제목의 단어 가중치를 이용한 중등학교 공문서 자동분류시스템

강 현 희* · 진 민**

*경남대학교 교육대학원 전자계산교육전공

**경남대학교 정보통신공학부

요 약

현재 일선 학교와 교육기관의 공문서 분류는 아직도 수작업으로 처리되고 있어 많은 시간이 소요된다. 이러한 문제점을 해결하기 위해 본 논문은 문서 제목의 단어 정보를 이용한 자동 문서 분류 방법을 제안한다. 먼저 기존 문서의 제목 단어 중에서 의미 있는 단어를 추출하여 각 단어에 대해 범주별로 역문헌 빈도(IDF) 가중치를 계산한 후 단어 가중치 사전을 구축한다. 문서의 분류 요구가 들어오면 구축된 단어 가중치 사전을 이용하여 문서 제목에 포함된 단어들의 범주별 가중치 합을 비교하여, 범주별 가중치 합이 최대인 범주로 문서를 분류한다. 실제 중등학교에서의 공문서를 대상으로 제안된 방법의 분류 성능을 평가하였다.

An Automatic Classification System of Official Documents in Middle Schools Using Term Weighting of Titles

Hyun-hee Kang* · Min Jin**

*Computer Science Education, Graduate School of Education, Kyungnam University

**Div. of Information and Communication Engineering, Kyungnam University

ABSTRACT

It takes a lot of time to classify official documents in schools and educational institutions. In order to reduce the overhead, we propose an automatic document classification method using word information of the titles of documents in this paper. At first, meaningful words are extracted from titles of existing documents and Inverse Document Frequency(IDF) weights of words are calculated against each category. Then we build a word weight dictionary. Documents are automatically classified into the appropriate category of which the sum of weights of words of the title is the highest by using the word weight dictionary. We also evaluate the performance of the proposed method using a real dataset of a middle school.

본 연구는 경남대학교 학술논문게재연구비 지원에 의하여 이루어졌음

1. 서론

급속히 발전하는 통신망의 영향으로 업무상 처리해야 하는 온라인 전자 문서의 양도 따라서 늘어나고 있으며 문서의 양이 늘어날수록 처리 부서 담당자가 원하는 문서를 정확하고 신속하게 전달받기 위한 분류, 관리의 필요성이 더욱 커지게 된다. 하지만, 현재 일선학교에서의 교육 공문서 분류는, 분류담당자(교감)가 접수된 문서를 수작업으로 분류한 후에 각 문서를 부서별 처리담당자에게 전달하는 방식을 취하고 있다. 만약 전자문서관리시스템 내에서 접수된 문서를 자동으로 분류하여 준다면, 여러 단계를 거치는 문서 처리의 번거로움을 피할 수 있어 효율적인 문서 처리와 관리가 가능해질 것이다.

자동 문서분류란 문서의 내용에 근거해서 컴퓨터가 자동으로 미리 정의된 여러 개의 범주 중에서 가장 관련이 있는 범주에 문서를 할당하는 작업을 말하며, 문서 분류 시스템은 문서를 분류하는데 필요한 자질을 추출하는 작업과 추출된 자질을 기반으로 범주를 결정하는 문서 분류 작업으로 이루어진다[1].

문서의 범주를 결정하는 문서 분류 방법에는 베이지안 확률 모델(Naive Bayesian probability), 벡터 유사도 계산(Vector similarity), 결정트리(Decision tree), K-최근접이웃 기법(K-nearest neighbor), 규칙 기반(Rule based model), 신경망(Neural networks) 등의 방법이 있다[1, 6, 9, 10].

자질 추출 방법에는 단어 빈도(Term frequency), 카이제곱 통계량(χ^2 statistics), 상호정보측도(Mutual information), 기대 상호정보측도(Expected mutual information), 정보 획득량(Information gain) 등의 방법이 있다[1, 2].

일반적으로 문서의 제목은 문서의 내용을 대표한다고 볼 수 있다. 그런데 뉴스와 전자메일 등의 비형식적인 문서는 문서 제목이 문서 내용을 대표하기보다는 불필요하거나 오히려 모호성을 키우는 결과를 가져오는 경우도 있다. 그러나 교육공문서인 경우 대부분 형식을 갖추고 있을 뿐만 아니라 문서 제목이 문서 내용을 축약하여 표현하고 있는 정도가 다른 문서에 비해 강한 편이다. 본 논문은 이러한 점에 착안하여, 문서 제목에 나타난 단어의 출현 빈도를 기반으로 가중치를 계산하여 문서를 분류하는 방법을 제안하고 실제 문서를 이용한 실험을 통하여 이 방법의 정확도를 측정하였다.

2. 관련 연구

본 논문에서는 단어 가중치(Term weights) 기법을 사용하여 중등학교 공문서를 분류하는 시스템을 제안하고 있다. 그런데 저자가 조사한 바로는 교육공문서를 대상으로 한 자동 분류에 대한 연구는 알려져 있는 것이 없는 실정이다. 따라서 이 장에서는 본 논문에서 사용하고 있는 단어 가중치 기법에 관련된 내용을 살펴보기로 한다. 단어 가중치를 구성하는 요소는 단어 빈도(Term frequency: TF), 역문헌 빈도(Inverse document frequency: IDF), 문서 길이 정규화(Document length normalization)이다. 문서에서 단어의 중요도를 결정하는 두 가지 주 요소는 단어 빈도와 역문헌 빈도이다[11].

2.1 단어 빈도

단어 가중치 계산 방법에는 단순히 단어빈도를 그대로 사용하는 단순 단어빈도, 출현빈도가 1인 단어의 지나치게 낮은 영향력을 보충하고, 출현빈도가 높은 단어의 지나친 영향력을 낮추기 위한 로그 단어빈도, 이진 단어빈도, 보정 단어빈도, 오가피 단어빈도 등 다양한 공식이 제안되어 있다[4, 8, 11].

2.2 역문헌 빈도

단어빈도를 문헌빈도로 나누어 빈도값을 표준화시킨 상대빈도값으로, 보통 이 역문헌빈도에 용어빈도를 곱하여(IDF ×TF) 단어에 가중치를 부여한다.

단어 W_i의 문서 j에서의 빈도수를 $freq_{ij}$, 총 문서의 개수를 N, 단어 k_i가 있는 문서의 개수를 n_i라고 할 때, 특정 범주에서 TF-IDF에 의한 가중치는 다음과 같이 계산된다[2, 5, 7].

$$idf_i = \log \frac{N}{n_i}$$

$$W_{ij} = \frac{freq_{ij}}{\max_k freq_{kj}} \times idf_i = \frac{freq_{ij}}{\max_k freq_{kj}} \times (\log \frac{N}{n_i})$$

문서 d_j에서의 색인어 k_i의 가중치는 $W_{ij} \geq 0$이고, 문

서 내에 한번도 출현하지 않은 색인어의 가중치는 0이 된다.

단어의 중요도는 그 단어가 출현하는 전체 문헌수에 대해 반비례한다고 보고 출현빈도가 낮은 저빈도 단어일수록 높은 가중치를 주고, 출현빈도가 높은 고빈도 단어일수록 낮은 가중치를 주어 색인어를 선정한다.

2.3 역카테고리 빈도

역카테고리 빈도(Inverse Category Frequency: ICF)란 문서의 분류를 위해 범주의 분리 능력이 우수한 색인어에 높은 가중치를 주는 방법으로, 단어 W_i의 범주 j에서의 빈도수를 $freq_{ij}$, 총 범주의 개수를 M, 단어 W_i를 포함하는 범주의 개수를 m_i 라고 할 때 TF-ICF에 의한 가중치는 다음과 같이 계산된다[2, 9].

$$icf_i = \log \frac{M}{m_i} + 1$$

$$W_{ij} = freq_{ij} \times icf_i = freq_{ij} \times \left(\log \frac{M}{m_i} + 1 \right)$$

ICF는 IDF와 기본 원리는 같지만, IDF는 문서간의 분리도가 높은 단어에 높은 가중치를 주는 것이고, ICF는 카테고리간의 분리도가 높은 단어에 높은 가중치를 주는 점에 차이가 있다. 즉 소수의 카테고리에 많이 나온 단어에 대해 높은 가중치를 주고, 여러 카테고리에서 고르게 나오는 단어에 대해서는 낮은 가중치를 주는 것이다[5, 9].

2.4 문헌 길이 정규화

일반적으로 TF×IDF 가중치가 단어의 중요도를 결정하는 좋은 추정치이지만 적절하지 못한 측면이 있다. 이 값 자체는 단어의 중요성을 결정하는데 기여한 문서의 길이를 무시하고 있다. 실제로 분류하고자 하는 문서의 길이가 다양한데 이러한 성질은 다음의 이유로 긴 문장의 범주 유사성을 증대시키게 된다. 첫째, 긴 문서는 동일 단어를 반복적으로 사용하는 경향이 있다. 결과적으로 긴 문헌일수록 단어 빈도가 커지게 되어 범주 유사도를 높이게 된다. 둘째, 긴 문서는 많은 단어를 가지게 되어 유사도를 증가시켜 짧은 문헌에 비해 상대적으로 많은 기회를 갖게 된다.

따라서 문헌의 길이의 차이로 인한 유사도를 정규화를 통하여 보정해 주어야 할 필요가 있다. 정규화에는 코사인 정규화(Cosine normalization), 최대 단어 빈도 정규화(Maximum tf normalization), 바이트 길이 정규화(Byte length normalization) 등의 기법이 있다[8, 11].

2.4 형태소 분석(Morphological analysis)

자연언어는 다음과 같은 계층적 구조를 가지고 있다.

- 음소(phoneme) : 인간의 의미(의지) 전달에서 음성을 어떻게 사용하는가를 기초로 생각한 음의 단위
- 형태소(morpheme) : 의미를 가진 최소의 언어 단위, 하나 이상의 음소로 된다.
- 단어(word) : 하나의 의미의 총합을 이루며, 문법상 하나의 기능을 가진 최소의 언어 단위, 하나 이상의 형태소로 구성된다.
- 문장(sentence) : 전달하고자 하는 내용을 가지며, 완결된 언어 단위, 하나 이상의 단어로 된다.
- 텍스트(text) : 전달하고자 하는 내용을 표현하기 위하여 문장이 순서대로 모여진 집합

형태소 분석이란 어절을 입력으로 받아 형태소로 분리하고 그 형태소의 품사정보 및 분석에 필요한 정보를 추출하는 것을 의미한다[12].

본 논문에서 사용되는 단어가중치사전은 국민대학교의 HAM version 5.0.0.a를 이용하여 형태소 분석을 수행하여 문서에서 분류 키워드가 될 명사들을 추출하여 만들어졌다.

3. 제목의 단어 가중치를 이용한 자동 문서 분류

본 논문에서는 단어 중요도에 기반을 둔 단어 가중치를 이용하여 중등학교의 교육문서를 분류하는 시스템을 설계하고 구현한다. 일반적으로 단어의 중요도를 결정하는 주 요소는 단어빈도와 역문헌 빈도이다. 따라서 최종 가중치는 TF×IDF로 구해진다[11]. 또 하나의 요소는 문서길이 정규화[8, 11]인데 이는 문서 검색시스템에서 길이가 긴 문서일수록 단어의 출현 빈도가 높고 출현하는 단어의 종류가 많다는 이유로 짧은 문서에 비해 검색될 확률이 높기 때문에 정규화해야 한다는 것이다. 그런데 본 시스템에서는 문서의 제목에 대해서만 단어의 가중치를 계산하기 때문에 이를 고려할 필요 없다.

문서의 분류에 있어 역카테고리 빈도(ICF) 방법이 범주간의 구분에 도움이 되는 색인어의 중요도가 높다고 할 수 있

어 이 방법이 의미 있는 가중치 계산 방법이 될 수 있다[9]. 따라서 본 논문에서는 이 방법으로도 가중치를 계산하여 성능을 비교하기로 한다.

본 논문에서 다루고 있는 중등학교의 공문서는 일반 문서에 비해 제목이 내용을 대표하는 성질이 강하므로 제목에 대해서만 단어를 추출한다. 먼저 학습집단의 교육공문서의 제목으로부터 의미 있는 단어를 추출하여 문헌 빈도 가중치와 역문헌 빈도 가중치를 구하여 단어 가중치 사전을 구성하고, 이를 이용하여 분류할 문서의 제목의 단어에 대해 범주별 가중치를 계산하여 해당 범주로 분류한다.

3.1 단어 가중치 사전

문서를 분류하기 위해서는 구문분석, 의미분석이 선행되어야 정확한 분류가 가능하나, 현재까지 만족할 만한 성능에 이르지 못하고 있다. 교육공문서는 일반문서와는 달리 분류가 필요한 텍스트의 정보량이 많지 않다. 웹 상에서 전송되고 있는 교육공문서의 제목에서 형태소 분석과정을 통해 추출된 단어(체언 정보)에 가중치를 부여함으로써, 아주 적은 정보(단어)량으로도 문서를 분류할 수 있는 방법을 제안한다.

실제 중학교에서 사용된 총 1578개의 교육공문서를 단어 가중치 사전 구축에 사용하였고, 각 문서의 제목에서 형태소 분석을 통해 추출된 체언(단어) 중에서 불용어를 제거한 후 분류 키워드가 될 542개의 명사들을 재추출하였다. 이렇게 추출된 단어들로 가중치를 계산하여 가중치 사전을 구축하여 분류에 사용한다. <표 2>와 <표 3>은 각각 역문헌 빈도와 역카테고리 빈도를 이용하여 구성된 단어 가중치 사전을 보여주고 있다.

<표 1> 단어와 범주간의 문서빈도 테이블의 예

word	교무	서무	교육정보	연구	예체능	인성	학생	학습자료	환경	총빈도수
상담	1	1	0	0	0	9	0	0	0	11
자원	0	0	0	1	0	5	0	0	2	8
연수	85	21	7	6	6	6	2	2	1	136
교재	1	0	8	4	0	5	0	0	0	18
교육공무원	18	5	0	0	0	1	0	0	0	24
입학	15	2	0	0	2	0	0	0	0	19
:	:	:	:	:	:	:	:	:	:	:

단어의 가중치는 추출된 단어가 문서에 나타나는 빈도수와 단어가 나타난 전체 문서의 수를 기반으로 계산되는 TF-IDF 가중치 계산식에 의하여 다음과 같이 구해진다.

$$W_{ij} = \frac{freq_{ij}}{\max_k freq_{kj}} \times (\log \frac{N}{n_i})$$

위 식에서의 결과값이 단어 k_i의 가중치이다.

본 논문에서, 문서의 총 개수 N은 1578이고, n_i는 단어의 총 빈도(출현)수, $\max_k freq_{kj}$는 단어의 범주별 최고 빈도수를 의미한다.

예를 들어, 식(5)를 이용하여, <표 1>의 단어와 범주간의 문서빈도 테이블로 단어의 가중치를 계산해보면,

'상담'의 교무부에서의 가중치 = $\frac{1}{9} \times \log \frac{1578}{11} = 0.7960$

'상담'의 서무부에서의 가중치 = $\frac{1}{9} \times \log \frac{1578}{11} = 0.7960$

⋮

'입학'의 예체능부에서의 가중치 = $\frac{2}{15} \times \log \frac{1578}{19} = 0.8500$

가 된다.

<표 2> 역문헌 빈도에 의한 단어 가중치사전의 예

word	교무부 가중치	서무부 가중치	교육정보부 가중치	연구부 가중치	예체능부 가중치	인성부 가중치	학생부 가중치	학습자료부 가중치	환경부 가중치
상담	0.7960	0.7960	0	0	0	7.1644	0	0	0
자원	0	0	0	1.5247	0	7.6238	0	0	3.0495
연수	3.5364	0.8737	0.2912	0.2496	0.2496	0.2496	0.0832	0.0832	0.0416
교재	0.8067	0	6.4539	3.2269	0	4.0337	0	0	0
교육공무원	6.0380	1.6772	0	0	0	0.3354	0	0	0
입학	6.3750	0.8500	0	0	0.8500	0	0	0	0
:	:	:	:	:	:	:	:	:	:

<표 3> 역카테고리 빈도에 의한 단어 가중치사전의 예

word	교무부 가중치	서무부 가중치	교육정보부 가중치	연구부 가중치	예체능부 가중치	인성부 가중치	학생부 가중치	학습자료부 가중치	환경부 가중치
상담	2.5850	2.5850	0	0	0	23.2647	0	0	0
자원	0	0	0	2.5850	0.0000	12.9248	0	0	5.1699
연수	85.0000	21.0000	7.0000	6.0000	6.0000	6.0000	2.0000	2.0000	1.0000
교재	2.1699	0	17.3594	8.6797	0	10.8496	0	0	0
교육공무원	46.5293	12.9248	0	0	0	2.5850	0	0	0
입학	38.7744	5.1699	0	0	5.1699	0	0	0	0
:	:	:	:	:	:	:	:	:	:

3.2 역문헌 빈도 가중치를 이용한 문서 분류

역문헌 빈도에 의해 가중치가 부여된 사전을 이용하여 제

목에 나온 단어들의 범주별 가중치의 합을 각각 구하여 이들 중에서 가장 큰 값을 갖는 범주로 문서를 분류한다. 즉, 문서 t의 범주는 다음과 같이 결정된다.

$$TF\text{-}IDFmax(t)=max\{\sum W_{i1},\sum W_{i2},\cdots,\sum W_{i9}\}$$

예를 들어, 분류할 문서의 제목이 '2002 상담 자원봉사자 연수교재 배부'라 하자. 이것을 형태소 분석을 하면 {상담, 자원봉사자, 자원, 봉사자, 연수교재, 연수, 교재, 배부}의 단어가 추출된다. 단어 가중치 사전에서 검색된, "상담", "자원", "연수", "교재"로 부서별 가중치 합을 구해서 가장 높은 값을 갖는 부서로 이 문서는 분류된다. <표 2> 역문헌 빈도에 의한 단어 가중치 사전을 이용해서 단어의 부서별 가중치 합을 구하면 다음과 같다.

교무부의 가중치 합 : $\sum W_{i1}=5.1391$
서무부의 가중치 합 : $\sum W_{i2}=1.6697$
교육정보부의 가중치 합 : $\sum W_{i3}=6.7451$
연구부의 가중치 합 : $\sum W_{i4}=5.0012$
예체능부의 가중치 합 : $\sum W_{i5}=0.2496$
인성부의 가중치 합 : $\sum W_{i6}=19.0715$
학생부의 가중치 합 : $\sum W_{i7}=0.0832$
학습자료부의 가중치 합 : $\sum W_{i8}=0.0832$
환경부의 가중치 합 : $\sum W_{i9}=3.0911$
$\Rightarrow TF\text{-}IDFmax(t)=max\{\sum W_{i1},\sum W_{i2},\cdots,\sum W_{i9}\}$
$\Rightarrow \sum W_{i6}=19.0715$

따라서 위 문서는 인성부로 분류된다.

마찬가지 방법으로, <표 3>의 역카테고리 빈도에 의한 단어 가중치사전을 이용해서 단어의 부서별 가중치 합으로 TF-ICFmax(t)를 구하면,

$\Rightarrow TF\text{-}ICFmax(t) =max\{\sum W_{i1},\sum W_{i2},\cdots,\sum W_{i9}\}$
$\Rightarrow \sum W_{i1}= 89.7549$ 이다.

이 방법으로 하면 위 문서는 교무부로 분류된다. 이와 같이 단어빈도 가중치 기법을 달리함에 따라, 계산에 적용되는 가중치와 가중치 합이 달라져서 분류되는 문서의 범주가 달라질 수 있다.

4. 시스템 설계 및 구현

4.1 시스템 설계

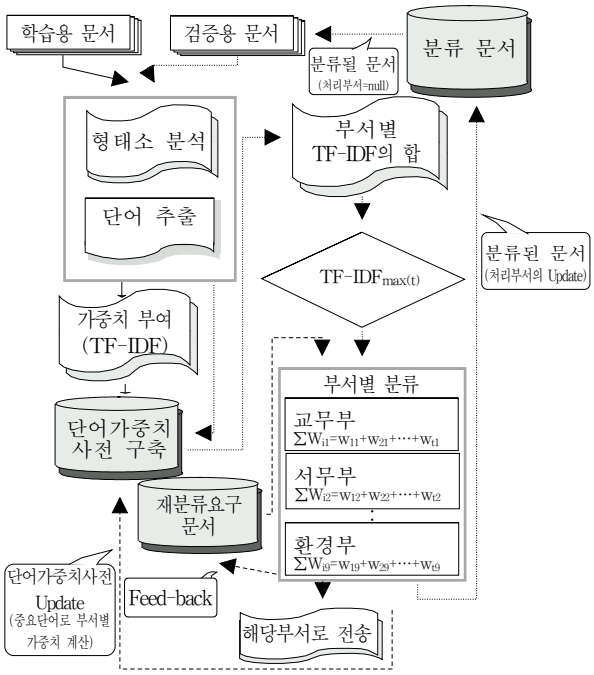
부서별 역문헌 빈도 가중치 합을 계산하는 모듈은 문서의 제목을 입력받아서, 현재 저장된 단어들의 목록(가중치사전의 단어)을 탐색하여 일치하는 단어가 나오면 가중치를 구하게 된다. 가중치를 가지는 제목의 모든 단어에 대해 부서별 가중치를 구해 이를 누적함으로써 이 문서의 부서별 가중치를 구한다. 부서별 가중치의 합이 구해지면 TF-IDFmax(t)에 의해 가중치 합이 가장 큰 부서로 분류한다.

재분류요구 문서를 분류하는 모듈은 재분류 처리 요구 문서를 입력받아서, 처리 부서를 갱신시켜주고, 관련 정보는 별도로 저장하여 차후에 단어 가중치 사전의 갱신에 사용한다.

4.2 구현

4.2.1 자동문서분류시스템의 메인 화면

초기화면의 다음 화면으로 분류 처리할 문서가 있을 때에는 처리할 문서의 건수를 보여준다. 상단 메뉴의 문서분류를 선택해도 같은 화면이 나온다.

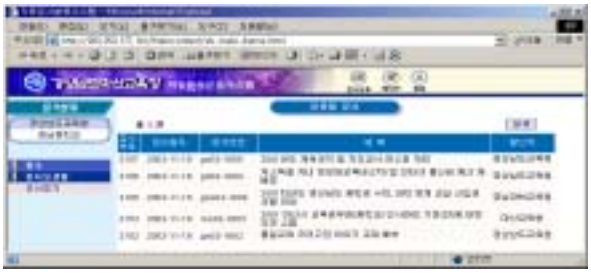


<그림 1> 시스템 설계 구성도

<그림 2> 자동문서분류시스템의 메인 화면

4.2.2 분류할 문서 목록 화면

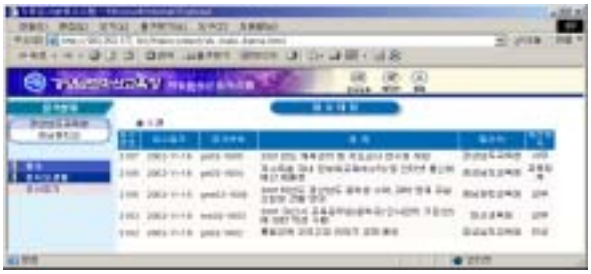
왼편메뉴의 접수를 선택하면 앞에서 보여준 분류할 문서 건수에 해당하는 문서 목록을 보여준다. 이 화면에서 다시 분류할 문서 목록 상단의 '분류'를 선택하면 자동문서분류작업을 하게된다.



<그림 3> 분류할 문서 목록 화면

4.2.3 분류된 문서의 출력 화면

이전 화면에서 '분류'를 실행시켜서 분류작업이 끝나면 바로 다음 화면으로 처리를 끝내고 분류된 문서의 목록을 보여준다.



<그림 4> 분류된 문서의 출력 화면

4.2.4 재분류요구 문서의 처리

재분류 요구 문서에 대해 처리 부서를 갱신하고 새로운 처리 부서로 분류하게 한 중요단어를 입력한다. 입력된 중요단어와 부서명, 문서의 제목은 데이터베이스에 저장되여 차후에 단어 가중치 사전의 갱신에 사용한다.

<그림 5> 재분류요구 문서의 처리 화면

5. 성능평가

본 논문에서는 경남 마산양덕중학교에서 1년간(2001년 1월~12월) 접수 처리된 후 분류 담당자(교감)에 의해 9개의 범주(부서-교무부, 서무부, 교육정보부, 연구부, 예체능부, 인성부, 학생부, 학습자료부, 환경부)로 분류된 총 2104개의 교육공문서를 대상으로 실험하였다. 각 범주마다 임의로 75%의 문서를 추출하여 학습집단을 구성하고(1578건), 나머지 25%(526건)를 검증집단으로 하였다.

실험의 성능평가 척도로는 정확도(accuracy)를 사용하였는데, 정확도는 다음과 같이 구해진다[2, 13].

<표 4> 정확도 분할표

	적합문헌(r)	부적합문헌(F)	전체문헌(N)
분류된 문헌	적중(a)	잡음(b)	a+b
분류되지않은문헌	누락(c)	배제(d)	c+d
전체문헌	a+c	b+d	N

※모든 문서가 분류되면, c+d= 0이다.

$$\text{정확도} = \frac{\text{분류된적합문헌수}}{\text{분류된문헌의총수}}$$

역문헌 빈도 가중치와 역카테고리 빈도 가중치를 사용한 두 경우에 대해 정확도를 측정하였는데 각각의 장확도는 다음과 같다.

- 역문헌 빈도 가중치 적용 시 정확도(a=434, b=92, c=d=0)
  $\frac{434}{526} \times 100 = 82.5\%$
- 역카테고리 빈도 가중치 적용 시 정확도
  (a=379, b=147, c=d=0)
  $\frac{379}{526} \times 100 = 72.04\%$

위 실험 결과는 중등학교 교육공문서의 자동분류 시스템에서는 역카테고리 빈도보다는 역문헌 빈도를 이용해서 단어에 가중치를 부여하는 방법이 분류 성능이 더 높음을 보여 주고 있다. 또한 일반적으로 사용되는 검색시스템이나 분류시스템은 문서 전체에 포함된 단어(정보)들을 대상으로 하여 분류하여 정확도가 75%이상이면 좋은 시스템으로 평가하는데[3, 7, 10], 본 논문의 역문헌 빈도 가중치를 이용한 분류 방법은 제목에 나오는 적은 양의 단어로도 이보다 우수한 성능을 보여 주고 있음을 알 수 있다.

6. 결론

정보통신기술의 발전, 특히 인터넷 서비스의 영향으로 온라인으로 많은 양의 문서가 유통되고있으며, 교육분야에서도 예전의 우편이나 인편발송에 비하여 편리함과 처리 속도 면에서 우수한 전자문서를 주로 사용하고 있다. 전자문서관리 시스템에서 유통되는 교육공문서의 분류를 현재의 수작업 방법 대신 업무의 효율성 증대를 위해 자동분류방법의 개발이 요구되고 있다.

이에 본 논문에서는 중등학교에서의 공문서 분류를 위해 문서 제목의 단어에 대해 역문헌 빈도 가중치의 적용을 제안하고 이를 기반으로 자동분류시스템의 설계하고 구현하였다.

실제 중등학교 공문서에 역문헌 빈도 가중치와 역카테고리 빈도 가중치를 적용시켜, 본 논문에서 설계한 시스템의 성능을 평가하였다. 역카테고리 빈도 가중치를 적용시켜 분류한 것보다 역문헌 빈도 가중치를 적용시켜 분류한 것이 정확도가 높아 중등학교 공문서 분류에는 역문헌 빈도 가중치를 이용하는 분류 기법이 우수함을 보여 주었다.

그리고, 본 논문의 분류시스템은 제목에 나오는 적은 양의 단어 정보를 사용하고도 문서 전체 내용을 기반으로 하는 일반적으로 사용되는 검색시스템이나 분류시스템보다 우수한 성능을 보여 주고 있다.

향후 다양한 중등학교 교육공문서 데이터 집합에 대해 이 논문에서 제안된 시스템의 성능을 보다 객관적으로 검증할 필요가 있다. 뿐만 아니라 분류 성능을 더욱 향상시키기 위하여 충분한 실험을 통하여 이 논문에서 대상으로 삼고 있는 중등학교의 공문서의 특성을 고려한 단어 가중치 계산 방법을 찾아야 할 것이다. 아울러 오분류된 문서를 이용한 체계적인 시스템의 성능 개선 방안에 대한 연구도 필요하다.

참고 문헌

[1] 고영중, 박진우, 서정연(2002), 문장 중요도를 이용한 자동 문서 범주화, 정보과학회논문지: 소프트웨어 및 응용, 29-6, 417-424

[2] 국민상, 정영미(2000), 자질 선정에 따른 Naive Bayesian 분류기의 성능 비교, 한국정보관리학회 학술대회 논문집, 7, 33-36

[3] 김진상, 신양규(2000), 베이지안 학습을 이용한 문서의 자동분류, 한국데이터 정보과학회논문지, 11-1, 19-30

[4] 노현아, 김민수, 김수형, 박혁로(2002), 단어 빈도 가중치를 이용한 자동 문서 분류, 정보처리학회 추계학술발표대회, 9-2, 581-584

[5] 류근호, 이제환(2000)역, 정보저장 및 검색, 시그마프레스

[6] 방선이, 양재동(2001), K-NN과 객체 지향 시소러스를 이용한 웹 문서 자동 분류, 정보과학회 추계학술대회, 28-2, 145-147

[7] 이경찬, 강승식(2002), 범주 대표어의 가중치 계산 방식에 의한 자동 문서 분류 시스템, 한국 정보과학회 학술발표논문집(B), 29-2, 475-477

[8] 이재윤, 최보영, 정영미(2000), 문헌 자동분류에서 용어가중치 기법에 대한 연구, 제7회 한국정보관리학회 학술대회 논문집, 41-44

[9] 조광제, 김준태(1997), 역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류, 한국정보과학회 봄 학술발표논문집, 4-2, 508-510

[10] 한광록, 선복근, 한상태, 임기욱(2000), 인터넷 문서 자동 분류 시스템 개발에 관한 연구, 한국정보처리학회 논문지, 7-9, 2867-2875

[11] Amitabh Kumar Singhal(1997), Term Weighting

Revisited, Ph.D. Dissertation, Cornell University.
[12] http://www.gurugail.com/NLP/page.html?subject=morphAnal.html
[13] http://www.khsme.com/info_search/chap10/page_6.htm

저자 소개

강 현 희

1994년 창원대학교 통계학과(이학사)
2003년 경남대학교 교육대학원 전자계산교육학과(교육학석사)
관심분야: 데이터베이스 응용, 문서자동분류

진 민

1982년 서울대학교 계산통계학과(이학사)
1984년 한국과학기술원 전산학과(공학석사)
1997년 University of Connecticut, Computer Science and Eng. (Ph.D.)
1985년~ 경남대학교 정보통신공학부 교수
관심분야: 객체지향 데이터베이스,
데이터 모델링, XML 데이터베이스

주요어: 중등학교 공문서 분류, 문서제목 단어가중치, 역문헌빈도가중치