자동 문서분류기의 개발동향 및 구축

임 희 석*

◈ 목 차 ◈

1. 서 론

자동문서 분류(automatic automatic text categorization)란 미리 정의되어 있는 범주를 입력 문서의 내용에 근거하여 컴퓨터가 자동으로 그 문서가 속하는 범주를 할당하는 작업을 의미하며, 문서 분류를 수행하는 시스템을 자동 문서 분류기라고 한다[Apte94, Manning99].

문서 분류기는 전통적으로 문서 분류를 위해 요구되었던 수작업량을 감소시키는 데 결정적인 역할을 할 뿐만 아니라 최근 인터넷이 폭넓게 보급되어 온라인 상에서 얻을 수 있는 텍스트 정보의 양이 증가하고 다루어야 할 정보의 양이 급증함에 따라 효율적인 정보 관리 및 검색을 위하여 매우 중요한 요소로 부각되고 있다. 현재 분류 검색 서비스를 제공하는 대부분의 국내 업체들은 수동 문서 분류에 의존하고 있으나, 정보의 생성 속도에 비하여 가공되는 속도가 지나치게 뒤쳐져 웹 공간에서의 정보 순환 주기에 적응하지 못할 뿐만 아니라 인건비 등 경제적으로도 많은 비용을 요구하게 된다. 이와 같은 비효율성을 감소시키기 위하여 문서 분류기의 개발은 매우 의미 있는 일이다. 비단 검색 업체에서의 문서 분류뿐만 아니라 온라인을 통해서 민원이나 서비스 신청을 받는 관공서 또는 기업의 경우 고객이 의뢰한 내용을 자동으로 분류하고 민원의 내용을 처리할 수 있는 해당 부서로 자동으로 전달하는 문서 라우팅(document routing)이나 사용자의 정보 요구에만 적합한 문서를 제공하는 문서 여과(document filtering)에서도 문서 분류기가 결정적인 역할을 한다.

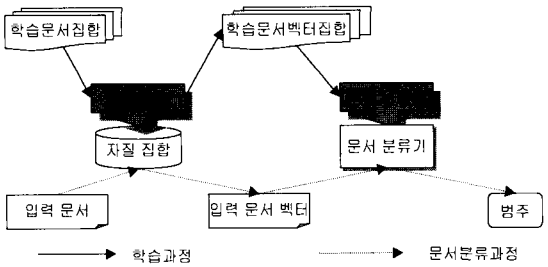
본 논문은 2장에서 문서 분류기의 개발에 필수적인 분류기의 기계 학습 방법과 자질 추출 방법 등 문서 분류기 개발에 관한 국내외 동향을 소개하고, 3장에서는 한국어 문서 분류기 개발을 위한 메모리 기반 학습에 관하여 설명한다. 4장에서는 메모리 기반 학습을 이용한 한국어 문서 분류기의 성능 향상 방법을 소개한다. 최종적으로 5장과 6장에서는 메모리 기반 학습을 이용한 한국어 문서 분류기의 성능 실험 결과와 결론에 대하여 서술한다.

2. 자동 문서 분류기의 개발 동향

2.1. 국내외 연구 동향

일반적으로 자동 문서 분류기 개발을 위해서는 그림 1과 같이 크게 두 가지 과정이 필요하다. 첫 번째는 일련의 단어 열로 구성된 입력 문서를 기계 학습이나 또는 범주 할당을 위한 정보로 사용하기 위한 자질(feature)을 추출 과정이고 두 번째는 자질로 표현된 학습 문서를 사용하여 일정 시간 내에 정확한 범주를 할당을 가능하게 하는 문서 분류기의 학습 과정이다. 자질 추출 과정은 문서 분류를 위하여 사용할 특성을 추출하는 과정으로 문서를 가장 잘 나타낼 수 있는 자질을 선택하는 것이 매우 중요하며 분류기 학습

* 천안대학교 정보통신학부교수



(그림 1) 문서 분류기 개발을 위한 개념적 과정

과정은 기계 학습 방법으로 어느 방법을 사용할 것인가가 중요한 관건이 된다.

문서 분류를 위한 자질로서는 정보 이득(information gain), TF/IDF, 상호 정보(mutual information) 등이 사용될 수 있으며 분류기 학습을 위해서는 신경망, 결정트리(decision tree), naive-bayesian 등의 방법들이 적용될 수 있다[Aptec94, Joachims98, Weiner95, Yang94, Yang97]. 영어의 경우 각각의 학습 방법과 자질을 사용한 분류기들이 개발되고 평가도 상당히 이루어져 영어의 특성을 반영할 수 있는 문서 분류기 개발을 위한 노력이 많이 이루어졌다.

반면, 국내에서는 이제 한국어 문서 분류에 대한 중요성과 필요성에 대해서 인식하고 몇몇 검색 엔진 업체 또는 인터넷 기업을 중심으로 연구가 시작되고 있으나 영어권에서 개발될 방법론을 한국어에 그대로 적용하는 수준에 머무르고 있다. 한국어는 영어와는 달리 한국어대로의 특성을 가지고 있다. 따라서 높은 정확률과 성능을 갖는 한국어 문서 분류기 개발을 위하여 타 언어에서 성공적으로 적용된 방법이 한국어 문서 분류를 위하여 효과적일 것이라고 단정지을 수 없으며 한국어 나름대로의 특성을 반영할 수 있는 자질의 선택과 분류 방법에 대한 고찰이 선행되어야 한다.

우수한 성능의 한국어 문서 분류기 개발을 위해서는 분류기 개발을 위한 알고리즘의 개발뿐만 아니라 분류기를 학습시키고 평가하기 위하여 사용될 수 있는 문서 범주 부착 코퍼스(document category annotated corpus) 구축이 절실히 요구된다. 문서 범주 부착 코퍼스란 문서가 가지고 있는 언어적인 오류가 제거되고 그 문서에 해당하는 범주가 부착된 문서 집합을 의미하며 문서 분류기 개발을 위한 학습 단계와 평가에 반드시 필요한 것이다. 많은 연구자들에 의해서 각기 다른 학습 문서 집합과 평가 집합을 이용한 결과로는 문서 분류기들의 공정한 평가가 이루어 질 수 없으므로 분류기 성능의 비교 판단이 어렵다. 또한 기존 연구의 문제점을 보완한 후속 연구의 진행과 그에 따른 성능 향상 결과를 비교하기가 어렵다. 이러한 이유로 영어의 경우, 20,000건의 로이터 신문기사로 구성된 로이터 코퍼스(Reuters corpus)가 구축되어 많은 연구자들이 공통적으로 사용하고 있고 평가를 위한 표준 집합으로 사용되고 있다. 한국어의 경우 로이터 코퍼스와 같은 표준화된 데이터가 전무한 상태이며 그에 따라 한국어 문서 분류기의 올바른 평가가 제대로 이루어지지 못하고 있는 실정이다.

2.2. 문서 표현에 대한 연구

문서 분류를 위해서는 문서를 벡터화하여 문서를 대표할 수 있는 형태로 변환하는 작업이 필요하다. 이 때 모든 단어를 자질 정보로 사용할 경우 기계학습으로는 처리하기 힘든 수준의 방대한 벡터가 생성되어 학습이 불가능해 질 수 있다. 따라서 분류기를 학습하는데 적절한 단어들을 자질로 선택하는 것이 중요하다. 일반적으로 많이 사용되는 자질로는 문서 빈도, 정보획득량, 상호정보, x2통계량이 있으며 이에 대한 간단한 설명과 연구 동향은 아래와 같다.

- 문서 빈도

이 방법은 특정 자질이 나타난 문서의 개수를 구하여 특정 개수 이하의 문서에서 출현하는 자질들은 제거하는 방법이다. 이는 거의 출현하지 않는 단어일수록 문서 분류에 도움이 되지 않는다는 가정에 의한 방법으로 가장 간단한 방법이지만 전통적으로 정보검색에 있어서 문서 빈도 값이 낮을수록 색인어로서 가중치를 높게 할당하는 것과 정면 배치되는 방법이다.

- 정보획득량(information gain)

이 방법은 기계 학습 분야에서 단어의 유용성을 측정하는 전통적인 기준으로 특정 단어의 출현 여부가 그 문서가 소속될 범주의 예측에 어느 정도의 정보량을 제공하는지 측정하는 방법으로 엔트로피 이론을 배경으로 하고 있다.

▪ 상호정보

상호정보란 정보이론 분야에서 전통적으로 단어간의 연관정도를 측정하는데 사용된 기준으로 두 단어의 상호 정보는 두 단어 중 한 단어가 출현했다는 사건이 다른 단어의 출현 여부를 예측하는데 기여하는 정도를 수치적으로 나타낸 값이라 할 수 있다. 단어 x와 y와의 상호 정보(MI)는 다음과 같이 계산된다.

$$MI(x,y) = \log \frac{p(x,y)}{p(x) \times p(y)}$$

▪ x2 통계량

이 방법은 단어와 범주간의 무관성(the lack of independence)을 측정하고 이를 자유도 1의 x2 분포와 비교, 그 치우쳐진 정도를 판단하는 기법으로 기대도수와 관측도수의 차이가 유의한지를 판단하는 기준또는 방법이 되는 검정 통계량이다[김우철94]. 이를테면 어느 교수가 통계학과목을 수강한 학생을 대상으로 나중에 자신이 개설할 과목을 또 듣겠냐는 질문을 했을 때 이 질문에 대한 대답이 학생이 받은 성적과 무관한지, 아니면 관련이 있는 지를 x2 통계량을 사용하여 측정할 수 있다. [Yang97]에서는 이러한 자질값 측정 방법을 사용하여 kNN에 의한 분류 실험을 한 결과 단어 빈도나 상호 정보 척도에 비하여 x2 통계량과 정보 획득량을 사용하는 것이 효과적임을 입증한 바 있다.

2.3. 문서 분류 학습에 대한 연구

자동 문서 분류의 학습에 대한 기존 연구는 규칙기반 방법, 확률기반 방법, 그리고 귀납적 기계학습 방법으로 구분할 수 있다.

▪ 규칙기반 방법

규칙기반의 접근 방법은 가장 먼저 시작된 방법으로 일종의 전문가 시스템을 구축하기 위한 방법이라 볼 수 있다[Hayes90]. 규칙기반 방법을 이용한 시스템으로는 Carnegie Group에서 개발한 CONSTRUE 시스템을 들 수 있으며 이 시스템은 로이터셋[Hayes90]에 대해서 실험한 결과 정확률과 재현율 모두 90% 이상의 훌륭한 결과를 보였다. 그러나 CONSTRUE와 같은 규칙기반 방법은 수작업으로 구축한 규칙을 사용하므로 다른 영역으로의 이전 또는 시스템 확장시 많은 시간과 비용을 요구한다.

▪ 확률기반 방법

이 방법은 확률에 기반하여 문서에 범주를 할당하는 것으로 학습 문서에 나타난 어휘들이 특정 범주의 문서에 나타날 확률을 계산하여 새로운 문서의 범주를 예측하는 것이다. [Lewis94]에서는 단순 베이지언 분류기(naive bayesian classifier)에 의한 asn서 범주화 실험을 통하여 결정 트리에 의한 범주화와 유사한 성능을 나타냄을 보였다. [Nigam99]에서는 최대 엔트로피 모델을 문서 범주화에 적용하여 좋은 성능을 나타내기도 하였으나 상당한 학습 시간을 요구한다는 문제점이 제기 되었다.

▪ 귀납적 기계학습(Inductive Machine Learning) 방법

귀납적 기계학습 기반의 연구은 보통 수작업에 의해 범주가 할당된 대량의 학습 문서에서 분류기를 학습한다. 이를 위해서 학습 문서를 벡터로 표현하는 일과, 벡터화된 학습 문서를 사용하여 분류기를 학습하는 과정이 요구된다. 귀납적 학습 방법은 문서 분류기 구축을 위한 방법 중 최근에 가장 각광을 받고 있는 방법이다. 귀납적 학습을 통한 자동 문서 분류의 접근은 기존에 기계 학습 영역에서 정립되어 온 다양한 기계학습 방법들을 실용적으로 적용할 수 있다는 점에서 그 의미가 크며 확률이나 규칙에 기반한 방법에 비해 우수한 성능을 보이고 있다. 귀납적 학습방법으로는 귀납적 규칙학습(inductive rule learning), 예제기반학습(example-based learning), 신경망 등이 있다.

귀납적 규칙 학습은 가장 먼저 연구되었던 방법 중 하나로서 [Aptec94]에서는 학습 문서 집합에서 귀납적 방법에 의해 DNF(Disjunctive Normal Form) 분류 규칙을 유도해 냈다. 예제기반 학습은 새로 분류할 문서와 유사한 학습 문서들을 추출하여 추출된 학습 문서에 할당된 범주를 새 문서에 할당할 범주의 결정에 사용하는 방법으로 [Yang94]에서는 kNN 방법을 문서 분류에 적용하여 높은 정확도를 보였다. 그러나 이 방법은 범주 할당에 필요한 비용이 많이 든다는 단점이 제기

되었다. 선형 분류기를 학습하기 위해서 [Lewis96]에서는 Rocchio 알고리즘과 LMS 알고리즘을 사용하여 문서 분류에 사용할 수 있는 분류기를 학습하고 실험하였으며 [Weiner95]에서는 퍼셉트론(perceptron) 기반의 신경망을 문서 분류에 적용하였다. 최근에는 지지벡터기계(support vector machine)에 기반한 문서 분류가 활발히 연구되고 있는데, 이 방법은 학습 자료들에서 나타나는 패턴으로 지지벡터를 자동 생성하고 이를 문서 분류에 사용하는 것으로 우수한 성능을 보이고 있어 최근 패턴 인식 연구분야에서 많이 이용되고 있으며[Joachims98], [Yang99b] 등에서 비교 실험을 통해 다른 범주화 학습 방법에 비하여 가장 우수한 성능을 보였다.

이 밖에도 [Larkey96]에서는 kNN과 베이지언 분류기가 생성하는 문서의 범주 할당 신뢰도에 가중치를 부여하여 합산한 값을 사용, 문서에 범주를 할당하는 분류기의 결합 방법을 제안하였고 [Lam98]에서는 kN과 선형 분류기의 단점을 보완할 수 있도록 학습 문서 집합을 일반화시킨 벡터들을 새로 kNN의 학습 벡터 GIS (generalized instance set)로 사용하는 방법을 제안하였다.

국내에서도 한국어 문서를 중심으로 최근 연구가 활발히 진행되고 있다. [강원석98]에서는 개념에 기반한 문서 분류 모델을 제안하였고, [정성화98]은 웹문서의 구조 정보를 이용하여 링크 정보를 이용하는 하이퍼텍스트 문서 분류모델을 제안하였으며 [이경순99]에서는 구축된 전문용어사전을 활용하여 문서 범주화에 사용하였다. [김상범99]는 선형분류기에 의해서 계산된 문서와 학습 문서집합을 이용하여 미리 계산된 범주별 클러스터를 이용한 범주들간의 상호 관련성을 고려하여 기존 문서 분류기의 성능을 개선하고자 하였다.

3. 메모리 기반 학습을 이용한 한국어 문서 분류

3.1. 메모리 기반 학습

메모리 기반 학습 방법 중 kNN 기계 학습을 이용한 문서 분류는 예제 기반 방법(instance-based method)으로 일반적인 목적 함수(target function)를 학습하는 기계 학습 방법을 사용하는 것과는 다르게 예제들만을 색인하는 것으로 모든 학습 과정이 끝나며, 문서 분류

(표 1) kNN 학습 및 문서 분류 알고리즘

- 학습 알고리즘
  - 모든 문서를 특징 벡터, x로 변환
  - 문서, x와 문서의 범주, c(x)에 대해서 1), 2) 과정을 반복
  1) <x, c(x)>를 저장
  2) x를 구성하고 있는 자질(단어)들을 색인
- 문서 분류 알고리즘
  - 입력 문서를 특징 벡터, x로 변환
  - 유사도 계산식에 따라 x와 유사한 k개의 이웃 선택
  - k개의 이웃과 범주결정 함수에 의해서 범주결정

시에는 입력 문서와 유사한 k개의 예제들을 이용하여 문서의 범주를 할당한다. kNN 기계 학습을 이용한 문서 분류기의 학습 알고리즘과 문서 분류 알고리즘은 표 1과 같다.

위의 학습 알고리즘에서 학습 문서의 색인 구조는 주로 정보검색에서 사용되는 역화일(inverted file)이 이용되며 문서를 특징 벡터로 변화시킬 때 벡터의 차원과 벡터를 구성하는 자질들의 결정이 매우 중요한 요소이다. 문서 분류 알고리즘에서는 어떤 유사도 계산식을 어느 것을 사용하는 지와 k개의 이웃을 이용하여 범주를 결정하는 함수를 무엇으로 쓰는 지가 문서 분류기의 성능을 좌우하게 된다. 즉 kNN 기계 학습을 이용한 문서 분류기의 성능을 좌우하는 파라미터는 다음과 같이 세 가지로 정리될 수 있다.

- 문서 표현을 위한 자질 선택 방법 및 자질 집합 크기(벡터차원)
- 이웃 문서 결정을 위한 유사도 계산 함수
- k개의 이웃 문서를 이용한 범주 결정 함수

다음 장부터는 kNN 학습을 이용한 한국어 문서 분류기와 위에서 제시된 세 가지 파라미터를 이용한 성능 개선 방안에 대하여 설명한다.

3.2. 자질 추출 및 가중치 계산

문서 분류를 위해서는 입력 문서를 벡터화하여 문서를 대표할 수 있는 형태로 변환하는 작업이 필요하며 자질은 문서 벡터를 구성하는 요소들로 사용된다.

자질의 종류와 자질 집합의 크기는 분류기 성능에 많은 영향을 미친다. 또한 선택된 자질이 입력 문서에서 차지하는 비중을 나타내는 가중치 부여 방법도 성능을 좌우하는 요소이다. 일반적으로 많이 사용되는 자질로는 문서 빈도, 정보 획득량, 상호정보, x2통계량이 사용되고 있다.

본 논문은 영어 문서 분류기의 평가[8]에서 비교적 우수한 성능을 보인 문서 빈도(DF)와 본 논문이 제안하는 새로운 자질 정보인 DF/ICF (Document Frequency/ Inverse Category Frequency) 값을 사용하는 방법을 소개한다. DF/ICF 값은 식 1과 같이 정의한다.

$$DF/ICF_i = DF_i \times \frac{1}{CF_i} \quad \text{(식 1)}$$

DF/ICF 값은 단어 i가 나타난 문서의 수가 많을수록 그리고 i가 나타난 문서들이 속한 분류 범주의 개수가 적을수록 높은 값을 가지게 된다. DF/ICF 값의 사용은 전통적으로 정보검색에 있어서 문서 빈도 값이 높을수록 색인어로서 가중치를 낮게 할당하는 것과 정면 배치되는 DF값 사용의 문제점을 보완할 수 있을 것으로 기대된다.

입력 문서를 분류기의 입력으로 변환하기 위해서는 선정된 자질들이 문서에서 차지하는 비중을 계산하는 작업이 수행되어야 한다. 이를 위해서는 이진 값 부여 방법과 가중치 계산 방법을 사용할 수 있다. 이진값 부여 방식은 특정 자질이 나타났으면 1, 그렇지 않으면 0을 부여하는 방식이며 이 방식은 계산이 간단하지만 문서 내에서의 자질의 역할을 올바르게 나타내지 못할 수 있다. 가중치 계산 방법은 특정 자질이 문서 내에서 차지하는 비중을 전통적인 정보 검색 방법에서 사용하는 TF/IDF(Term frequency/Inverse document frequency) 값을 이용하여 계산한다. TF/IDF 값의 계산은 식 2와 같다.

$$w_i = TF_i \times IDF_i$$
$$where\ TF_i = \frac{F_i}{\max_l F_l},$$
$$IDF_i = \log \frac{N}{DF_i} \quad \text{(식 2)}$$

식 2에서 TF_i는 단어 i가 문서 내에서 나타난 빈도 F_i값을 정규화하기 위하여 문서 내에서의 최대 F_l 값으로 나눈 값이며 IDF_i는 총 문서의 개수 N을 단어 I가 나타난 문서의 수 DF_i로 나눈 값이다. TF/IDF 값은 특정 단어가 특정 문서 내에서만 고빈도로 출현한 경우 가중치 값이 증가하는 함수로 이 값이 높은 단어는 다른 문서와 현재 입력 문서를 구분하는 능력이 큰 자질이라고 간주한 것이다.

3.3. 유사도 계산과 범주 결정 함수

범주 결정 함수는 유사한 k개의 문서를 이용하여 입력 문서의 범주를 계산하는 함수이고, 유사도 계산 함수는 입력 문서와 유사한 k개의 이웃(neighbor) 문서를 계산하는 함수를 의미한다. 본 논문은 유사도 계산하는 함수로는 두 벡터 사이의 거리를 계산하는 함수와 코사인 유사도 계산 함수를 사용하고자 한다. 두 개의 문서 벡터, d_i와 d_j의 거리 계산과 코사인 유사도 계산 함수는 각각 식 3과 식 4와 같다.

$$\mathrm{sim}(d_i, d_j) = -\sqrt{\sum_{r=1}^{n}(d_{ir} - d_{jr})^2} \quad \text{(식 3)}$$

$$\mathrm{sim}(d_i, d_j) = \frac{\vec{d_i} \cdot \vec{d_j}}{\left|\vec{d_i}\right| \times \left|\vec{d_j}\right|} \quad \text{(식 4)}$$

k개의 이웃 문서를 이용하여 문서 범주를 할당하는 범주 결정 함수로는 discreet-valued function(DVF), similarity-weighted function (SWF), 그리고 average-similarity-weighted function(ASWF)을 사용하고자 한다. DVF는 k개의 이웃 문서 중 가장 많은 개수의 문서w가 속한 범주를 할당하는 함수로 (식 5)와 같이 정의한다.

$$TC(\vec{d_q}) \Leftarrow \underset{c_j \in C}{\operatorname{argmax}} \sum_{\vec{d_i} \in kNN}^{k} y(\vec{d_i}, c_j) \quad \text{(식 5)}$$

식 5에서 함수 y는 $\vec{d_i}$의 범주가 c_j인 경우 1을 그렇지 않은 경우 0을 리턴하는 함수이다. 식 5를 사

용하는 DVF 방법은 매우 간단하지만 문서들간의 거리 또는 유사도 정보를 문서 분류에 사용하지 못하는 문제점이 있을 수 있다. DVF의 이러한 문제점을 보완하기 위해서는 문서들간의 유사도 정도를 반영할 수 범주 결정 함수가 필요하며 SWF는 이웃 문서들의 유사도 정도를 반영한 범주 계산 함수로 식 6과 같이 정의한다.

$$TC(\vec{d}_q) \Leftarrow \underset{c_j \in C}{\operatorname{argmax}} \sum_{\vec{d}_i \in kNN}^{k} \mathrm{sim}(\vec{d}_q, \vec{d}_i)\, y(\vec{d}_i, c_j) \qquad \text{(식 6)}$$

식 6과 같이 정의되는 SWF는 같은 범주에 속하는 이웃 문서들의 유사도를 모두 합하여 그 합이 가장 큰 범주를 문서의 범주로 할당하는 방법으로 sim $(\vec{d}_q, \vec{d}_i)$는 식 3 또는 식 4를 이용하여 계산하게 된다. SWF 방법은 충분히 많은 양의 학습 데이터가 제공되는 경우 잡음 데이터에 견고한 특성을 보이나 학습 데이터의 양이 충분하지 않은 경우 잡음 데이터로 인해서 잘못된 분류 결과를 초래할 수도 있으며 이러한 문제를 완화시킬 수 있는 방법은 잡음 데이터의 영향을 제거할 수 있는 ASWF 방법이다. ASWF 방법은 SWF 값을 해당 범주의 문서 개수로 나누어 평균값을 계산하는 방법으로 (식 7)과 같이 정의한다.

$$TC(\vec{d}_q) \Leftarrow \underset{c_j \in C}{\operatorname{argmax}} \frac{\sum_{\vec{d}_i \in kNN}^{k} \mathrm{sim}(\vec{d}_q, \vec{d}_i)\, y(\vec{d}_i, c_j)}{\sum_{\vec{d}_i \in kNN}^{k} y(\vec{d}_i, c_j)} \qquad \text{(식 7)}$$

kNN 문서 분류는 범주 결정 함수에서 이웃 문서로 사용되는 파라미터인 k에 따라 전역적 방법(global method)과 지역적 방법(local method)로 구분할 수 있다. 전역 방법은 학습 데이터에 있는 모든 문서를 이웃 문서로 사용하는 방법이며 지역 방법은 전체 학습 문서 중 유사도 계산에 따라 일정한 수의 k값을 사용하는 방법이다. 범주 결정 함수로 DVF를 사용하는 경우에는 입력 문서와 전혀 상관없는 sax 문서가 분류 결정에 영향을 미칠 수 있으므로 전역 방법을 사용하기 어려우며, 본 논문은 SWF와 ASWF 방법에 대해서 전역 방법과 지역 방법을 적용하여 k값의 변화에 따른 분류기의 성능을 평가한다.

4. 실험 및 평가

본 논문의 실험을 위해서 웹상에서 추출한 4,294개의 문서를 수집하였고 수집된 문서는 대범주 8가지, 중범주 20가지, 소범주 90가지 범주에 의해서 수작업으로 분류하여 분류 코퍼스를 구축하였다. 분류 코퍼스 중 데이터의 90%는 kNN의 학습을 위하여 사용하였고, 10%는 분류기의 성능 평가를 위한 실험 데이터로 사용하였다[1]. 4,294개의 문서에 나타난 유일한 단어 개수는 약 260,000개로 이중 실제 자질로 사용되는 단어는 문서 빈도를 이용한 자질 선택 방법에 따라 선정하였다.

분류기의 성능 평가를 위해서는 전통적으로 정보 검색 시스템의 성능 평가를 위해서 많이 사용되는 정확도(P)와 재현율(R)을 하나의 수치로 나타낼 수 있는 F_1 값을 사용하였으며, F_1값의 계산식은 식 8과 같으며 정확도와 재현율 계산은 식 9에 의해서 계산하였다.

$$F_1 = \frac{2RP}{R+P} \qquad \text{(식 8)}$$

$$P = \frac{b}{a+b}, \quad R = \frac{a}{a+c} \qquad \text{(식 9)}$$

식 9의 정확도와 재현율 계산을 위한 a, b, c값의 의미는 아래의 표와 같다.

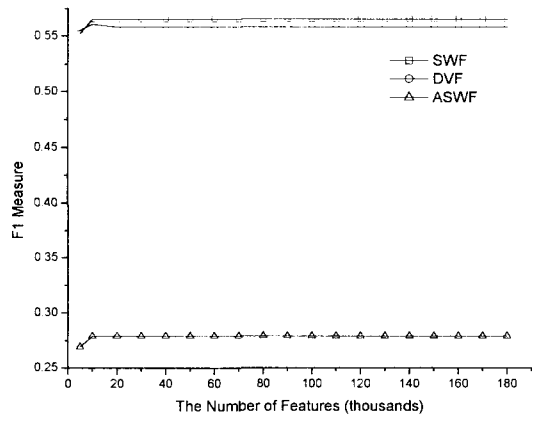
# of correct answer / #of system output	Yes	No
Yes	a	b
No	c	d

그림 2는 범주 결정 함수 DVF, SWF, ASWF를 사

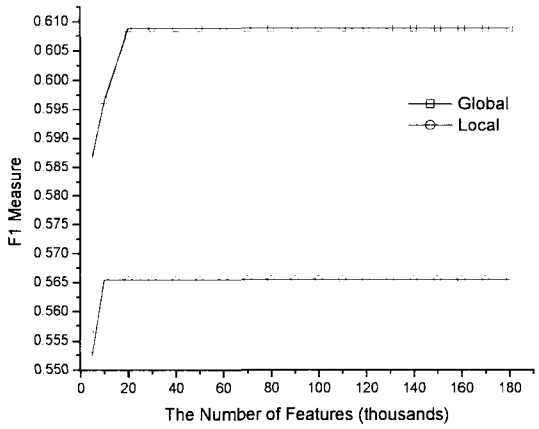
1) 본 논문에서 소개한 한국어 문서 분류기는 http://infocom.chonan.ac.kr/~limhs/cgi-bin/doccat/doccat.html에서 데모 중이다.

용하였을 때의 성능을 보이고 있다. SWF을 이용한 결과가 자질의 집합의 크기에 상관없이 가장 높은 F_1값을 보였고, ASWF 값이 가장 낮은 성능을 보였다. 또한 그림 2를 보면 어떤 범주 함수를 사용하는 경우에도 자질 집합의 크기가 100,000까지 증가하다가 그 이후에는 일정하게 유지됨을 보이고 있는데, 이는 전체 자질 집합의 약 10%정도만을 사용하여도 높은 분류 성능을 보일 수 있으므로 성능에 지장을 미치지 않고 나머지 90%에 해당하는 자질을 제거하여 자질 집합을 축소할 수 있음을 나타내는 것이다. 이 결과는영어 문서 분류의 자질 축소에 관한 [3, 9] 연구 결과와도 일치하는 것이다.

그림 3은 분류에 사용되는 이웃 문서의 개수에 따른



(그림 2) 자질 집합 크기에 따른 분류 결정 함수의 성능

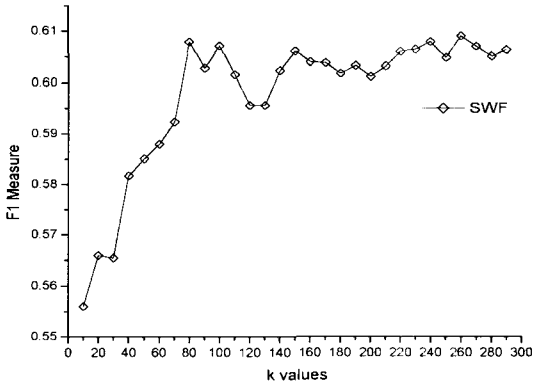


(그림 3) 전역 방법과 지역 방법의 비교

성능을 범주 결정 함수 중 가장 높은 성능을 보인 SWF를 이용하여 실험한 결과이다. Global의 방법은 학습 코퍼스의 전체 개수를 k로 사용한 전역적 방법을 나타내고 Local은 k값을 30으로 사용한 지역적 방법의 결과이다. 그래프에 서 보이듯이 모든 크기의 특성 집합에 대해서 전역적 방법이 지역적 방법보다 우수한 성능을 보였는데, 이는 학습 코퍼스의 분류가 균형있게 작성 된 것과 고정된 k값을 사용한 이유라고 판단된다.

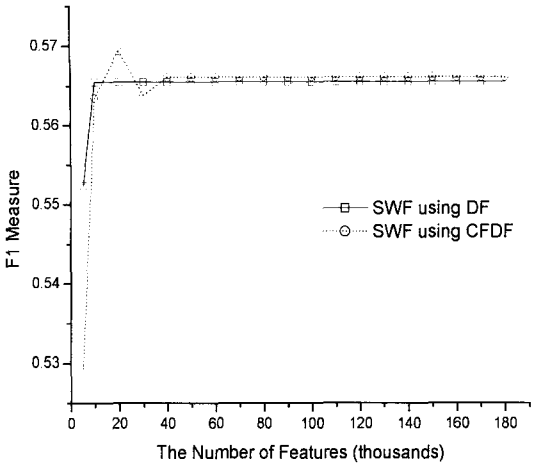
k값의 변화에 따라 지역적 방법도 전역적 방법과 유사한 성능을 보일 수 있을 것으로 예상된다. k값의 변화에 따른 성능을 보기 위하여 SWF 함수를 이용한 지역적 방법의 성능을 실험하였으며 그림 4의 그래프가 그 결과를 보이고 있다. 실험 결과 지역적 방법에서 이웃 문서의 개수를 결정하는 k값이 성능에 매우 중요한 영향을 미치는 것으로 나타났으며, 적당한 k값을 선정하는 경우[2]에는 전역 방법과 같은 많은 계산 비용을 소모하지 않고도 전역 방법과 유사한 성능을 보일 수 있음을 알 수 있다.

그림 5는 본 논문에서 특징 추출의 방법으로 제안한 DF/ICF의 성능을 DF값만을 이용한 경우와의 실험 결과이다. 실험 결과 이웃 문서의 분류 정보를 사용한 DF/ICF값을 사용한 경우가 거의 모든 크기의 자질 집합을 이용한 실험에서 높은 결과를 보였다. 한가지 흥미로운 사실은 자질 집합의 크기가 20,000의 경우에 DF/ICF값을 이용한 경우의 성능이 월등히 높음을 확인할 수 있었는데, 자질 집합 크기 20,000은 그림 2의 실험에서 모든 분류 함수의 성능이 급격하게 증가한 위치였다.



(그림 4) k 값에 따른 성능 변화

2) 본 실험 결과에서는 80개임.



(그림 5) DF 자질과 DF/ICF 자질의 성능

5. 결론

본 논문은 웹문서 자동 분류, 문서 라우팅 및 문서 여과 등에서 활용될 수 있는 자동 문서 분류기에 대한 고찰과 메모리 기반 학습 방법인 kNN 기법을 이용한 한국어 문서 분류기의 개발 사례를 소개하였다. kNN 기법을 이용한 문서 분류기의 성능 향상을 위한 방법으로 자질 선택의 방법, 다양한 분류 결정 함수에 따른 성능 분석, 분류를 위하여 사용하는 이웃 문서의 수에 따른 성능 변화를 실험하였다. 실험 결과, 이웃 문서의 분류 정보를 고려한 DF/ICF값을 사용한 경우와 분류 함수로는 이웃 문서들의 유사도의 가중치를 이용한 SWF 방법이 우수한 성능을 가질 수 있음을 보였다. 분류에 고려하는 이웃 문서의 범위로는 학습 데이터 전체를 사용하는 전역적 방법이 지역적 방법보다 우수함을 알 수 있었으나, 문서 분류기의 특성에 따른 적절한 k값을 사용한 지역적 방법도 많은 계산량을 필요로 하는 전역적 방법과 유사한 성능을 보일 수 있음을 알 수 있었다.

본 논문에서 문서 분류기의 성능 향상에 대한 소개는 kNN 기법에서의 성능 향상에만 초점이 맞추어져 있으나 향후에는 다른 신경망, svm, 그리고 결정 트리 등을 이용한 다른 기계 학습 방법을 이용한 한국어 문서 분류기의 성능 향상을 위한 연구가 계속 수행되어야 할 것이다.

참고문헌

[1] C. Apte and F. Damerau, "Automated learning of decision rules for text categorization", ACM Transactions on Information Systems, Vol. 12, No. 3, pp.233-251, 1994.

[2] Thorsten Joachims, "Text categorization with support vector machines : Learning with many relevant features", In International Conference on Machine Learning(ICML), 1998.

[3] D. D. Lewis and M. Ringuette, "Comparison of two learning algorithms for text categorization", In Proceedings of the 13rd Annual Symposium on Document Analysis and Information Retrieval, pp. 81~93, 1994.

[4] C. D. Manning, H. Schutze, Foundation of Statistical Natural Language Processing, The MIT Press, 1999.

[5] T. M. Mitchell, Machine Learning, McGraw-Hill companies, Inc., 1997.

[6] E. Weiner, J. O. Pedersenm and A. S. Weigend, "A neural network approach to topic spotting", In Proceedings of the 14th Annual Symposium on Document Analysis and Information Retrieval, 1995.

[7] Y. Yang. "Expert network : Effective and efficient learning from human decisions in text categorization and retrieval", In Proceedings of the 17th International Conference on Research and Development in Information Retrieval (SIGIR '94), pp. 13~22, 1994.

[8] Y. Yang, J. O. Pederson, "A Comparative study on feature selection in text categorization", In Proceedings of the 14th International Conference on Machine Learning, 1997.

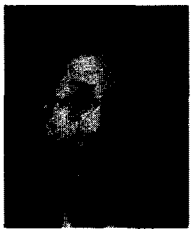
[9] Y. Yang, "An evaluation of statistical approaches to text categorization", Information Retrieval Vol 1. N o. 1/ 2, pp. 69~90, 1999.

[10] 강원석, 강현규, 김영섬, 시소러스 도구를 이용한 실시간 개념 기반 문서분류 시스템, 한국 정보과학회지논문지, 26 권, 1호 , 1999.

[11] 김상범, 범주간의 상호관계를 고려한 자동 문서 범주화의 개선, 고려대학교 컴퓨터학과 석사학위 논문, 1999.

[12] 서울대학교 자연과학대학 계산통계학과, 통계학 개론, 영지문화사, 1994.
[13] 이경순, 최기선, 전문용어 및 정보추출에 기반한 문서분류 시스템, 제 11 회 한글 및 한국어 정보처리학술대회, 1999.
[14] 정성화, 이종혁, 문서 구조 정보에 기반한 웹페이지 범주화 모델, 제10회 한글 및 한국어 정보처리학술대회, 1998.

◐ 저 자 소 개 ◑



임 희 석

1988년 3월~1992년 2월 고려대학교 컴퓨터학과(이학사)
1992년 3월~1994년 2월 고려대학교 컴퓨터학과(이학석사)
1994년 3월~1997년 8월 고려대학교 컴퓨터학과(이학박사)
1997년 9월~1999년 2월 삼성종합기술원 선임연구원
1999년 3월~현재 : 천안대학교 정보통신학부교수
주 관심분야 : 인공지능, 자연어처리, 인터넷 정보 시스템, 기계 학습, 정보검색