정보관리연구, vol. 33, no. 2, 2002, pp. 19-32

문서관리를 위한 자동문서범주화에 대한 이론 및 기법

An Automatic Text Categorization Theories and Techniques for Text Management

고 영 중*· 서 정 연**

Youngjoong Ko· Jungyun Seo

차 례

초 록

최근 디지털 도서관이 등장하고 인터넷이 폭 넓게 보급되어 온라인 상에서 얻을 수 있는 텍스트 정보의 양이 급증함에 따라 효율적인 정보 관리 및 검색이 요구되고 있다. 자동 문서 범주화란 문서의 내용에 기반하여 미리 정의되어 있는 범주에 문서를 자동으로 할당하는 작업으로써 효율적인 정보 관리 및 검색을 가능하게 하는 동시에 방대한 양의 수작업을 감소시키는데 그 목적이 있다. 문서 분류를 위해서는 문서들을 가장 잘 표현할 수 있는 자질들을 정하고, 이러한 자질들을 통해 분류할 문서를 색인 과정을 통해 표현한다. 또한, 문서 분류기를 통해 문서를 목적에 맞게 분류한다. 본 논문에서는 자동 문서 범주화를 수행하기 위한 각 단계를 소개하고 각 수행 단계에서 사용되는 여러 가지 기법들을 소개하고자 한다.

키 워 드

문서 범주화, 자질 선택, 색인, 문서 분류기

* 서강대학교 컴퓨터학과 박사과정
(Department of Computer Science, Sogang University, kyj@nlpzodiac.sogang.ac.kr)
** 서강대학교 컴퓨터학과 교수
(Professor, Department of Computer Science, Sogang University, seojy@ccs.sogang.ac.kr)

ABSTRACT

With the growth of the digital library and the use of Internet, the amount of online text information has increased rapidly. The need for efficient data management and retrieval techniques has also become greater. An automatic text categorization system assigns text documents to predefined categories. The system allows to reduce the manual labor for text categorization. In order to classify text documents, the good features from the documents should be selected and the documents are indexed with the features. In this paper, each steps of text categorization and several techniques used in each step are introduced.

KEYWORDS

Text Categorization, Feature Selection, Indexing, Text Classifier

1. 서 론

최근에는 전통적인 종이 매체인 신문이나 잡지와 같은 미디어로부터 인터넷과 디지털 도서관과 같은 전자 매체까지 다양한 경로에서 정보를 습득할 수 있게 되었다. 특히, 인터넷의 확산과 디지털 도서관의 등장을 통해 이러한 여러 형태의 정보를 하나로 통합하여 사용자에게 제공함으로써 보다 쉽고 편리하게 정보를 얻고 활용하는 단계에 이르렀다. 이와 같이 온라인상에서 얻을 수 있는 텍스트 정보의 양이 급증함에 따라 효율적인 정보 관리 및 검색이 요구되고 있다.

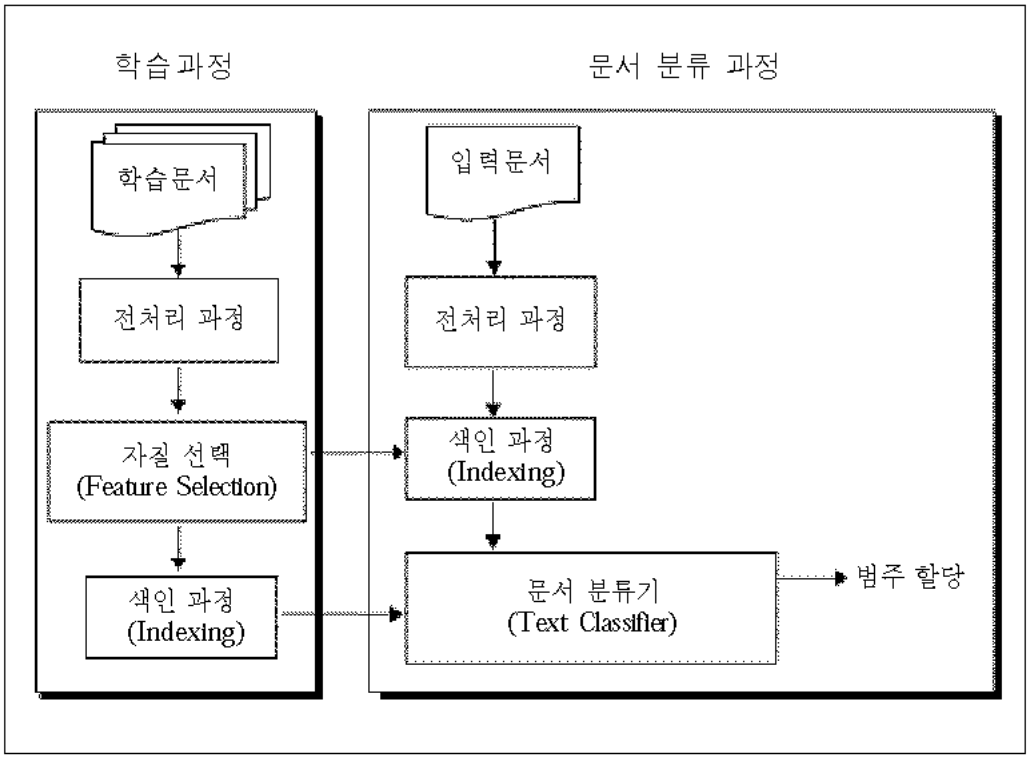
자동 문서 범주화(automatic text categorization)는 미리 정의된 범주에 문서를 자동으로 할당하는 기법과 관련된 연구 분야이다. 이전에는 수작업으로 문서마다 범주를 지정해 주는 방식이 사용되었으나 이 경우에는 사람의 노력, 시간, 비용 면에서 심각한 어려움을 초래할 수 있다. 이러한 작업을 자동 분류 시스템으로 교체하거나 보조 시스템으로 활용하면 비용을 크게 줄일 수 있을 것이다. 그러므로, 자동 문서 범주화는 대량의 문서를 효율적으로 관리하고 검색할 수 있게 하는 동시에 방대한 양의 수작업을 감소시키는 데 그 목적이 있다(Yang and Pederson 1997).

2. 자동 문서 범주화

일반적으로 지도 학습 (supervised learning)을 기반으로 한 문서 범주화 시스템은 학습 과정을 필요로 하기 때문에 <그림 1>과 같이 학습 과정과 문서 분류 과정으로 나뉜다.

문서 범주화 시스템의 각 과정을 구성하기 위한 처리 단계는 다음과 같다.

· 전처리 단계 : 수집된 문서를 기계적



<그림 1> 문서 범주화 시스템의 전체 구성도

처리가 가능하도록 변환하고 문서의 내용이나 특징을 잘 반영하는 내용어(content word)를 추출하는 단계

- 자질선택 단계 : 앞의 전처리 단계를 통해 학습 문서에 나타나는 내용어 중에서 범주화 학습에 유용하게 학습될 만한 내용어만을 자질(feature)로서 선택하는 단계
- 문서색인 단계 : 선택된 자질을 통해 어떻게 문서를 표현할 것인가에 대한 색인 단계
- 문서분류 단계 : 기계 학습에 사용되는 알고리즘을 사용하여 입력된 문서를 정해진 범주로 분류하는 단계

이후의 장에서는 위에서 소개한 문서 범주화 시스템의 4가지 처리 단계에 대해서 자세히 알아본다.

2.1 전처리 단계

전처리 단계는 다음과 같이 문서 정규화와 내용어 추출로 나누어 질 수 있다.

1) 문서 정규화

수집된 문서를 자동 문서 범주화 시스템에서 사용하기 위해서는 우선 기계적 처리가 가능하도록 변환하여야 한다. 이를 위해서 먼저 문서에서 특수 문자 등을 제거

해야 하는데, 웹 문서인 경우에는 HTML 문서에 포함되어 있는 태그와 특수 문자를 제거하고 한자어는 그에 해당하는 한글로 변환한다. 그리고, 형태소 분석기를 사용하기 위하여 문서의 내용을 문장 단위로 분할한다.

2) 내용어 추출

문서의 내용이나 특징을 잘 반영하는 단어를 내용어(content word)라고 한다. 이러한 내용어를 추출하기 위해서는 먼저 형태소 분석기를 사용하여 문장을 각 형태소별로 나누어 품사를 결정한다. 명사는 개념을 도입하고 설명하는데 쓰이므로 내용어로 가장 많이 등장하는 중요한 품사로서 특히, 한국어에서는 동작성 명사에‘ ~하다’,‘ ~되다’ 등의 동사 파생 접미사가 붙어서 동사가 되는 경우가 많으므로 명사의 비중이 그만큼 크다고 할 수 있다. 이렇게 추출된 내용어 중에는 여러 문서에서 공통적으로 많이 나타나기 때문에 별다른 정보를 주지 못하는 불용어(stop word)들이 있다. 예를 들어, 영어에서는 주로 관사(a, the), 접속사(that 등), 대명사(it 등), 그리고 be 동사(is, are 등) 등이 이들에 속한다. 이들 불용어를 처리하기 위해 불용어 사전을 정의하고 내용어 추출 시 불용어에 해당하는 용어들을 제거한다.

2.2 자질 선택 단계

자질 선택은 전처리 단계에서 문서에 나타나는 내용어들 중에서 범주화 구분에 유용하게 사용될 만한 내용어를 선택하는 작업이다. 학습 문서에 나타나는 내용어의 수는 수만에서 수십만에 이르기 때문에 모든 내용어가 자질로 선택된다면 학습 및 분류 시간이 매우 오래 걸리게 된다. 그러므로, 문서 범주화 성능의 저하 없이 자질의 수를 줄이기 위하여 학습 문서에 나타나는 내용어의 정보량을 계산하고 정보량이 큰 내용어만을 자질로 선택하려는 연구가 활발히 진행되어 왔다.

본 장에서는 자질 선택 기법 중에 대표적인 방법을 간략히 소개한다. 이들 기법 중에 정보 획득량과 카이 제곱 통계량이 가장 좋은 성능을 보이는 것으로 평가되고 있다(Yang and Pederson 1997; Wiener, Pderson and Weigend 1995; Church and Hanks 1989; Vapnik 1996).

1) 문서 빈도(Document Frequency)

문서 빈도란 어떤 용어가 나타난 문서의 빈도를 말하는 것으로, 학습 문서에서 그 용어가 나타나는 문서의 빈도를 계산한 후에 일정 빈도 이상의 용어만을 자질로 선택하는 기법을 말한다. 이 기법의 기본 가정은 출현 문서 빈도가 적은 용어는 문서 범주화의 성능에 기여하지 못한다는 것이다. 이 기법은 가장 간단하고 계산량이 적은 장점이 있으나“ 적은 빈도의 문서 빈도를 갖는 용어가 정보량이 많다”라는 정보 검색에서 널리 받아들여지는 기본 가정에 대치되어 잘 사용되지는 않는다(Yang and Pederson 1997).

2) 상호 정보 척도(Mutual Information)

상호 정보 척도는 통계적 언어 모델

(statistical language model)에서 일반적으로 사용되는 기법이다(Yang and Pederson 1997; Mitchell 1996). 용어 t의 범주 c에서의 정보량은 다음의 식으로 나타낸다. 즉, 범주 c에서 용어 t가 많이 출현할 수록 범주 c에서의 용어 t의 정보량은 크다.

$$I(t,c) = \log \frac{\Pr(t \wedge c)}{\Pr(t) \times \Pr(c)} \quad (1)$$

$$I(t,c) \approx \log \frac{A \times N}{(A+C) \times (A+B)} \quad (2)$$

식(2)는 식(1)의 근사값을 계산하기 위해 사용하는 식으로서 A는 범주 c에 속해 있는 문서 중 용어 t를 포함하는 문서의 수이고, B는 범주 c외의 범주에 속해 있는 문서에서 용어 t가 출현할 빈도이며, C는 범주 c에서 용어 t를 포함하지 않는 문서의 수이다. 그리고, N은 전체 학습 문서의 수이다. $I(t,c)$는 용어 t와 범주 c가 완전히 독립적이면 0의 값을 가진다. 각 범주에 대한 용어의 정보량을 계산한 후 전체 학습 문서에서의 용어 정보량을 계산하기 위해서 다음의 식 중에서 하나를 선택하여 사용한다.

$$I_{avg}(t) = \sum_{i=1}^{m} \Pr(c_i) I(t,c_i)$$
$$I_{max}(t) = \max_{i} \{I(t,c_i)\} \quad (3)$$

상호 정보 척도의 약점은 식(4)에서 알 수 있듯이 같은 조건부 확률값($Pr(t|c)$)을 갖는 단어라도 전체 출현 빈도(Pr(t))가 적은 용어의 상호 정보량이 상대적으로 더 높게 나온다는 것이다.

$$I(t,c) = \log \Pr(t|c) - \log \Pr(t) \quad (4)$$

3) 정보 획득량(Information Gain)

정보 획득량은 기계 학습 분야에서 자주 사용되는 기법이다(Mitchell 1996). 이 기법의 특징은 문서에서의 출현 빈도뿐만 아니라 출현하지 않은 빈도까지 고려해서 각 범주에서의 용어 정보량을 계산한다는 것이다. $\{c_1, c_2, \ldots, c_m\}$를 범주 집합이라 할 때 용어 t의 정보 획득량은 다음과 같은 식으로 구해진다.

$$G(t) = -\sum_{i=1}^{m} \Pr(c_i) \log \Pr(c_i) + \Pr(t) \sum_{i=1}^{m} \Pr(c_i|t) \log \Pr(c_i|t) + \Pr(\bar{t}) \sum_{i=1}^{m} \Pr(c_i|\bar{t}) \log \Pr(c_i|\bar{t}) \quad (5)$$

식(5)에서의 정보 획득량은 모든 범주의 평균값으로 계산된다. 학습 문서가 주어지면 학습 문서에서 나타나는 모든 용어들의 정보 획득량을 계산하여 일정 임계값 이상의 값을 갖는 용어들만을 자질로 선택하게 된다. 위에서 설명한 상호 정보 척도와 비교를 위해 식(5)를 변환하면 다음과 같다.

$$M(t,c) \quad (6)$$

정보 획득량은 상호 정보 척도와는 달리, 용어의 출현 빈도를 고려한 상호 정보 척도의 평균값과 용어가 출현하지 않은 빈도의 상호 정보 척도의 평균값의 합으로 계산된다. 이러한 특징으로 인해 정보 획득

량은 상호 정보 척도보다 문서 범주화 기법에서 더 좋은 성능을 보인다.

4) 카이 제곱 통계량(χ^2 statistics)

카이 제곱 통계량은 용어 t와 범주 c와의 의존성(dependency)을 측정하는 것으로서 자유도 1인 카이 제곱 분포와 비교될 수 있다. 카이 제곱 통계량을 계산하는 식은 식(7)과 같다.

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (7)$$

위 식(7)에서 A는 범주 c에 속해 있는 문서 중에서 용어 t를 포함하고 있는 문서의 수이고, B는 범주 c의의 범주에 속해 있는 문서 중에서 용어 t를 포함하고 있는 문서의 수이다. 또한, C는 범주 c에 속해 있는 문서 중에서 용어 t를 포함하지 않는 문서의 수이며, D는 범주 c의의 범주에 속해 있는 문서 중에서 용어 t를 가지고 있지 않은 문서의 수이다. 그리고, N은 전체 학습 문서의 수이다.

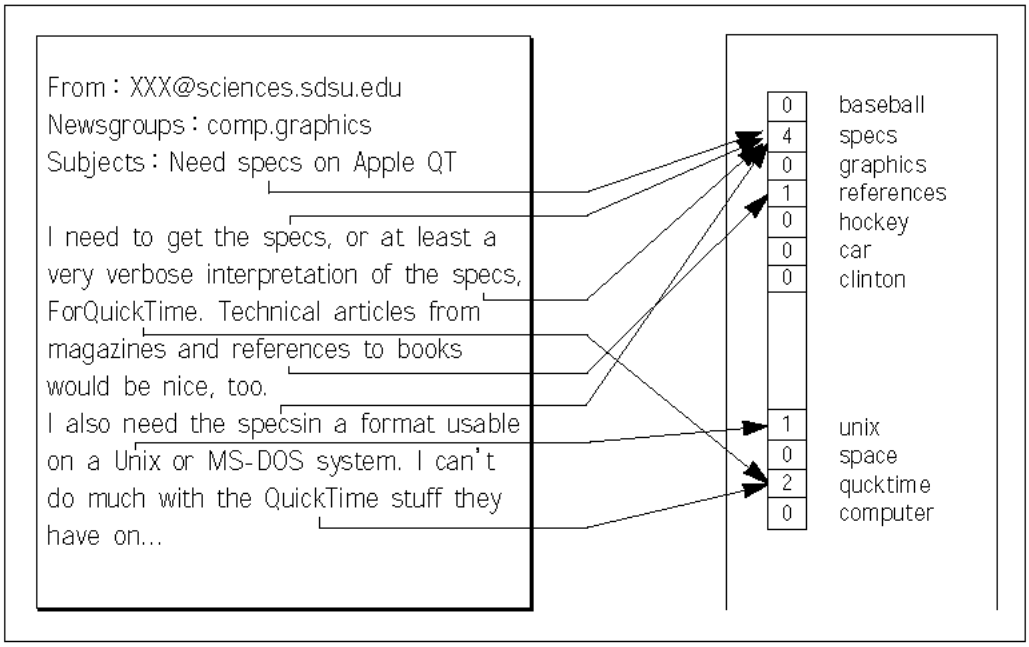
카이 제곱 통계량은 용어 t와 범주 c가 완전히 독립적이면 0의 값을 가진다. 각 범주에 대해서 용어의 정보량을 계산한 후에 전체 학습 문서에서의 용어 정보량을 계산하기 위해서 다음의 식 중에 하나를 선택하여 사용한다.

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} \Pr(c_i)\chi^2(t,c_i)$$
$$\chi^2_{max}(t) = \max_{i=1}^{m}\{\chi^2(t,c_i)\} \quad (8)$$

2.3 색인 단계: 문서 표현

색인이란 선택된 자질을 사용하여 어떻게 문서를 표현할 것인가에 대한 단계이다. 문서의 표현은 문서 범주 시스템의 전체적인 일반화 성능에 큰 영향을 미치므로 각 문서를 학습에 적합한 형태로 표현해야 한다. 자질 선택 단계에서 선택된 자질을 색인 용어로 사용하기 위해서 문서에서 단어의 순서는 큰 문제를 일으키지 않는다는 가정을 한다. 그러면, 문서는 더 이상 순서(sequence)로 표현하는 객체가 아니라 단어 주머니(bag-of-words)형태로 표현된다. 이러한 표현법은 기계 학습 분야에서 사용되는 <자질:값> 표현법과 동일하다. 별개의 단어는 자질이 되고 문서 내에서의 빈도수 등을 이용한 단어 가중치(term weight)가 값이 되는 것이다(Frakes and Yates 1997). 다음의 <그림 2>는 단어 가중치로 단순하게 단어 출현 빈도(TF: Term Frequency)를 사용했을 경우의 어떤 문서에 대한 <자질:값> 표현법의 예를 나타낸다.

가장 일반적으로 사용되는 문서 표현 방법은 이른바 벡터 공간 모델이다. 이것은 문서 전체에 나타난 각 자질의 출현 빈도(TF)를 이용하여 문서를 하나의 벡터로 표현하는 것으로, 보통 자질의 출현 빈도와 역문헌빈도(IDF: Inverse Document Frequency) 혹은 역범주빈도(ICF: Inverse Document Frequency)를 이용하여 가중치를 줌으로써 문서를 표현한다(Salton, Fox and Wu 1983; 조광제, 김준태 1997; Youngjoong Ko and Jungyun Seo 2000). 문서 d_j를 표현하는 문서 벡터 $\vec{d_j}$를 식으로



<그림 2> 문서에 대한 자질 값의 표현

나타내면 다음과 같다.

$$\vec{d}_j = (d_{j1}, d_{j2}, \ldots, d_{jk}) \qquad (9)$$

식(9)에서 d_{jk}는 문서 d_j에서 자질 k의 가중치이다. 문서에서 각 자질의 가중치는 해당 자질의 빈도와 역문헌빈도, 또는 역범주빈도 등을 이용하여 다음과 같이 다양한 방법으로 결정된다.

1) 이진 가중치(Boolean Weighting)

가장 단순한 방법으로, 만약 문서에 해당 자질이 출현하면 1로 표현하고 그렇지 않다면 0으로 표현하는 방법이다. j번째 문서에서 자질 k의 가중치는 다음의 식으로 나타낸다.

$$d_{jk} = \begin{cases} 1 & \text{if } f_{jk} > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (10)$$

여기서 f_{jk}는 문서 d_j에서 출현한 자질 k의 빈도이다. 이 방법은 다음에 설명하는 방법들에 비해 계산량이 적은 장점이 있으며, 영역에 따라서는 다음에 설명하는 좀 더 복잡한 방법들보다 더 좋은 성능을 나타낸다는 연구 결과도 있다(Dumais et al. 1998).

2) 단어 빈도 가중치(Word Frequency Weighting)

단어 빈도 가중치 방법은 문서에 출현한 자질의 빈도를 그대로 문서 표현에 적용하는 방법이다. j번째 문서에서 자질 k의 가중치는 다음의 식으로 나타낸다.

$$w_{ik} = f_{ik} \qquad (11)$$

확률 모델의 경우 문서에서 해당 단어가 출현할 확률을 문서에서 출현한 단어의 빈도로 나타내는 경우가 많다. 확률 모델을 사용하는 단순 베이지언 확률 모델은 일반적으로 이 방법을 이용하여 문서를 표현한다.

3) TF-IDF 가중치(TF-IDF Weighting)

TF-IDF 가중치 방법은 가장 많이 알려진 방법 중의 하나이다. 문서에서 각 자질의 가중치는 해당 문서에서 각 자질의 빈도와 역문헌빈도(IDF)의 곱으로 나타낸다 (Salton and McGill 1983). i 번째 문서에서 자질 k의 가중치는 다음의 식으로 나타낸다.

$$w_{ik} = f_{ik} \times \log\left(\frac{N}{n_k}\right) \qquad (12)$$

여기서 N은 전체 문서의 수이며, n_k는 자질 k가 출현한 문서의 수이다.

TF-IDF 가중치 방법은 여러 가지 변형이 존재한다. 그 중 다음과 같은 대표적인 두가지 방법을 소개한다. 문서에 출현하는 자질의 빈도는 문서의 길이에 따라 영향을 받는다. 즉 길이가 긴 문서에서는 길이가 짧은 문서에서보다 자질의 빈도가 높아진다. 식(13)의 가중치 방법은 TF-IDF 가중치 방법으로 구해진 각 자질의 가중치를 문서의 크기로 정규화 하는 방법이다 (Salton and Buckley 1998). i번째 문서에서 자질 k의 가중치는 다음의 식으로 나타낸다.

$$w_{ik} = \frac{f_{ik} \cdot \log\left(\frac{N}{n_k}\right)}{\sqrt{\sum_{k=1}^{K} \left| f_{ik} \cdot \log\left(\frac{N}{n_k}\right) \right|^2}} \qquad (13)$$

여기서 K는 문서를 표현하는 데 사용되는 전체 자질의 수이다.

다음 방법은 식(13)의 가중치 방법의 변형으로, 로그를 이용하여 자질들간의 지나친 빈도 차이를 줄여주기 위한 것이다 (Buckley et al. 1994). i번째 문서에서 자질 k의 가중치는 다음의 식으로 나타낸다.

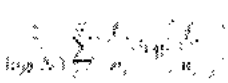
$$w_{ik} = \frac{(\log f_{ik} + 1.0) \log\frac{N}{n_k}}{\sqrt{\sum_{k=1}^{K} \left[(\log f_{ik} + 1.0) \log\frac{N}{n_k}\right]^2}} \qquad (14)$$

TF-IDF 가중치 방법은 정보 검색 분야에서 가장 기본적으로 사용하는 방법으로서 문서 분류기 중 k-최근린법, Rocchio, 그리고 지지 벡터 기계 모델에 이 방법을 적용하여 문서를 표현한다.

4) 엔트로피 가중치(Entropy Weighting)

엔트로피 가중치 방법은 가장 세련된 가중치 방법 중 하나로, 단어 빈도 가중치 방법에 비해 약 40%의 성능향상을 보였다는 연구 결과도 있다(Dumais 1991). i번째 문서에서 자질 k의 가중치는 다음의 식으로 나타낸다.

$$w_{ik} = \log(f_{ik} + 1.0) \cdot \left(1 - \frac{1}{\log(N)} \sum_{j=1}^{N} \frac{f_{ij}}{n_k} \log\frac{f_{ij}}{n_k}\right) \qquad (15)$$

여기서  는 자질 k의 평균 불확실성 또는 엔트로피라고 하며, 이 값은 만약 자질이 모든 문서에서 동일한 분포를 갖는다면 -1 값을 가지고, 단지 하나의 문서에서만 나타난다면 0의 값을 갖게 된다.

5) 역범주 빈도 가중치

역문헌빈도가 문서간의 분리도가 높은 자질에 높은 가중치를 주는 것이라면, 역범주빈도는 범주간의 분리도가 높은 자질에 높은 가중치를 주는 방법이다. 즉, 소수의 범주에 많이 나온 자질에 대해서는 높은 가중치를 주고, 여러 범주에 고르게 나오는 자질에 대해서는 낮은 가중치를 주는 방법이다(조광제, 김준태 1997; Youngjoong Ko, Jungyun Seo 2000). i번째 문서에서 자질 k의 가중치는 다음의 식으로 나타낸다.

$$w_{ik} = f_{ik} \cdot \log\left(\frac{C}{c_k} - 1\right) \quad (16)$$

여기서 C는 전체 범주의 수이며, c_k는 자질 k가 출현한 범주의 수이다.

2.4 문서 분류 단계: 문서 범주화 모델

많은 양의 문서를 관리하고 이를 효율적으로 검색하기 위한 문서 분류 모델에는 기계 학습 분야에서 사용되는 알고리즘들이 사용되는데, 크게 규칙 기반 모델(Rulebased Model)과 연역적 학습 모델(Inductive Learning Model), 그리고 검색을 활용한 모델로 나뉘어진다. 먼저 규칙 기반 모델은 학습 문서들에서 나타나는 범주간의 구별된 규칙을 전문가가 찾아주거나 학습을 통해 추출된 규칙을 이용하여 문서를 분류하는 모델이다(Apte, Damerau and Weis 1994). 연역적 학습 모델은 학습 문서에서 자질을 추출하여 이를 확률적인 접근 방법으로 사용한 베이지언 확률 모델(Lewis 1998; McCallum and Nigram 1998)과 트리 구조로 표현하여 자질의 유무로 범주를 결정하는 결정 트리 모델(Lewis and Ringuette 1994), 학습 문서를 통해 생성된 양성 자질(positive feature)과 음성 자질(negative feature)을 벡터 공간으로 표현하고 이들의 차이를 극명하게 하는 지지 벡터(support vector)를 찾는 지지 벡터 기계(Cortes and Vapnik 1995; Joachims 1998) 등이 있다. 또한, 정보 검색 관점에서 분류할 대상 문서를 질의로 보고 이와 유사한 문서를 찾는 방법인 k-최근린법(Yang et al. 2002)과 선형 분리기(Lewis et al. 1996) 등이 있다. 이들 중에서 지지 벡터 기계와 k-최근린법이 가장 좋은 성능을 보이는 것으로 보고되고 있다(Yang and Liu 19991). 본 장에서는 이중 가장 대표적인 네 가지를 간략히 설명한다.

1) 베이지언 확률 모델(Bayesian Probability Model)

베이지언 확률 모델은 주로 기계 학습(machine learning)에서 연구되어 왔다(Mitchell 1996). 이 모델은 주어진 입력 문

서의 각 범주에 할당될 확률을 구하기 위해서 문장에 속해 있는 용어들과 범주와의 결합 확률값을 사용하는 방법이다(Young-joon Ko and Jungyun Seo 2000; Lewis 1998). 구하고자 하는 확률은 식(17)로 나타낸다.

$$\Pr(c_k \mid d) \quad \Pr(c_j) \cdot \frac{\prod \Pr(x \mid c_j)}{\Pr(d)} \qquad (17)$$

여기서 d는 문서를 나타내고 c_j는 범주를 나타낸다. 이 식은 구하고자 하는 확률($\Pr(c \mid d)$)을 Bayes' rule을 사용하여 바꾼 것이다. 이 식을 계산 가능하게 하기 위하여 문서 d를 문서 안에 포함되어 있는 용어(x_i)의 집합($d = \{x_1, \ldots, x_{\ldots}, \ldots, x_n\}$)으로 표현하고 용어 간에 독립 발생을 가정하면 식(18)을 유도할 수 있다.

$$\Pr(c_k \mid d) \quad \Pr(c_j) \cdot \frac{\prod \Pr(x \mid c_j)}{\Pr(d)} \qquad (18)$$

식(18)에서 계산되는 확률 값을 문서가 각 범주에 할당될 확률로 보고 가장 확률이 높은 범주로 문서를 할당한다.

2) k-최근린법(k-Nearest Neighbor)

최근린법은 40년 동안 패턴 인식에서 연구가 이루어진 잘 알려진 예제 기반의 학습 방법으로서 문서 범주화 연구가 시작된 초기부터 적용되어 왔으며 가장 좋은 성능을 가진 분류기중의 하나이다(Dasarathy 1991). 최근린법의 기본 알고리즘은 매우 간단하다. 즉, 실험 문서가 주어졌을 때 학습 문서 중에서 테스트 문서와의 유사도가 가장 높은 k개의 문서를 추출하고 그들을 사용하여 각 후보 범주의 순위를 매기는 방법이다. k개의 추출된 학습 문서는 미리 정해진 범주가 있으므로, 각 범주의 테스트 문서와의 유사도는 각 범주별로 추출된 k개의 문서와 테스트 문서와의 유사도의 합으로 계산된다. 최근린법의 결정 규칙은 식(19)로 나타낼 수 있다.

$$y(x, c_j) = \sum_{d_i \in kNN} sim(x, d_i) y(d_i, c_j) \qquad (19)$$

여기서 $y(d_i, c_j)$는 문서 d_i가 범주 c_j에 속하는가에 해당되는 값을 갖는 함수이고, $(sim(x, d_i))$는 테스트 문서 x와 학습 문서 d_i와의 유사도 값으로 두 문서 벡터 간의 코사인 유사도 값을 사용한다.

k-최근린법은 모든 범주의 문서를 하나의 공간에 표시하면서 범주의 영역을 나누려는 노력을 할 필요가 없으므로, 선형 분리 문제로 고생하지 않으며, 지지 벡터 기계 모델과 더불어 문서 범주화 분야에서 가장 좋은 성능을 보이는 것으로 알려져 있다(Joachims 1998; Yand and Liu 1999). 그러나, 각 입력 문서에 대해서 모든 학습 문서를 비교해야 하기 때문에 수행 속도가 느리다는 치명적인 단점을 가지고 있으며, 이를 극복하기 위한 연구가 활발히 진행 중이다.

3) 지지 벡터 기계(SVM : Support Vector Machines)

지지 벡터 기계는 두개의 범주를 구분하는 문제를 해결하기 위해 1995년에

Vapnik에 의해 소개된 비교적 최근의 학습 기법으로 두 개의 클래스의 구성 데이터들을 가장 잘 분리할 수 있는 결정면(decision surface)을 찾는 모델이다(Vapnik 1995). 지지 벡터 기계에서의 결정면은 식(20)와 같이 나타낼 수 있다.

$$\vec{w} \cdot \vec{x} - b = 0 \qquad (20)$$

여기서 $\vec{x}$는 분류하고자 하는 문서의 벡터이며 $\vec{w}$와 b는 학습 데이터로부터 학습되어 나온 결과이다. 학습 문서 집합을 $D = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)\}$과 같이 나타냈을 때 각각의 학습 문서 벡터($\vec{x}_i$)가 임의의 범주에 속한 문서이면 y_i의 값에 +1을 할당하고, 범주에 속하지 않은 문서에는 -1을 할당한다. 결국 지지 벡터 기계는 식(21)을 만족시키는 $\vec{w}$와 b를 찾는 문제이다.

$$\begin{aligned} \vec{w} \cdot \vec{x}_i - b &\geq +1 \quad \text{for} \quad y_i = +1 \\ \vec{w} \cdot \vec{x}_i - b &\leq -1 \quad \text{for} \quad y_i = -1 \end{aligned} \qquad (21)$$

지지 벡터 기계는 직선으로 나눌 수 있는 문제에 사용되는 알고리즘이지만, 다차원의 부드러운 곡선을 이용하여 경계면을 설정하거나, 실제 데이터 벡터를 새로운 자질을 포함한 새로운 벡터 공간에 매핑하는 방법을 통해서 직선으로 나눌 수 없는 문제도 해결할 수 있다. Joachims는 최근에 지지 벡터 기계 모델을 문서 범주화에 적용하여 좋은 성능을 보였다(Joachims 1998).

4) 선형 분류 모델(Linear Classification Model)

선형 모델은 범주 c의 가중치 벡터 $\vec{W_c}$와 문서 자질 벡터 d의 내적(inner product) 값을 문서 d에 대한 범주 c의 할당 여부 결정에 사용하는 기법이다(Lewis et al. 1996). $\vec{W_c} = (w_1, \ldots, w_n)$를 $\vec{W} = (w_1, w_2, \ldots, w_n)$로, d를 $d = (x_1, x_2, \ldots, x_n)$로 나타낼 때 범주 c에 대한 문서 d의 선형 분류 함수식은 다음 식(22)와 같다.

$$f(d) = \vec{W} \cdot d = \sum_{i=1}^{k} w_i x_i \qquad (22)$$

식(22)는 일반적으로 정보 검색 모델에서의 검색 결과에 대한 순위 결정에 사용되나, 문서 범주화에서는 문서에 대한 범주의 할당 여부를 결정하기 위한 선형 분류기로 사용된다. 선형 분류기의 가중치 벡터를 학습시키기 위하여 여러 가지 알고리즘이 사용되는데, 전체 학습 문서에 대해 한 번의 계산으로 가중치 벡터를 생성해 내는 Rocchio 알고리즘과 개별 학습 문서들을 가중치 벡터 조정에 참여시키는 Widrow-Hoff 알고리즘이 대표적인 것이다(Lewis et al. 1996).

3. 결 론

본 논문에서는 문서를 효율적으로 관리할 수 있는 자동 문서 범주화 기법에 대해서 소개하고 자동 문서 범주화 시스템의 각 단계에서 사용할 수 있는 기술에 대해서 설명하였다. 물론 자동 문서 범주화 시

스템을 구축하고 사용하는 데는 적지 않은 노력과 시간이 필요하다. 특히, 문서 분류기를 구축하기 위해서는 기계 학습에 대한 전반적인 이해와 지식이 필요하다. 하지만, 영어권에서는 인터넷상에 온라인으로 오픈되어 있는 toolkit들이 있다. 이들 toolkit을 참조하면 조금 더 쉽게 문서 분류기를 구축할 수 있을 것이다. 이들을 간단히 소개하면 다음과 같다.

먼저 CMU(Carnegie Mellon University)에서 개발한 Rainbow toolkit은 베이지언 확률 분류기와 최근린법 분류기 그리고 선형 분류기 중 하나인 Rocchio 분류기가 포함되어 있다. 다음의 사이트에서 Rainbow toolkit을 얻을 수 있다.

· http://www-2.cs.cmu.edu/~mccallum/bow/rainbow/

문서 범주화 외에도 분류 문제 영역에서 가장 좋은 성능을 보이고 있는 지지 벡터 기계를 위한 toolkit인 SVMlight toolkit은 Joachims가 개발한 것으로 다음의 사이트에서 얻을 수 있다.

· http://svmlight.joachims.org/

참고 문헌

조광제, 김준태. 1997.“ 역카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동분류.”『한국정보과학회 봄 학술발표논문집(B)』, 507-510.

Buckley, C. G. Salton, J. Allan and A. Singhal. 1994.“ Automatic Query Expansion Using SMART: TREC 3” *Proceedings 3rd Text Retrieval Conference*,NIST.

Cortes, C. and V. Vapnik. 1995.“ Support vector networks.” *Machine Learning 20(3)*: 273-297.

Chidanand Apte, Fred Damerau, and Sholom M. Weis. 1994.“ Towards language independent automated learning of text categorization models.” *Proceeding of the 17th annual international ACM-SIGIR*

Church, Kenneth Ward and Patrick Hanks. 1989.“ Word association norms, mutual information and lexicagraphy,” *Proceedings of ACL* 27: 76-83, Cancouver, Canada .

Dasarathy, Belur V. 1991.“ Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.” *McGraw-Hill Computer Science Series* CA: IEEE Computer Society Press.

Dumais, S. T. 1991.“ Improving the retrieval information from external sources.” *Behaviour Research Methods, Instruments and Computers*, 23(2): 229-236.

Dumais, S. T. J. Platt, D. Heckerman and M. Sahami. 1998.“ Inductive learning algorithms and representations for text categorization.” *Proceedings of ACM-CIKM98*,Nov. 148-155.

Frakes, W. B. and R. B. Yates. 1997. *Information Retrieval Data Structures & Algorithms* Prentice-Hall.

Joachims, T. 1998. " Text Categorization with Support Vector Machines: Learning with Many Relevant Features". *European Conference on Machine Learning(ECML)*.

Ko, Youngjoong Jungyun Seo. 2000. " Automatic Text Categorization by Unsupervised Learning", *Proceedings of The 18th International Conference on Computational Linguistics (COLING 2000)*, 453-459

Lewis, David D. and Marc Ringuette. 1994. " A comparison of Two Learning Algorithms for Text categorization." *Proceeding of the 3rd Annual Symposium on Document Analysis and Information Retrieval*.

Lewis, David D. Robert E. Schapire, james P. Callan and Ron Papka. 1996. " Training Algorithms for Linear Text Classifiers." *Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR'96)*, 289-297.

Lewis, David D. 1998. " Naive (bayes) at forty: The independence assumption in information retrieval." *European Conference on Machine Learning*.

McCallum, Andrew and Kamal Nigram. 1998. " A comparison of Event Models for Naive Bayes Text Classification". *AAAI' 98 workshop on Learning for Text Categorization*.

Mitchell, Tom 1996. Machine Learning. McCraw Hill.

Salton, G. E. A. Fox and H. Wu. 1983 " Extended boolean information retrieval." *Communications of the ACM*, 26(12): 1022-1036.

Salton G. and M. J. McGill. 1983. *An Introduction to Modern Information Retrieval*, McGraw-Hill.

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

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Wiener, E. J. O. Pedersen, and A.S. Weigend. 1995. " A neural network approach to topic spotting." *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR' 95)*.

Yang, Y. and J. O. Pederson. 1997. " A comparative study on feature selection in text categorization." *Proceedings of the 14th International Conference on Machine Learning*

Yang, Y. and Xin Liu. 1999. " A re-examination of text categorization methods". *Proceedings of Conference on Research and Development in Information Retrieval (ACM SIGIR' 99)*.

Yang, Y. S. Slattery, and R. Ghani. 2002. "A study of approaches to hypertext categorization", *Journal of Intelligent Information Systems*, 18(2).