자질 중요도 계산 기법에 의한 자동 문서 범주화

이경찬°, 강승식
국민대학교 컴퓨터학부, 첨단정보기술연구센터
{ayin, sskang}@cs.kookmin.ac.kr

Automatic Document Categorization by the Importance of Features

Kyung-Chan Lee°, Seung-Shik Kang
School of Computer Science, Kookmin University and AITrc

요 약

문서 범주화를 위해 자질을 선별하는 기법으로는 자질의 출현 빈도에 따라 범주를 대표하는 자질들을 선별하는 것이 일반적이다. 출현 빈도에 의한 자질을 선별하는 통계적인 기법은 문서의 내용을 대표하는 용어들의 중요도를 간과하는 문제가 발생한다. 본 논문에서는 학습 문서 및 실험 문서에서 자질의 중요도에 의해 범주 대표어를 선별하는 문서 범주화 기법을 제안하였으며, 역범주 빈도 및 카이제곱 통계량에 의해 자질을 선별하는 방법과 비교-실험을 하였다. 문서 범주화 모델로는 나이브 베이지언 확률 모델을 이용하였으며, 성능 평가를 위해서 웹 디렉토리에서 수집된 데이터를 이용하여 실험하였다. 본 논문에서 제안한 자질 중요도에 의한 자질 선별 기법은 용어의 출현 빈도 및 카이제곱 통계량에 의해 자질을 선별한 방법보다 더 나은 성능을 보였다.

1. 서 론

정보화 시대의 대량으로 생겨나는 문서들에 대해서 검색과 관리의 효율성의 향상을 위해서 문서들에 대한 효과적인 관리가 필요해 졌다. 자동 문서 범주화(automatic document categorization)는 미리 정의된 두 개 이상의 범주에 대해서 입력된 문서의 범주를 자동으로 할당해 주는 작업으로써 수작업에 의한 문서 분류 비용을 줄이고 실시간으로 문서를 분류하기 위한 목적으로 사용된다.

문서 범주화에 관한 연구는 효과적인 범주화 모델의 계산 방법과 학습 자질의 추출 방법이라는 두 가지 문제를 중심으로 발전되어 왔다. 범주화 방법으로는 확률 모델, 기계학습 모델, 정보검색 모델 등을 이용하여 다양한 방법들이 사용되고 있으며, 범주 대표어로 사용되는 자질의 추출 방법은 학습 문서에서 형태소 분석기를 이용하여 명사들을 추출하고 자질의 출현 빈도를 이용하여 해당 자질에 대한 가중치를 계산하는 것이 일반적이다. 단순 출현 빈도에 의한 가중치 계산 기법은 문서를 대표하는 용어를 선별하는데 부족한 점이 있다. 따라서 역문헌 빈도(IDF) 혹은 역범주 빈도(ICF)를 빈도에 이용하여 가중치를 계산하는 방법이 사용되고 있다. 이 방법으로 어느 정도 성능의 향상을 보였으나, 문서 자체를 분류하는 문서 범주화 시스템에서 문서 구문간의 특징을 이용하지 못한 빈도 자체는 성능 향상에 한계가 있다. 본 논문에서는 이러한 자질의 가중치 계산에 있어서 기존의 빈도만을 이용하는 방법과는 달리 문서내의 구문론적 특성과 기타 어절의 위치 정보 등을 고려한 값을 이용하여 자질의 가중치 값을 달리하여 문서 범주화 방법에 적용하였다.

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

본 논문의 구성은 다음과 같다. 2장에서는 자질 선정과 범주화 방법에 대한 관련 연구에 대해 살펴보고, 3장에서는 문서 범주화 시스템 구조와 자질 중요도 값에 대한 내용에 대해 설명한다. 그리고 4장에서는 실험 및 평가를 하고, 5장에서는 결론 및 향후 연구과제에 대해서 기술한다.

2. 관련 연구

문서 범주화는 범주를 대표하는 자질을 선별하고 각 자질들에 대해 가중치를 부여하여 입력 문서에 대해 미리 정의되어 있는 범주로 할당시키는 범주화 방법이다. 범주 내용을 대표하는 자질의 선별 및 자질들의 가중치 계산을 위해 단어의 출현 빈도와 더불어 역문헌 빈도 혹은 역범주 빈도를 사용하여 단순 출현 빈도에 의존하는 가중치 계산 기법을 보완하는 방법이 이용되고 있으며, 자질들의 개수를 줄이기 위하여 카이제곱 통계량(x^2 statistic), 상호 정보량(Mutual Information), 기대 상호 정보량(Expected MI), 정보 획득량(Information Gain) 등의 방법을 이용하여 범주에 대하여 중요한 자질을 선별하는 기법이 사용된다.

문서 범주화 모델로는 확률을 이용한 나이브 베이지언 분류법(Naive Bayesian Classifier), k-NN(k-Nearest Neighbor) 분류법[1,3], 기계 학습을 이용한 신경망(Neural Network), SVM(Supported Vector Machine) 분류법[1,4] 등이 있다.

3. 자질 중요도를 이용한 문서 범주화

3.1. 문서 범주화 기법

본 연구에서 사용한 문서 범주화 기법은 나이브 베이지언 확률 모델이다. 입력 문서가 범주에 속할 확률을 구해서 가장 확률이 높은 범주로 할당하는 방법으로 이

모델을 다음과 같은 수식으로 표현된다.

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \quad (1)$$

$$= P(c) \prod_{i=0}^{T} P(\omega_i|c)^{N(\omega_i|d)}$$

$P(\omega_i|c)$:범주에서 자질 ω_i가 나타날 확률

$N(\omega_i|d)$:입력문서에서 출현한 자질 ω_i의 개수

T : 총 자질의 개수

3.2. 문서 범주화 시스템의 구조

문서 범주화 시스템은 크게 두 가지 부분으로 나눌 수 있다. 범주화된 문서 혹은 문서들의 자질들을 선별하고 가중치를 계산하는 과정과 선별된 자질과 범주화되어질 입력 자질들을 문서 분류 과정을 통해 가장 유사한 범주로 할당하는 범주화 과정이다.

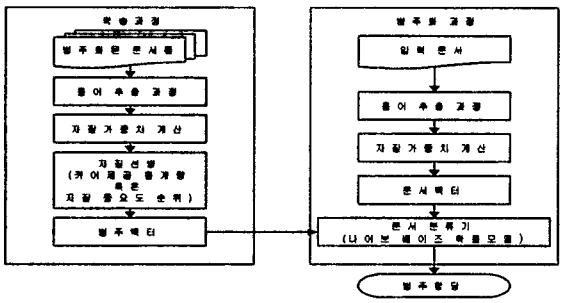

그림 1 . 문서 범주화 시스템

용어 추출 과정에서는 형태소 분석기를 이용 용어를 추출한다. 자질 가중치 계산 과정에서는 추출된 용어를 통해 일반적으로 빈도를 이용하지만 본 시스템에서는 빈도에 구문론적 특징을 반영한 가중치를 계산하였다. 가중치가 계산된 자질들에 대해서 카이제곱 통계량을 이용하여 자질을 선별하는 작업이 수행된다. 최종적으로 선정된 자질들을 벡터 공간 모델로 나타낸다. 범주화 과정에서는 입력 문서에 대한 과정을 학습 과정과 마찬가지로 수행한 후 문서 벡터를 만들게 되며, 분류기를 통해 문서를 해당 범주에 할당하게 된다.

3.3. 자질 중요도 계산

자질 중요도 계산 기법은 일반적으로 이용되는 문서내의 용어 빈도를 이용하는 것 이외에 품사 정보와 격 정보 등 어절 단위의 용어 특성과 문장을 단위로 하는 용어의 구문론적 기능, 문서 내에서 문장의 위치 등을 이용하여 용어에 가중치를 부여하는 방식이다[5].

3.4. 자질 선별 기법

자질 선별이란 자질들이 학습 시에 이용되어지는 자질들의 각 범주별 중요 용어를 얻는 방법이다. 자질 선별 방법으로 카이제곱 통계량 방법이 좋은 성능을 내는 것으로 알려져 있다[1]. 카이제곱 통계량은 중요 자질을 순위화하여 학습할 문서가 다량일 경우 벡터 차원을 줄일 수 있고, 성능 또한 우수한 자질 선별 방법으로 식은 다음과 같다.

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

A:범주c에 속해있는 문서중 용어 t를 포함하고 있는 문서수
B:범주c에 속하지 않은 문서중 용어 t를 포함하고 있는 문서수
C:범주c에 속해있는 문서중 용어 t를 포함하고 있지않은 문서수
D:범주c에 속하지 않은 문서중 용어 t를 포함하고 있지 않은 문서수
N:학습에 사용된 총 문서수

χ^2 값으로 선별된 자질들을 범주별 하나의 자질로 선별하는 작업이 필요하며 그 식은 다음과 같다.

$$\chi^2_{max}(t) = \max\{\chi^2(t, c_1), \chi^2(t, c_2), ..., \chi^2(t, c_n)\} \quad (3)$$

해당 식을 이용하여 자질에 대한 유일한 값을 중요도별로 순위화 할 수 있다. 실험에서 학습 문서에 대해 자질 중요도 값으로 순위화 선별 방법과 같이 비교하였다.

4. 실험 및 평가

4.1. 성능 평가 방법

자질 선정 및 가중치 계산 기법으로 기존의 출현 빈도에 의한 방법과 본 논문에서 제안한 자질 중요도 방법에 의해서 계산된 가중치 계산 방식을 비교해 보고 자질 선별에 있어서 기존 이용되던 카이제곱 통계량 방법과 본 논문에서 제안한 자질 중요도 값을 적용한 방법을 실험하였다. 본 실험에서는 입력된 문서는 하나의 범주를 갖는 단일 범주화 방식을 채택하였으며, 성능 평가를 위해 정확률(precision), 재현율(recall), F_1-measure를 계산하였다.

4.2. 실험 데이터

실험에 사용된 데이터는 고려대학교에서 웹 디렉토리를 이용하여 만들어진 데이터를 이용하였다. 문서수는 총 4,800문서로 이루어져 있으며, 이에 학습 문서로 3,200개, 테스트 문서로 1,600개를 이용하였다. 학습되는 범주 구조는 각 범주 당 3단계의 계층으로 이루어져 있으며, 최하위 단계에 총 80개의 범주로 구성되어 있다. 80 개의 범주들은 아래의 표와 같다. 각 범주 당 학습 문서의 개수는 40개이고, 테스트 문서는 20개씩으로 구성되어 있다.

표 1 . 범주표

범주 (최상위범주)	하위 범주개수	학습 문서	테스트문서	총계
컴퓨터,인터넷	10	400	200	600
사회생활	9	360	180	540
경제	8	320	160	480
문화	11	440	220	660
건강,의학	14	560	280	840
레크레이션, 스포츠	11	440	220	660
과학,공학	13	520	260	780
교육	4	160	80	240
총계	80	3200	1600	4800

4.3. 실험 결과

기존 자질 가중치를 나타내는 빈도에 의한 방법과 본 논문에서 제시한 자질 중요도를 이용한 방법을 비교하였으며, 실험 결과는 그림 2와 같다. 이 실험에서는 학습 문서에서 추출된 총 175,175개의 자질을 모두 사용하였다. 이 실험 결과에 의하면 나이브 베이지언 모델을 이용하여 모든 자질을 사용했을 때, 단순 빈도(tf-idf)보다는 본 논문에서 제안한 자질 중요도에 의한 기법의 성능이 정확률과 재현율, 그리고 F_1-measure 값의 비교에서 각각 5%, 2%, 5%정도의 성능 향상을 보이고 있음을 볼 수 있다.

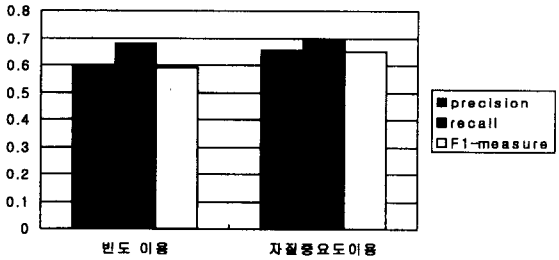

그림 2. 빈도와 자질 중요도에 의한 비교-실험

가중치 계산 및 자질 선별 기법을 달리 하여 (1) 출현 빈도(tf-idf)에 의한 가중치 계산 및 카이제곱 통계량에 의한 자질 선별 기법, (2) 자질 중요도에 의한 가중치 계산 및 카이제곱 통계량에 의한 자질 선별 기법 (3) 자질 중요도 및 IDF에 의한 자질 선별 기법을 적용한 실험 결과는 그림 3과 같다(표 2).

표 2 . 가중치 계산 및 자질 선별 기법의 실험 결과

방법 \ 자질비율		30%	50%	70%	100%
카이제곱& 빈도이용	precision	0.505	0.548	0.588	0.596
	recall	0.611	0.630	0.669	0.673
	F_1-measure	0.501	0.545	0.582	0.592
카이제곱& 자질 중요도 이용	precision	0.561	0.607	0.636	0.656
	recall	0.623	0.668	0.685	0.698
	F_1-measure	0.563	0.615	0.631	0.651
자질 중요도 순위화	precision	0.627	0.652	0.659	0.656
	recall	0.662	0.690	0.700	0.698
	F_1-measure	0.620	0.646	0.653	0.650

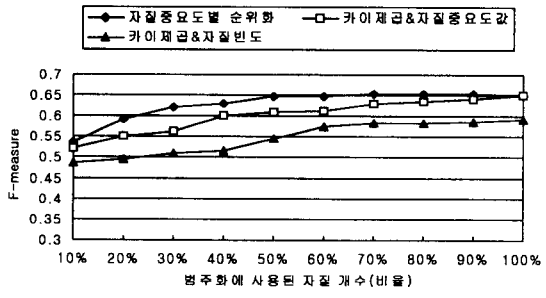

그림 3 . 선별된 자질의 개수에 따른 성능 비교

이 실험 결과를 보면 자질 중요도별 순위화를 이용한 실험(3)에서 자질의 비율이 70%일때를 보면, 나머지 실험(1)(2)에서 모든 자질을 사용했을 때와의 성능 비교에서 F1-measure값이 각각 5.9%, 0.2%의 성능 차이가 났다. 카이제곱 통계량으로 자질을 선별한 실험에 비해 높은 성능을 보이고 있다. 고빈도에 친화적인 방법이라고 알려져 있는 카이제곱 통계량 기법의 특성상 문서내 자질의 빈도만을 의존하지 않고 기타 구문의 어절 위치 정보 등이 추가된 방식으로 가중치를 달리했을 때 자질 선별에 따른 성능 차이가 있음을 알 수 있다.

5. 결론 및 향후 과제

본 논문에서는 자질 중요도에 따른 가중치 계산 기법을 이용한 자동 문서 범주화 시스템을 제안하였다. 기존의 출현 빈도에 문서의 내용을 대표하는 주제어의 가중치를 계산하였으며, 가중치 값의 순위화로 자질을 선별하였다. 본 연구에서 제안된 자질 추출 및 가중치 계산 기법은 출현 빈도에 의한 방법보다 좀 더 나은 성능 향상을 보였으며, 자질선별에 따른 성능 또한 차이를 보였다. 향후 과제로는 현재 실험 데이터 집합이 웹 디렉토리 문서들로 한정되어 있었지만 좀 더 다양한 문서, 신문기사나 기타 두괄식, 미괄식 형태의 실험과 이에 대한 적용 방법에 대한 연구가 필요할 것이다.

참고 문헌

[1] Yang, Y. and Xin Liu, "A re-examination of text categorization methods", In Proc. of Conference on Research and Development in Information Retrieval(SIGIR 99), pp.42-49, 1999.

[2] Y. Yang and J. P. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," In Jr. D H. Fisher(Ed.), the 14th International Conference on Machine Learning, Morgan Kaufmann, pp.412-420, 1997.

[3] F. Sebastiani, "Machine Learning in Automated Text Categorization", Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT, 1999.

[4] V. Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag, 1995.

[5] 강승식, 이하규, 손소현, 홍기채, 문병주, "조사 유형 및 복합명사 인식에 의한 용어 가중치 부여 기법", 한국정보과학회 가을 학술발표논문집, 28권 2호, pp.196-198, 2001.

[6] 고영중, 박진우, 서정연, "문장 중요도를 이용한 자도 문서 범주화", 정보과학회 논문지: 소프트웨어 및 응용, 29권 6호, pp.417-423, 2002.

[7] 이지행, 조성배, "전자우편 문서의 자동분류를 위한 다중 분류기 결합", 정보과학회 논문지: 소프트웨어 및 응용, 29권 3호, pp.192-201, 2002.