

문서 요약 기법을 이용한 자동 문서 범주화

박진우⁰ 고영중 서정연
서강대학교 컴퓨터학과
[\[jwpark, kyj\]@nlprep.sogang.ac.kr](mailto:[jwpark, kyj]@nlprep.sogang.ac.kr), seojy@ccs.sogang.ac.kr

Automatic Text Categorization Using Text Summarization Techniques

Jinwoo Park⁰ Youngjoong Ko Jungyun Seo
Dept. of Computer Science, Sogang University

요약

자동 문서 범주화란 문서의 내용에 기반하여 미리 정의되어 있는 범주에 문서를 자동으로 분류하는 작업이다. 문서 분류를 위해서는 문서들을 가장 잘 표현할 수 있는 자질들을 정하고, 이러한 자질들을 통해 분류할 문서를 표현해야 한다. 기존의 연구들은 문장간의 구분 없이, 문서 전체에 나타난 각 자질의 빈도수를 이용하여 문서를 표현한다. 그러나 하나의 문서 내에서도 중요한 문장과 그렇지 못한 문장의 구분이 있으며, 이러한 문장 중요도의 차이는 각각의 문장에 나타나는 자질의 중요도에도 영향을 미친다.

본 논문에서는 문서 요약에서 사용되는 중요 문장 추출 기법을 문서 분류에 적용하여, 문서 내에 나타나는 각 문장들의 문장 중요도를 계산하고 문서의 내용을 잘 나타내는 문장들과 그렇지 못한 문장들을 구분하여 각 문장에서 출현하는 자질들의 가중치를 다르게 부여하여 문서를 표현한다. 이렇게 문장들의 중요도를 고려하여 문서를 표현한 기법의 성능을 평가하기 위해서 뉴스 그룹 데이터를 구축하고 실험하였으며 좋은 성능을 얻을 수 있었다.

1. 서론

자동 문서 범주화(automatic text categorization)는 미리 정의된 범주(category)에 문서를 자동으로 할당하는 기법과 관련된 연구분야로서, 대량의 문서의 효율적인 관리 및 검색을 가능하게 하는 동시에 방대한 양의 수작업을 감소시키는 데 그 목적이 있다.

자동 문서 범주화 과정은 문서를 어떤 자질을 통해 표현할 것인가를 다루는 자질 추출(feature extraction) 과정과 추출된 자질로 표현된 문서를 어느 범주로 할당할 것인가를 결정하는 문서분류(text classification) 과정으로 구성된다.

자질 추출과정에는 추출된 자질로 어떻게 문서를 표현할 것인가에 대한 색인(indexing) 과정이 포함되며, 가장 일반적인 색인 방법은 이른바 벡터 공간 모델(vector space model)이다. 이 모델은 문장의 구분 없이 전체 문서에 출현한 각 자질의 빈도수(TF)를 가지고 표현하는 방법이다. 그러나 문서 내에 나타나는 문장들 중에는 해당 문서의 핵심 내용을 잘 나타내는 문장과

그렇지 못한 문장들이 있으며, 이러한 문장 중요도의 차이는 각 문장에 나타나는 자질의 중요도에도 영향을 미친다. 그러므로, 본 논문에서는 문서 요약 기법의 적용을 통해 중요한 문장과 그렇지 못한 문장을 구분하고, 각 자질이 어느 정도의 중요성을 지닌 문장으로부터 출현했는지를 색인 과정에 적용한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 자동 문서 분류 시스템과 자동 문서 요약 분야에서 수행됐던 관련 연구들에 대해 살펴보고, 3장에서는 본 논문에서 제안하는 자동 문서 분류 시스템 모델에 대하여 설명한다. 그리고 4장에서 실험 및 평가를 하며, 5장에서는 결론 및 향후 연구를 기술한다.

2. 관련연구

문서 분류 시스템은 미리 정의된 2개 이상의 범주에 문서의 내용을 파악하여 가장 관련이 있는 범주로 할당하는 시스템이다.

이는 문서 집합을 색인하고 분류하는 데 필요한 자질

을 추출하는 자질 추출 과정과 추출된 자질을 기반으로 어떤 범주로 할당할 것인가를 결정하는 문서 분류 과정으로 나눌 수 있다. 후자의 문서 분류 방법에는 규칙 기반 모델(Rule-based Model), 확률적인 접근 방법의 단순 베이저언 확률 모델(Naive Bayesian probabilistic approach), 기계 학습(Machine Learning) 방법을 이용한 지지 벡터 기계(Support Vector Machine: SVM), 그리고 정보 검색 기법을 이용한 최근린법(k-Nearest Neighbor) 등이 있다[1,2]. 전자의 자질 추출 과정은 다시 자질 선택(feature selection) 단계와 색인(indexing) 단계로 나눌 수 있다.

자질 추출 과정 중 자질 선택 단계는 문서에 나타난 여러 단어들 중 범주화에 유용하게 사용될 만한 단어들을 선택하는 과정으로 문서 빈도(document frequency), 상호 정보 척도(mutual information), 카이 제곱 통계량(χ^2 static), 정보 획득량(information gain) 등의 기법이 있다[3].

자질 추출 과정 중 색인 단계는 선택된 자질을 통해 문서를 표현하는 단계로서, 일반적으로 벡터 공간 모델이 사용된다[4]. 이것은 문서 전체에 나타난 자질들을 이용하여 문서를 하나의 벡터로 표현하는 것으로 보통 자질의 빈도수와 역 문헌 빈도수(IDF) 혹은 역 범주 빈도수(ICF)를 사용하여 문서를 표현한다[5]. 그러나 이러한 기존의 방법은 문서가 가진 자질의 위치정보나 문장간 구분 등의 구조적 정보는 고려되지 못한다는 단점이 있다.

이러한 한계를 극복하기 위하여 다양한 연구가 있었는데, 먼저 문서의 구조적 정보를 살리기 위해 단어의 위치나 출현한 문장의 위치에 따라 가중치를 차등 적용한 방법이 연구되었으나, 모든 문서를 두괄식 또는 미괄식으로 가정하였기 때문에 신문 기사(article) 등 형식적인 문서를 제외하곤 그 적용이 힘들다[6]. 또한 제목이 있는 문서의 경우에 일반적으로 제목이 문서 전체의 내용을 대표하는 경우가 많으므로 제목에 가중치를 주는 방법이 있었지만, 제목이 문서를 대표할 만큼의 중요한 의미를 포함하고 있지 않다면 불필요하거나 오히려 모호성을 키우는 결과를 가져오기도 한다[7]. 이

러한 현상은 뉴스그룹(newsgroup)이나 이메일(e-mail) 등 비형식적인 문서에서 더 크게 부각된다.

하지만 중요한 것은 의미적으로 완전하지 못한 제목들도 문서를 작성한 의도와 목적을 포함하는 경우가 있다는 것이다. 따라서 제목에서 내포한 의도와 부합되면서 부가적인 설명을 하고 있는 문장이 본문에 존재한다면, 이 문장은 문서내의 다른 문장에 비해 더 중요하게 고려되어야 한다. 그러므로 본 논문에서는 문서내의 전체 문장들 중에서 중요한 문장들을 찾아내기 위해 문서 요약 기법의 적용을 제안한다.

자동 문서 요약에 대한 기존 연구들은 방법론에 따라 크게 통계 정보에 기반한 방법과 언어학적 지식에 기반한 방법으로 나누어진다[8,9,10]. 본 논문에서는 자동 문서 분류를 하는 데 있어서 추가 부담을 줄이기 위해 적은 비용으로 빠르게 어느 정도 신뢰할 만한 결과를 얻을 수 있는 통계 정보를 이용한 방법을 사용하였으며, 이는 자동 문서 분류를 위한 학습과정에서 얻어진 정보를 그대로 통계 정보로 활용할 수 있다는 장점이 있다.

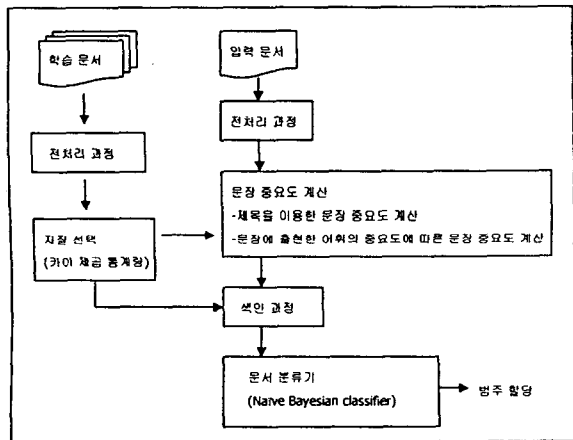
본 논문에서는 여러 가지 문서 요약 기법 중에서 다음의 두 가지 문서 요약 기법을 적용하였다. 첫째는 제목과 문서의 모든 문장을 비교하여 문장간 유사도를 구하고 제목과 유사할수록 중요한 문장으로 결정하는 방법[8]이고, 둘째는 각각의 문장에 출현한 자질의 중요도에 따라 해당 문장의 중요도를 결정하는 방법[9]이다.

3. 문서 요약 기법을 이용한 자동 문서 범주화 시스템

전체 시스템은 크게 학습 과정과 문서 분류 과정으로 나뉜다. 학습 과정에서는 전처리 과정을 통해 학습 문서로부터 내용을 추출하고, 자질 선택 과정에서는 카이 제곱 통계량을 이용하여 정보량에 따라 내용을 순위화한다. 자질 선택 과정의 결과 중에서, 각 내용의 정보량은 문장 중요도를 계산하는 데 사용되며, 정보량의 순위는 색인 과정에서 자질의 수를 제안하는 데 사용된다.

입력 문서는 전처리 과정을 통해 문장 단위로 내용이 추출되며, 문장 단위로 추출된 내용을 통해 문서 내의 모든 문장들을 문장 벡터로 표현한다. 문장 중요

도 계산 과정에서는 두 가지 방법으로 문장 중요도를 계산한다. 첫째는 제목과의 유사도를 구하여 높은 유사도를 가질수록 높은 중요도를 주는 방법이고, 둘째는 문장에 중요한 자질이 많이 나오면 높은 중요도를 주는 방법이다. 결국 문장의 중요도는 이 두 가지 방법으로 얻어진 문장 중요도를 통합함으로써 얻어지고 이렇게 얻어진 문장 중요도는 색인 과정에 적용된다. 색인 과정에서는 전처리 과정에서 만들어진 문장 벡터들을 모두 합함으로써 하나의 문서 벡터를 만드는데, 이때 문장 벡터에 해당 문장의 문장 중요도가 곱해진다. 따라서 문서 벡터에 나타나는 각 자질의 빈도수는 문장 중요도에 따라 실제 나타난 빈도수보다 더 높은 값을 가지게 된다. 이렇게 표현된 문서 벡터는 단순 베이지언 확률 모델을 사용한 문서 분류기를 통해 범주를 할당 받게 된다.



[그림 1] 제안한 시스템의 전체 구성도

3.1 전처리 과정

뉴스그룹의 문서는 총 7개의 필드(제목, 저자, 날짜, 그룹, 서버, ID, 본문)로 이루어져 있으며, 제안된 시스템에서는 제목과 본문만을 이용했다. 문장의 내용이나 특징을 잘 반영하는 단어를 내용어(content word)라고 하는데 본 시스템에서는 내용어로 형태소 분석기 dianaM¹의 결과 중 명사만을 고려하였다.

¹ dianaM은 (주)다이퀘스트닷컴(www.diquest.com)에서 만든 형태소 분석기이다.

전처리 과정을 통해 입력 문서는 문장단위로 내용어를 추출하게 되고, 추출된 내용어를 사용하여 문장 벡터들을 구성한다.

3.2 자질 추출

Yiming Yang은 [3]에서 여러 가지의 자질 추출 방법을 사용하여 실험을 한 결과 카이제곱 통계량과 정보 획득량을 사용하는 것이 가장 효과적임을 보였다. 본 논문에서는 이를 바탕으로 비교적 구현이 쉽고 고빈도 단어에 친화적인 카이 제곱 통계량을 사용하여 자질을 추출한다. 카이 제곱 통계량을 구하기 위한 식은 다음과 같다.

$$\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (1)$$

여기에서 A는 범주 c에 속해 있는 문서 중 용어 t를 포함하고 있는 문서의 수, B는 범주 c에 속하지 않은 문서 중 용어 t를 포함하고 있는 문서의 수, C는 범주 c에 속해 있는 문서 중 용어 t를 포함하지 않은 문서의 수, 그리고 D는 범주 c에 속하지 않은 문서 중 용어 t를 포함하지 않은 문서의 수이다.

각 범주별로 얻어진 카이 제곱 통계량 값은 다음과 같은 식에 의해 가장 큰 값이 해당 용어의 자질 값이 되며 이 값을 순위화하여 자질 값이 높은 순으로 추출한다.

$$\chi^2_{\max}(t) = \max_{i=1}^m \{ \chi^2(t, c_i) \} \quad (2)$$

또한 여기서 구해진 자질 값은 문장의 중요도를 구하기 위해 식 (4)에서 다시 사용된다.

3.3 문장 중요도 계산

문장 중요도는 두 가지 방법에 의해서 구해진다. 하나는 제목과의 유사도를 구하여 유사도가 높을수록 높은 중요도를 주는 방법이고, 다른 하나는 문장에 출현한 단어들의 정보량을 판단하여 정보량이 많은 단어들이 나온 문장에 높은 중요도를 주는 방법이다.

결국 문장 중요도는 이 두 가지 방법에 의해 구해진 문장 중요도를 합하여 결정된다.

3.3.1 제목을 이용한 문장 중요도 계산

일반적으로 제목은 본문에 비해 문서 전체에서 중요한 의미 또는 의도를 갖는 경우가 많다. 따라서 제목에 가중치를 부여하는 방법이 이전에 시도되었으나 제목이 문서 전체의 의미를 대표하지 못하는 경우 오류를 일으키거나 그 적용에 한계가 있었다. 이러한 문제점은 뉴스 그룹 문서나 이메일 등의 비형식적인 문서에서 더욱 심하게 나타난다. 다음은 뉴스 그룹 문서들의 제목과 본문의 일부만을 모은 예제이다.

- 1) 제목 : 노래 제목과 가수 이름을 알고 싶어요
본문: 제가 음악에 대해 잘 몰라서 질문 중...
- 2) 제목: 질문이 있는데요....
본문: 게임관련 자료를 얻으려면...
- 3) 제목: 자료를 구하고 싶은데...
본문: 다악은 우리 음악과 차의 향기가 어우러지는 공연으로...
공연내용은 연주자 사정상 바뀔 수 있습니다.
- 4) 제목: 공연정보

첫번째 예제 1)의 제목은 제목만을 보고서도 음악과 관련된 문서라는 사실을 알 수 있으며, 글을 올린 목적을 명확히 파악할 수 있다. 그러나 다음에 나오는 2), 3), 4)의 제목은 무슨 질문이 있는 건지, 무슨 자료를 원하는 것인지, 무슨 공연에 대한 정보인지를 명확히 알 수 없으며, 따라서 이러한 제목들은 제목 자체만으로는 문서를 분류하는 데 별 도움을 주지 못한다. 하지만 중요한 것은 이렇게 모호한 제목들조차도 문서 전체의 의도를 포함하고 있다는 것이다. 따라서 제목에서 내포한 의도와 부합되면서 부가적인 설명을 하고 있는 문장이 본문에 존재한다면 이 문장은 문서내의 다른 문장에 비해 더 중요하게 고려되어야 할 것이다. 예제 2)를 보면 "질문"이라는 단어를 공유한 본문의 문장에서 질문 내용이 음악에 관련한 것이라는 사실을 말하고 있으며, 예제 3)에서는 "자료"라는 단어를 공유한 본문의 문장에서 원하는 자료가 게임과 관련된 자료라는 것을

의미하고 있다. 마지막으로 예제 4)에서 제목에 나오는 "공연"이라는 단어는 본문에서 지속적으로 출현하면서 공연에 대하여 자세히 설명하는 문장들을 이끌고 있다. 따라서 제목과 유사한 문장에 출현하는 단어들은 해당 문서의 의도를 자세히 나타내고 있는 경우가 많으며, 이러한 단어들은 문서를 분류하는 데 중요한 기준이 된다.

본 논문에서는 제목의 의미 또는 의도를 포함하고 있는 문장을 중요한 문장으로 인정하고, 이를 판단하기 위해 제목과 문서내의 모든 문장들과 유사도를 구하여 제목과 유사할수록 높은 중요도를 부여한다.

제목과 문서의 각각의 문장들과 유사도를 구하기 위해 각 문장은 내용어의 벡터로 표현되고, 다음과 같이 내적을 통해 그 유사도를 계산한다. 이렇게 구한 제목과 각 문장의 유사도는 해당 문서에서 가장 높은 문장 유사도 값으로 나누어 줌으로써 0에서 1사이로 정규화한다.

$$Sim(S_i, T) = \frac{\hat{S}_i \cdot \hat{T}}{\max_{S_j \in D} (\hat{S}_j \cdot \hat{T})} \quad (3)$$

위의 식에서 D는 문서이고, T는 문서 D의 제목, S는 문서 D의 문장이다. $Sim(S_i, T)$ 는 제목 T와 i 번째 문장 S_i 의 유사도이며, \hat{T} 는 내용어로 이루어진 제목의 문장 벡터이고, \hat{S}_i 는 문장 S_i 의 문장 벡터이다.

3.3.2 출현 자질에 따른 문장 중요도 계산

만약 제목이 아무런 의미도 갖지 못함은 물론 문서의 의도조차도 내포하지 못한다면 3.3.1절에서 사용한 방법은 불필요하거나 문서를 왜곡시킬 수 있을 것이다. 또한 제목과 유사한 문장이라도 중요한 의미를 담고 있는 자질을 포함하고 있지 않다면 중요한 문장이 될 수 없으며, 반대로 제목과 유사성이 없는 문장이라도 중요한 의미를 포함한 자질들이 나타나는 문장이라면 중요하게 고려해야 한다.

자동 문서 요약 기법 중에는 문장에 출현하는 각 자질들의 중요도(Centroid)를 측정하고 이를 합하여 문장의 중요도를 결정하는 방법이 있다[9]. 이 방법을 통해

서 제목이 없는 경우나 제목이 있어도 전혀 무의미한 경우일 때에도 문장의 중요도를 구할 수 있다.

본 논문에서는 학습 과정을 통해 얻어진 문헌 빈도수와 카이 제곱 통계량을 이용하여 각 자질의 중요도를 구하고, 문장에서 나타난 자질들의 중요도를 합하여 해당 문장의 문장 중요도로 나타낸다. 이에 대한 식은 다음과 같다.

$$Cen(S_i) = \frac{\sum_{t \in S_i} C_t}{\max_{S_j \in D} \left(\sum_{t \in S_j} C_t \right)} \quad (4)$$

$Cen(S_i)$ 는 문장 S_i 의 문장 중요도이고, C_t 는 자질 t 의 자질 중요도이다. 여기서 자질 중요도는 각 자질의 빈도수, 카이 제곱 통계량, 그리고 역 문헌 빈도수의 곱으로 구한다.

위의 식을 통해 더 적은 수의 문서에 나타날수록, 카이 제곱 통계량에 의해 얻어진 정보량 값이 클수록 중요한 자질이고, 이러한 중요한 자질이 많이 나타난 문장일수록 문장 중요도가 크게 나타난다.

3.3.3 문장 중요도의 통합

두 가지 방법의 문장 유사도 계산이 끝나면 식(3)의 제목과의 문장 유사도와, 식(4)에서의 출현 자질들의 중요도를 이용한 문장 중요도는 다음과 같은 식에 의해 통합된다.

$$Score(S_i) = 1.0 + k_1 \times Sim(S_i, T) + k_2 \times Cen(S_i) \quad (5)$$

위의 식에서 k_1 과 k_2 는 두 가지 방법의 적용 정도와 비율을 조정하기 위한 실험값이며, 이 식을 통해 각 문장은 하나의 문장 중요도 값을 가지게 된다. 제목과의 문장 유사도를 구하기 위해서는 무엇보다도 제목이 있어야 하므로, 두 가지 방법을 이용하여 문장 중요도를 구하는 것은 본 논문에서의 실험 대상인 뉴스 그룹 문서와 같이 제목이 있는 경우에만 가능하다. 만약 제목이 없는 문서라면 자질 중요도를 이용하는 식(4)의 방법만을 이용하여 문장 중요도를 판단해야 할 것이다.

3.4 색인 과정

문장 중요도 계산 과정에서 얻어진 문장 중요도는 문서를 하나의 벡터로 표현할 때 자질의 빈도수에 가중치를 주기 위해 사용된다. 즉 문서에 출현한 자질의 빈도수는 각 문장에 출현한 자질의 빈도수의 합으로 구해지는데 이때 출현한 문장의 중요도에 따라 더해지는 빈도수가 달라지게 된다. 이를 나타내는 식은 다음과 같다.

$$N(t|d) = \sum_{S_i \in d} tf(S_i, t) \times Score(S_i) \quad (6)$$

위 식에서 $tf(S_i, t)$ 는 문장 S_i 에서 출현한 자질 t 의 빈도수이며, $N(t|d)$ 는 문장 중요도에 의해 가중치가 적용된 문서 d 에 출현한 자질 t 의 빈도수이다.

위 식에 따르면 각 자질은 출현한 문장의 중요도 ($Score(S_i)$) 만큼의 가중치를 받게 되므로, 중요한 문장에 나온 자질은 실제로 출현한 빈도수보다 높은 값을 가지게 된다.

3.5 문서 분류기

본 논문에서 사용한 문서 분류기는 단순 베이저언 확률 모델이다. 단순 베이저언 확률 모델은 대상 문서가 각 범주에 속할 확률을 구해 가장 큰 확률 값을 갖는 범주에 그 문서를 할당하는 기법이다. 이는 다음과 같은 식으로 나타낸다.

$$\begin{aligned} \arg \max_c [\Pr(c|d)] &= \arg \max_c \left[\frac{\Pr(c)\Pr(d|c)}{\Pr(d)} \right] \\ &= \arg \max_c \left[\Pr(c) \prod_{i=1}^T \Pr(t_i|c)^{N(t_i|d)} \right] \end{aligned} \quad (7)$$

여기서 $N(t_i|d)$ 는 문서 d 에서의 자질 t_i 가 출현하는 빈도수를 의미하고 T 는 전체 문서 집합내의 자질의 수를 나타낸다. 일반적으로 문서 d 에 자질 t_i 의 빈도수가 높고 범주 c 에 자질 t_i 가 많이 나타나면 문서 d 가 범주 c 에 속할 확률은 높아진다. 그러나 위의 식에서는 자질 t_i 가 문서 d 에서 출현 빈도수가 높을수록 오히려 $\Pr(c|d)$ 의 확률 값이 작아진다. 이러한 문제를 해결하기 위하여 Kulback-Leiber Divergence를 사용하여 다음과 같이 변환한다[11].

$$\Pr(c) \prod_{i=1}^T \Pr(t_i | c)^{N(t_i, c)} \quad (8)$$

$$\propto \frac{\log \Pr(c)}{n} + \sum_{i=1}^T \Pr(t_i | d) \log \left(\frac{\Pr(t_i | c)}{\Pr(t_i | d)} \right)$$

여기서 n 은 문서 d 에 출현하는 모든 자질의 빈도수의 합이고, $\Pr(c)$ 는 전체 학습 문서 집합에서의 해당 범주가 나타날 확률을 의미하며, $\Pr(t_i | c)$ 는 해당 범주에서 자질 t_i 가 나타날 확률을, 그리고 $\Pr(t_i | d)$ 는 대상 문서에서 자질 t_i 가 나타날 확률을 의미한다. 각각의 확률식은 다음과 같다[12].

$$\Pr(t_i | c) = \frac{N(t_i, c) + 0.5}{\sum_{j=1}^T N(t_j, c) + 0.5 \times T_c} \quad (9)$$

$$\Pr(t_i | d) = \begin{cases} \frac{N(t_i, d) + 0.5}{\sum_{j=1}^T N(t_j, d) + 0.5 \times T_d} \\ 0 \end{cases}$$

여기서 $N(t_i | c)$ 는 범주 c 에서의 자질 t_i 가 출현한 빈도이며 T_c 는 범주 c 의 자질의 총수이다. 그리고 $N(t_i | d)$ 는 3.4의 식(6)에서 구한 문서 d 에서 출현한 자질 t_i 의 빈도수이다.

4. 실험 및 평가

4.1 성능 평가 방법

본 논문에서 하나의 문서는 단지 하나의 범주로만 할당된다. 본 시스템의 평가 방법으로는 정보 검색 분야에서 일반적으로 사용되는 정확율(accuracy), 재현율(recall), 그리고 FI-measure를 사용하고, 모든 범주의 성능을 통합하여 평가하기 위한 기법으로는 micro-averaging 기법을 사용한다.

4.2 실험 데이터

실험에 사용된 데이터는 서강대학교 뉴스 서버로부터 수집한 문서이다. [표 1]과 같이 총 10,331개의 문서로 구성되어 있으며, 학습 문서로 7,224개를, 실험 문서로 3,107개를 사용하였다. 범주는 총 15개를 사용하였고 각 범주에 속해 있는 문서의 수는 각각 다르다. 각 문서는 하나의 범주만을 가지며 중복 할당을 허용하지 않

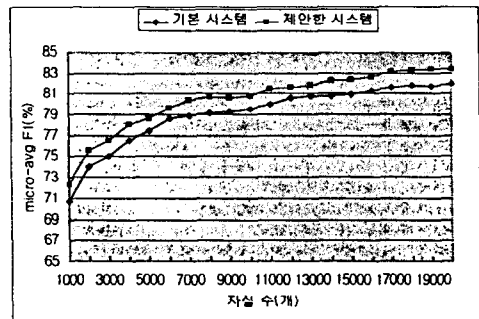
았다.

범주	학습 문서	실험 문서	총계
han.arts.music	315	136	451
han.comp.databases	198	86	284
han.comp.devtools	404	174	578
han.comp.lang	1387	595	1982
han.comp.os.linux	1175	504	1679
han.comp.os.windows	517	222	739
han.comp.sys	304	131	435
han.politics	1469	630	2099
han.rec.cars	291	126	417
han.rec.games	261	112	373
han.rec.movie	202	88	290
han.rec.sports	130	56	186
han.rec.travel	102	45	147
han.sci	333	143	476
han.soc.religion	136	59	195
총계	7224	3107	10331

[표 1] 실험 데이터의 구성

4.3 실험 결과

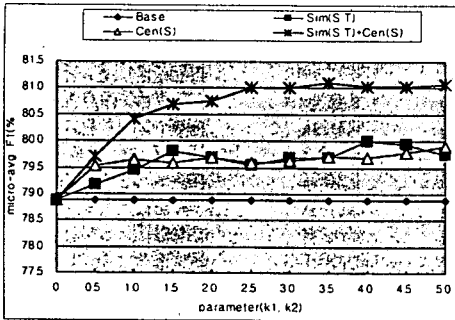
첫번째 실험은 학습 과정에서 얻어진 총 69,793개의 내용어 중 자질 선택 과정을 통해 1000개에서 20,000개까지 자질의 수를 제한하여 기존의 시스템과 제한한 시스템의 성능을 비교하였다. 이때 식(5)에서 정의한 상수 k_1 과 k_2 는 1.0으로 정하였으며 실험 결과는 다음과 같다.



[그림 2] 자질 수에 따른 성능 비교 그래프

위의 첫번째 실험 결과에서는 자질의 수를 증가시키기에 따라 급격한 성능의 향상을 보이다가 자질 7,000개 이상에서는 거의 수렴하면서 완만한 상승을 보인다. 이러한 현상은 기본 시스템과 제한한 시스템에서 비슷하게 나타나며, 자질의 수에 상관없이 제한한 시스템이 기본 시스템에 비해 평균 1.4%의 성능향상이 있음을 볼 수 있다.

다음 실험은 문장 중요도를 구하기 위해 사용한 두 가지 방법이 각각 성능 향상에 어떻게 영향을 미치는지 알아보기 위한 실험이다. 이 실험은 자질 수의 증가에 따른 성능향상이 1차적으로 수렴되고 있는 7,000개의 자질을 사용했을 때를 기준으로 실험했으며, 실험 상수 k_1 과 k_2 를 0에서 5까지 변화시키면서 비교 실험하였다.



[그림 3] 실험 상수 변화에 따른 성능 비교

위의 두번째 실험 결과는 제목과의 유사도를 이용해 문장 중요도를 구하는 방법(Sim(S,T))과 문장에 출현한 자질의 중요도를 합하여 문장 중요도를 구하는 방법(Cen(S))을 모두 적용한 경우가 가장 성능이 우수함을 나타냈으며, 두 가지 방법 중 하나만을 사용한 경우도 기본 시스템에 비해 좋은 성능을 보였다. 특이한 것은 두 가지 방법 중 하나만을 사용한 경우는 k_1 과 k_2 를 각각 증가시켰을 때 1.0 이상에서 약간의 불규칙적인 상승과 하락을 보이면서 불안정한 수렴을 하는 반면, 두 가지 방법을 같이 적용한 경우에는 2.5까지 꾸준한 성능 향상을 보이다가 안정적인 수렴을 하는 양상을 보인다. 이 실험의 결과를 통해 제안한 두 가지 문장 중요도 계산 방법 중 어느 방법을 사용해도 성능의 향상을 기대할 수 있으며, 두 가지 방법을 모두 사용하면 서로 보완을 통해 상승 효과(synergy)를 일으키며 추가적인 성능 향상이 일어난다는 것을 알 수 있다.

4.4 결과 분석

이번 절에서는 제안한 방법의 타당성을 확인하기 위하여 두 가지 예제를 통해 실험결과를 분석하고자 한다.

제목: 디아하다가 다운이 되요.

본문: 1) 2를 구해서 하고 있는데 배틀넷 들어갔다 나오면서 자꾸 다운이 되네요.

2) 우리집 그래픽카드가 세비지4인데..

3)이것 때문인지도 의심이가고 제이스텍에서 나온건데..

4) 보드는 유니텍BX보드이고요..

5) 아시는분 도와주세요.

위 예제는 실험문서 중 게임(han.rec.games)의 범주에 해당 하는 문서이다. 위의 문서는 한눈에 "디아"와 "배틀넷"이라는 단어를 보고 게임과 관련된 문서라는 사실을 알 수 있지만 실제로 기존 시스템은 "그래픽카드", "보드"등 시스템 관련 단어들 이 더 많이 나오므로 오히려 시스템(han.comp.sys)의 범주로 오인하는 결과를 보였다. 그러나 제안한 시스템에서는 문장 중요도 계산을 통해 "다운"이라는 단어를 제목과 공유하고, 중요도가 높은 단어인 "배틀넷"이라는 단어를 포함한 문장 1)이 가장 중요한 문장으로 인식되며, 따라서 "배틀넷"에 가중치가 월등하게 많이 적용되어 올바르게 게임 범주로 분류하는 결과를 볼 수 있다.

제목: 라쇼몽을 기억하며

본문: 1) 1951년 베니스 영화제 그랑프리를 수상하며 아시아 영화의 위상을 세계에 알린 구로사와 아키라 감독의 작품 라쇼몽을 기억하십니까?

2) 진실을 보고 싶어도 볼 수가 없지요.

3) 저는 다만 그 조선.중앙.동아 삼총사에 정부와 기타 잡다 언론들 상당수를 같이 끌어 넣어야 한다는 생각입니다.

위 예제는 정치(han.politics)의 범주에 속하는 문서이지만, "라쇼몽"이라는 영화에 빗대어 정치적 소견을 이야기하고 있다. 실제로 제목과의 유사도만을 이용하여 문장 중요도를 구한 경우 "라쇼몽"과 "기억"이라는 내용을 포함한 문장 1)이 가장 중요한 문장이 되고, 이 문장은 영화관련 자질을 많이 포함하고 있으므로 이 문서는 영화(han.rec.movie) 범주로 결정된다. 그러나 자질 중요도를 이용하여 문장 중요도를 구하는 방법을 적용

하면 "영화제", "영화", "감독" 등의 자질이 출현한 문장 1)보다 "조선", "중앙", "동아", "정부", "언론" 등의 자질을 포함한 문장 3)의 중요도를 더 높여주게 된다. 본 실험에서는 자질 중요도를 이용한 방법만을 사용할 경우나, 두 가지 방법을 모두 적용할 경우 올바르게 정치 범주로 결정해주고 있다. 따라서 이 예제를 통해 자질 중요도를 이용하는 방법이 제목과의 유사도를 이용한 방법에서 발생하는 오류를 보정해 주는 것을 확인할 수 있다.

5. 결론 및 향후 연구

본 논문에서는 문서 요약 기법을 적용한 자동 문서 분류 시스템을 제안했다. 제목을 이용하여 제목과의 문장 유사도를 통해 문장 중요도를 측정하는 방법과 문장에 출현한 자질들의 중요도를 더하여 문장 중요도를 구하는 방법을 이용하여 문서 전체의 문장 중요도를 구하였다. 이렇게 구해진 문장 중요도를 색인 과정에서 각 자질의 빈도에 차등 적용하였다.

제안한 방법은 뉴스 그룹 데이터에서 실험되었으며, 단순히 문서 전체에 출현한 단어의 빈도수를 이용하여 문서를 표현했을 때보다 더 좋은 성능을 보였다.

본 논문에서는 단순 베이지언 확률 모델을 이용하였으나, 현재 가장 좋은 성능을 보이고 있는 최근린법과 지지 벡터 기계 방법 등 다른 문서 분류 방법에서의 적용과 기사(article), 웹 문서 등 다양한 영역에 적용하는 연구가 필요하다.

6. 참고 문헌

- [1] Yang, Y. and Xin Liu.(1999) "A re-examination of text categorization methods." In Proc. of Conference on Research and Development in Information Retrieval (ACM SIGIR-99).
- [2] Yang, Y.(1999) "An evaluation of statistical approaches to text categorization." Information Retrieval Journal.
- [3] Yang, Y., Pedersen, J.O.(1997) "A Comparative Study on Feature Selection in Text Categorization", Proc. of The 14th International Conference on Machine Learning (ICML'97), pp. 412-420.
- [4] G. Salton, E. A. Fox and H. Wu(1983) "Extended boolean information retrieval." Communications of the ACM 26(12), 1022-1036.
- [5] Youngjoong Ko, Jungyun Seo (2000) "Automatic Text Categorization by Unsupervised Learning", Proc. of The 18th International Conference on Computational Linguistics (COLING 2000), pp453-459.
- [6] Masaki Murata, Qing Ma, Kiyotaka Uchimoto, Hiromi Ozaku, Hitoshi Isahara, and Masao Utiyama (2000) "Information retrieval using location and category information." Journal of the Association for Natural Language Processing,7(2)
- [7] Mock, K. J. (1996) "Hybrid hill-climbing and knowledge-based techniques for intelligent news filtering." In Proc. of The National Conference on Artificial Intelligence (AAAI'96).
- [8] J. Goldstein, M. Kantrowitz, V. O. Mittal, and J. Carbonell.(1999) "Summarizing Text Documents: Sentence Selection and Evaluation Metrics." In Proc. of SIGIR'99.
- [9] Radev, D. R, Jing, H., and Stys-Budzikowska, M.(2000) "Summarization of multiple documents: clustering, sentence extraction, and evaluation." Proc. of ANLP-NAACL Workshop on Automatic Summarization.
- [10] Daniel Marcu.(1999) "Discourse trees are good indicators of importance in text." Advances in Automatic Text Summarization. pp.123-136 The MIT Press.
- [11] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., & Slattery, S (1999) "Learning to Construct Knowledge Bases from the World Wide Web." to Artificial Intelligence.
- [12] Li H. and Yamanishi K.(1997) "Document Classification Using a Finite Mixture Model." The Association for Computational Linguistics(ACL'97)