 다른 것으로 붙인 경우와 다른 하나는 태그 명의 분류체계의 변경으로 인한 경우이다. 표 2 의 데이터를 보면 새로 구축되는 범주레이블에 과거의 것과 많이 다른 것을 알 수 있다. 즉 과거의 데이타에 부족한 점이 많았음을 나타내고 있다.

4. 결 론

본 논문에서는 우리가 진행하고 있는 문서분류 연구를 위한 테스트컬렉션 구축 작업에 대하여 살펴보았다. 우리 작업은 KRTC 데이터에서 출발하여 보다 양질의 언어자원을 구축하는 것을 목표로 하고 있다. 현재까지의 작업 결과를 볼 때 KRTC 의 재구축작업의 필요성이 입증되었다. 그러나 지금까지의 진행에 대한 문제점으로 생각되는 것은 분류체계의 정비가 충분치 못한 것 같다. 이 문제는 보다 깊은 연구 및 검토가 필요한 것으로 생각된다. 다른 문제로는 다중 범주를 붙이는데 있어서 일관성을 갖기 어려운 점이 관찰되었다. 즉 어느 정도의 관련성이 있을 때 관련 범주로 보아야 할지에 대한 기준이 명확하지 못한 실정이다. 또 하나의 문제점은 태깅 작업에 대한 검증 절차가 미흡하다는 점이다. 앞으로 이러한 점을 고려하여 구축 작업을 개선할 계획으로 있다.

▌참 고 문 헌▐

[1] Lewis, D., "Reuters-21578 text categorization test collection README file", Manuscript, Sep. 1997, (available at www. daviddlewis. com / resources / testcollections)

[2] Lewis, D., Yang, Y., Rose, T. and Li, F., "RCV1: A New Benchmark Collection for Text Categorization Research", J of Machine Learning Research, Vol. 5, pp. 361-397, 2004.

[3] Sebastiani, F., "Machine learning in automated text categorization", ACM Computing Surveys, 34(1), pp.1-47, 2002.

[4] http://www.kristalinfo/K-Lab/, "KRTC.2003.tar.gz"