인트라넷 기반의 최적의 웹문서 자동 분류기법 선정

The selection of Best suited Automatic Web Document Classification Based on Intranet

°김국희, *윤희병
국방대학교 전산정보학과

Kukhee Kim°, Heebyung Yoon*
°*National Defence University, Dept. of Computer & Information Science
E-mail : cookie122@hanmail.net

요 약

인트라넷에서는 증가하는 웹문서의 검색을 목적으로 웹 검색엔진의 도입이 활발히 진행 중이며 대부분 찾아야할 키워드를 알고 접근하는 검색엔진 형태이다. 그러나 사용자가 무엇을 찾아야 하는지 모르는 경우 웹문서 분류체계는 효율적인 방법을 제시할 수 있다. 일부 구축되어 있는 분류체계는 수작업에 의한 분류로 인해 증가하는 웹문서의 양에 효율적으로 대처하기 곤란하므로 자동분류기법을 활용한 분류가 더 효율적일 것이다. 본 논문에서는 국방인트라넷의 수작업으로 구축된 분류체계를 대상으로 용어 가중치를 계산하는 방법을 달리하여 다양한 분류기법을 적용하여 성능을 비교평가하고 웹문서 자동분류시스템에 적용하여 분류성능의 향상을 도모하고자 한다.

1. 서론

월드와이드 웹의 등장으로 인한 웹 문서의 폭발적인 증가와 더불어 대부분의 자료들이 디지털화됨에 따라 문서를 보다 효율적으로 관리하기 위한 문서 분류의 필요성이 증가하고 있다[1].

또한 인트라넷을 운영하는 기업에서도 증가하는 웹문서 검색을 목적으로 검색엔진 도입이 활발히 진행 중이나 대부분 찾아야할 키워드를 알고 접근하는 키워드형 검색엔진 형태만으로 구축되고 있다. 반면 사용자가 무엇을 찾아야 하는지 모르는 경우는 디렉토리형 검색엔진이 그 해결책을 제시 할 수 있으며, 이것은 곧 웹문서 분류체계를 가리킨다. 인트라넷에서 일부 구축되어 운용중인 분류체계는 사람에 의해 분류체계를 구축, 유지하며 유입자료를 직접 분류하는 방식을 취하고 있는 수작업에 의한 방법으로 과다 비용소요 및 잘못된 분류로 업무의 효율성이 저하되고 증가하는 자료량은 수작업 분류의 한계를 초래하게 된다. 그러한 노동집약적인 방법으로 정보량이 방대하고 분류주제가 매우 유동적인 정보 환경을 대처하기는 매우 힘들다. 그런 점에서 대량의 웹문서를 효율적으로 분류하기 위한 자동분류체계에 대한 연구가 필요하다.

본 논문에서는 현재 국방부내 일부 군/기관에서 국방차원의 표준화된 분류체계가 없이 각 군/기관별로 수작업에 의해 분류체계를 도입 운용하며, 자료 분류 또한 작업자 개인의 주관적 판단에 따라 일관적인 분류가 이루어지지 않는 점을 고려하여 기존 수작업에 의해 이루어진 분류체계를 활용하여 수작업에만 의존하던 문서 분류를 다양한 자동분류 기법과 두 가지의 용어 가중치 부여 방법, 특징수를 고려한 자동분류 실험을 통해 최적의 웹문서 자동분류기법을 선정하여 웹문서 자동분류시스템 구축시 분류기법 선정방안을 제시한다. 실험에서는 인트라넷에 존재하는 여러 가지 정보유형중 웹문서만을 대상으로 하였다.

2. 관련연구

2.1 분류기법

증가하는 문서를 자동분류하고 이를 효율적으로 검색하기 위한 문서 분류기법에는 기계학습 분야에서 사용되는 알고리즘들이 사용되는데, 크게 규칙기반모델(Rule-based Model)과 연역적 학습 모델(Inductive Learning Model), 그리고 검색을 활용한 모델로 나눈다[2][3].

먼저 규칙 기반 모델은 학습문서들에서 나타나는 범주간의 구별된 규칙을 전문가가 찾아주거나 학습을 통해 추출된 규칙을 이용하여 문서를 분류하는 모델이다. 연역적 학습 모델로는 학습 문서에서 특징을 추출하여 이를 확률적인 접근 방법으로 사용한 나이브 베이지안(Naive Bayesian)과 트리구조로 표현하여 특징의 유무로 범주를 결정하는 결정트리(Decision Tree), 학습 문서를 통해 생성된 양성 특성(positive feature)과 음성 특성(negative feature)을 벡터 공간으로 표현하고 이들의 차이를 극명하게 하는 지지 벡터(support vector)를 찾는 SVM(Support Vector Machines)이 있다. 정보 검색 관점에서는 분류할 대상 문서를 질의로 보고 이와 유사한 문서를 찾는 방법인 KNN(k-Naerest Neighbor)이 있다.

2.2 방법별 분류체계 구축

전통적인 문서자동분류는 학습문서집합과 입력문서집합으로 구분되며 양질의 학습문서가 충분히 확보되어 있어야 하지만 실제 인트라넷 환경에서 학습문서를 초기에 충분히 확보하는 것이 어렵다. 그러므로 적은 개수의 학습문서를 가지고서도 양질의 분류모델을 만들 수 있어야 한다. 이러한 제한을 감안하여 인트라넷 기반의 문서자동분류는 다음과 같은 초기의 분류체계 구축 방법으로 사람에 의한 문서의 분류작업 대신 통계 알고리즘을 이용하여 관련 문서들을 그룹으로 묶어주는 Clustering에 의한 분류체계와 전문가에 의해 수작업으로 분류된 문서의 예제로부터 분류체계를 '학습'하는 Classification에 의한 분류체계 구축 방법으로 나눌 수 있다.[4]

Clustering에 의한 방법은 어떤 범주에 분류되어야 할 문서에 반드시 그 관련 단어가 많이 포함되어 있는 것은 아니기 때문에, 단어나 문자에 의한 Clustering은 실제 문서의 분야를 나타내는 데에는 한계가 있다. 또한 관련 분야의 문서의 개수가 많다고 해서 기업에서 그 분야의 중요도가 높다고 할 수는 없다.

Classification에 의한 방법은 주제별 전문가의 최소한의 도움만으로도 Clustering에 의한 방법보다 뛰어난 정확도를 보일 뿐 아니라, business need에 맞게 분류범주를 설정할 수 있다.

3. 인트라넷 기반의 자동문서분류기법 선정

일반적으로 기계학습에 기반한 분류는 학습과정과 분류과정, 두 단계로 구성된다. 입력문서에 대해 범주를 할당하는 과정은 그림 1과 같다. 본 실험에서 사용한 학습 및 분류에 사용한 기법은 나이브 베이지안, SVM, KNN, 결정트리를 사용하였으며. 사용한 분석 Tool은 Java로 구현된 WEKA[5]를 사용하였다.

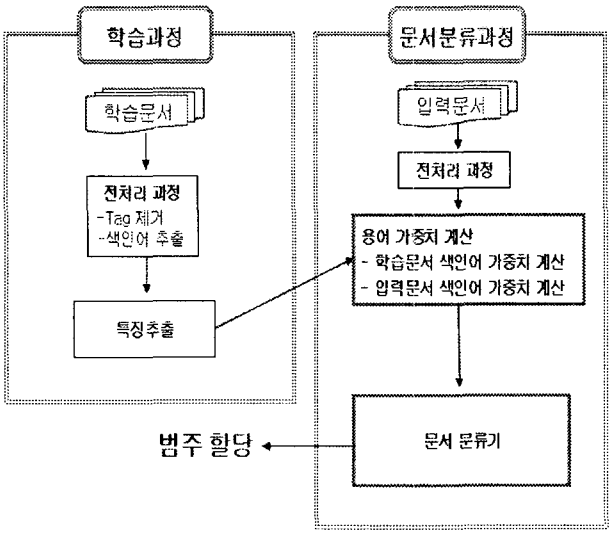

그림 1 문서 자동분류체계 구성도

3.1 문서집합 구성

각 문서집합은 국방인트라넷의 분류체계로부터 수집하였으며, 표 1과 같이 12개의 최상위 범주에 속한 웹문서 56개를 대상으로 하였다. 이중 70%의 문서를 학습문서로 30%의 문서를 입력문서로 사용하였다.

건강과 의학	교육과 학문	뉴스와 기상
동아리	법과 규정	복지와 생활
여가생활과 스포츠	컴퓨터와 인터넷	국방/기관
엔터테인먼트	게시판	자료실

표 1 최상위 범주의 분류

3.2 문서 자동분류체계 구성요소별 처리 절차

3.2.1 웹문서 전처리 과정

HTML Tag 제거

정확한 키워드 추출을 위해 먼저 HTML Tag의 제거 작업이 필요하며, Java로 구현하여 사용하였다.

색인어 추출

색인어 추출을 위해 한국어 형태소 분석기 KLT-v200[6]을 사용하여 5,560개의 키워드를 특징후보로 선택하고, 빈도수 2미만인 것과 한글명사만 추출하여 초기 800개의 특징을 추출하였다.

3.2.2 특징 추출(Feature Selection)

그림 2와 같이 전처리 과정을 거쳐 특징 구성수를 축소하여 각 학습기법에 적용 시는 특징추출 기법 중 문서 자동분류시 성능이 가장 좋은 카이제곱(χ^2) 통계량을 적용 하였고[7], 표 2는 학습과정에서 얻어진 특징의 순위를 나타낸다.

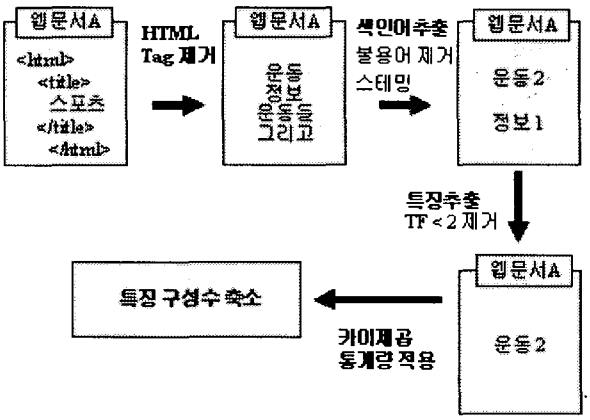

그림 2 전처리 및 특징추출 과정

순위	특징	χ^2	TF
1	운동	1,771,241,472	66
2	정보	1,764,384,336	63
3	과학	1,665,033,084	18
4	스포츠	1,665,033,084	71

표 2 학습과정에서 얻어진 특징들의 예

3.2.3 용어 가중치 계산 방법

문서 자동분류체계는 학습문서와 입력문서의 용어 가중치 계산 방법에 따라 그 성능이 달라 질 수 있다. 본 논문에서는 두 가지 방법으로 분류기법의 성능을 비교하였다.

TF*IDF 가중치

일반적인 벡터공간 모델에서 사용되는 방법으로 학습문서와 입력문서에 대한 가중치 계산이다. 범주 C에 대한 용어 t의 가중치를 $W_{t,c}$, 입력문서에 출현한 용어의 가중치를 $W_{t,d}$, N은 총 문서 범주의 개수, n_t는 용어t가 출현한 문서 범주의 개수라 할 때 식은 아래와 같고[8]. 표 3은 계산된 예이다.

$$W_{t,c},\ W_{t,d} = freq(t) \times \log \frac{N}{n_t}$$

특징	문서1	문서2	문서3	문서4	문서5	문서6	문서7	문서8
환영	0	0	0	0	0	0	0	0
방법	0	0.11	0	0	0	0	0	0.11
사람	0	0	0.65	0.65	0.87	1.09	0.22	0
교수	0	0	0	0	0	0	0	0

표 3 TF*IDF 가중치로 계산한 예

정규화 TF*IDF 가중치

TF는 문서에서 출현한 빈도수로 계산되므로 큰 문서에서는 값이 커지고, 작은 문서에서는 빈도수가 많이 나와도 그 값이 상대적으로 작아진다. 이런 편차를 줄이기 위해 정규화된 TF를 사용하며 식은 아래와 같고[8], 표 4는 계산된 예이다.

$$W_{t,c},\ W_{t,d} = \frac{freq(t)}{\max TF} \times \log \frac{N}{n_t}$$

특징	문서1	문서2	문서3	문서4	문서5	문서6	문서7	문서8
환영	0	0	0	0	0	0	0	0
방법	0	0.49	0	0	0	0	0	0.49
사람	0	0	0.67	0.67	0.89	1.11	0.22	0
교수	0	0	0	0	0	0	0	0

표 4 정규화 TF*IDF 가중치로 계산된 예

3.3 분류기법별 범주할당 결과

본 논문에서는 국방인트라넷 웹문서를 대상으로 나이브 베이지안, SVM, KNN, 결정트리의 네 가지 분류기법에 대하여 먼저 분류기법별로 학습문서 가중치 벡터 집합으로 학습을 하고, 학습된 각각의 분류기에 입력문서 가중치 벡터 집합을 계산하여 수작업으로 할당된 범주에 나타나는 정확율을 계산한 결과로 표 5, 6에서 나타내었다.

표 5는 TF*IDF 가중치 계산방법으로 특징 구성수를 100개, 300개, 500개, 800개로 정하고, 분류기법에 따른 정확율을 측정한 결과값이다.

(정확율:%)

분류기법 / 특징구성수	Naive Bayesian	SVM	KNN	Decision tree
100개	23.5	17.6	29.4	35.3
300개	29.4	17.6	11.8	35.4
500개	88.2	88.2	47.1	82.4
800개	94.1	88.2	47.1	82.4
평균	47.0	41.1	29.4	51.0
표준편차	37.5	40.8	16.9	27.1

표 5 TF*IDF 가중치에 따른 결과

표 6은 정규화 TF*IDF 계산방법으로 특징 구성수를 표 5와 같이 정하고, 분류기법에 따른 정확율을 측정한 결과값이다.

(정확율:%)

분류기법 / 특징구성수	Naive Bayesian	SVM	KNN	Decision tree
100개	88.2	82.4	82.4	82.4
300개	88.2	76.5	58.8	82.4
500개	88.2	58.8	41.2	82.4
800개	88.2	47.1	41.2	82.4
평균	88.2	66.2	55.9	82.4
표준편차	0	16.2	19.5	0

표 6 정규화 TF*IDF 가중치에 따른 결과

4. 비교 및 평가

그림 3, 4는 표5, 6의 결과에 따라 각각의 분류기법에 대한 비교를 나타내는 그래프이다.

그림 3에서는 분류기별 정확률이 비교적 낮은 오차의 성능을 유지하였으며, 나이브 베이지안이 보다 우수한 성능을 보였다. 또한 특징 구성수 증가시 정확률이 급격히 향상됨을 알 수 있다.

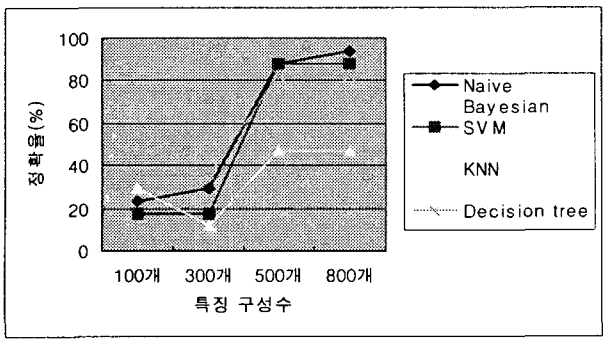

그림 3 TF*IDF 가중치 방식에서 분류기별 정확율 비교

그림 4에서는 나이브 베이지안과 결정트리는 특징 구성수에 상관없이 일관된 정확율을 나타내며, SVM, KNN은 오히려 특징 구성수를 축소시 정확율이 높게 나타난다.

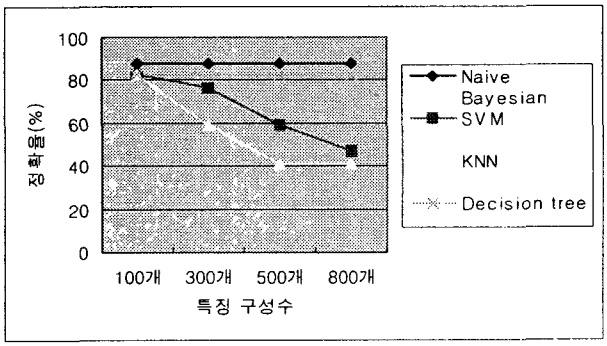

그림 4 정규화 TF*IDF 가중치 방식에서 분류기별 정확율 비교

분류기별 성능평가에서는 대상문서가 전체문서에 비해 현저히 적은 수의 문서집합과 특징 구성수가 적게 추출되더라도 높은 정확율을 나타낼 수 있으며, 전체문서에 대한 자동분류시스템에 적용시 분류성능 향상을 도모할 수 있다.

5. 결론

본 논문에서는 인트라넷 기반의 최적의 웹문서 자동분류기법 선정에 대한 실험을 통한 성능 비교 및 평가를 하였다. 문서분류 기계학습 이론 중 대표적인 4가지 방법에 대하여 실제 국방인트라넷 웹문서를 대상으로 어떤 학습기법이 우수한 결과를 보이는지에 대하여 측정하고 비교하였다. 다양한 분류기법 중에서 확률기반의 나이브 베이지안기법이 보다 우수 하였으며, 이 결과는 국방인트라넷에 자동문서분류시스템 도입시 적용하고자 하는 분류기법과의 우선하는 비교대상 요소로 활용할 수 있다.

향후 연구과제로서는 인트라넷의 정보량 증가에 따른 분류범주 변경시 재분류의 효율적인 방법이 제시되어야 할 것이다.

6. 참고문헌

[1] 강원석, 황도삼, 최기선, "의미의 상하위 정보를 이용한 웹문서 분류 시스템," 제11회 한글 및 한국어 정보처리 학술발표 논문집,pp.36~39,1999

[2] 박진우, "문장 중요도를 이용한 자동문서범주화," 서강대학교 컴퓨터학과 석사학위논문, 2002.

[3] Sahami, Mehran, "Learning dependence Bayesian classifier," KDD-96:Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp 335-338, 1996.

[4] Kamal Nigam, Andrew Kachites Maccallum, Sebastian Thrun, Tom Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM" To appear in the Machine Learning Journal, 1999.

[5] http://www.cs.waikato.ac.nz/ml/weka/

[6] http://nlp.kookmin.ac.kr/HAM/kor/

[7] Yang, Yiming, and J. O. Pedersen, "A comparative study on feature selection in text categorization," Machine Learning :Proceedings of the Fourteenth International Conference (ICML97), pp412-420,1997

[8] 이경찬, 강승식, "용어 가중치와 역범주 빈도에 의한 자동 문서 범주화," 제15회 한글 및 한국어 정보처리 학술발표 논문집. 2003.