Nearest Neighbor 방법을 이용한 문서 범주화에서 범주 자질의 평가†

권오욱, 이종혁, 이근배
포항공과대학교 전자계산학과

An Evaluation of Category Features in Text Categorization Using Nearest Neighbor Method

Oh-Woog Kwon, Jong-Hyeok Lee, Geunbae Lee
Dept. of Computer Science and Engineering, POSTECH

요 약

문서 범주화에서 문서의 내용에 따라 적합한 범주의 종류와 수를 찾는 문제를 해결하기 위해서는 문서 당 하나의 범주를 할당할 경우에 가장 좋은 성능을 보이는 모델이 효과적일 것이다. 그러므로, 본 논문에서는 문서 당 하나의 범주를 할당할 경우에 좋은 결과를 보이는 k-nearest neighbor 방법을 이용한다. 그리고 k-nearest neighbor 방법을 이용한 문서 범주화의 성능을 향상시키기 위해서, 문서 표현에 사용하는 단어들을 범주 자질의 성격을 갖는 단어들로 제한하는 방법을 제안한다. 제안한 방법은 Reuter 신문 일년치로 구성된 Reuter-21578 테스트 집합에서 breakeven point 82%라는 좋은 결과를 보였다.

1. 서론

많은 문서를 다루는 문서 저장 시스템(text storage system)과 문서 검색 시스템(text retrieval system)에서는 문서에 주제어 분류 코드(subject classification code)들을 할당하여 시스템의 성능을 향상시켰다. 주제어와 같은 범주(category)들을 문서에 할당하는 문제를 문서 범주화(text categorization)라고 한다.

문서 범주화 문제를 해결하기 위한 대부분의 기존 연구들은 학습 문서들(training documents)에서 학습한 결과를 이용하여 문서에 적합한 범주들을 할당하였다. 기존 연구들 가운데서 성공적인 결과를 보인 방법들을 크게 3 가지 접근 방법으로 분류할 수 있다. 먼저 학습 문서들에서 나타나는 범주들간의 상호 구별이 가능한 규칙을 전문가가 직접 찾거나 학습(learning)으로 추출한 규칙을 기반한 방법(rule-based method)이 있다[4][11]. 그리고, 학습 문서에서 추출한 범주 자질(category feature)을 이용하는 베지언 확률 모델(Bayesian probability model)[6][7]이나 결정 트리 학습 알고리즘(decision tree learning algorithm)[7]에 기반한 방법들을 들 수 있다. 마지막으로 학습 문서들을 문서에 범주를 할당하기 위한 예로 사용하는 k-nearest neighbor 방법이 있다[3][10][12]. 이러한 방법론들은 대부분 비슷한 결과들을 보였다.

문서 범주화는 자질 추출(feature extraction) 단계와 문서에 적합한 범주를 할당하는 단계로 나누어진다. 대부분의 문서 범주화 방법들은 범주 할당 단계에서 문서에 할당할 범주들을 순서화(ranking)한 후, 문서 당 n 개의 범주를 할당하거나 사용자가 정의한 임계치(threshold value)를 넘는 범주들만을 할당한다. 하지만 이러한 방법들은 문서에 따라 할당하여야 범주들을 수

† 이 논문은 '97 년도 한국통신(과제제목: 웹검색 서비스용 자동 문서분류 시스템 연구)의 지원비 지원에 의한 결과임

및 임계치 값이 변한다는 사실을 고려하지 않았다. [1]에서는 문서에 대한 가장 적합한 n 개의 범주들을 찾기 위해서 범주들의 조합으로 구성한 메타 범주(meta-category)라는 새로운 개념을 도입하여 범주할당 문제를 해결하였지만, 조합 문제를 풀기 위한 시간 복잡도(time complexity)가 증가하는 단점을 가졌다. 문서에 가장 적합한 범주들만을 할당하는 문제를 최적 범주 할당이라고 한다.

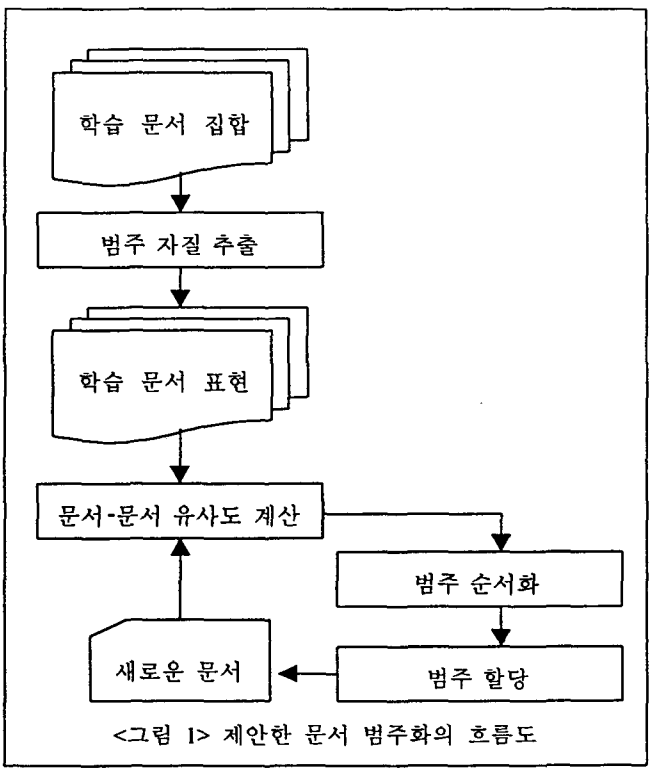
실용적인 최적 범주 할당 문제를 해결하기 위해서는 문서 당 하나 이상의 범주들을 할당한다는 문서 범주화의 가정에서 각 문서에 하나의 범주 할당이 성공적인 방법을 찾는 것이 중요하다. [7]에서 베지언 확률 모델, k-nearest neighbor 방법과 relevancy feedback 방법을 결합하였는데, 각 모델들의 특성을 실험 및 비교한 결과 k-nearest neighbor 방법이 문서 당 하나의 범주만을 할당할 경우에 가장 좋은 결과를 보였다. 그러므로, 본 논문에서는 k-nearest neighbor 의 성능 향상을 위한 방법론을 제시하고 이를 평가한 결과를 토대로 최적 범주 할당 문제를 해결하기 위한 방안을 모색하고자 한다.

k-nearest neighbor 방법은 학습 문서 집합에서 새로이 범주를 할당할 문서와 가장 관련성이 있는 k 개의 문서들을 추출한 후, 이 문서들이 가지는 범주들을 새로운 문서가 가져야 할 범주들이라고 생각하는 방법이다. 이러한 k-nearest neighbor 방법은 전통적인 문서 검색 시스템(text retrieval system)에서 질의(query) 대신 문서를 이용하는 것과 유사하다. 그러므로, 기존 k-nearest neighbor 방법들은 문서 검색의 색인과 동일한 방법만을 이용하여 문서를 표현하였다. 또한, 시스템의 성능 향상보다는 방법론 자체가 학습 집합의 모든 문서들과 비교하기 때문에 발생하는 검색 속도 저하 해결에 관심을 가졌다 [10][12].

본 논문에서는 k-nearest neighbor 방법에서 범주에 대한 자질을 추출하여 시스템의 성능을 향상시키고 자질들이 문서 범주화에 미치는 영향에 대해서 실험 및 평가하고자 한다.

2. k-nearest neighbor 문서 범주화

일반적으로 k-nearest neighbor 방법은 문서 검색 시스템의 역색인 화일(inverted index file)로 구현한다. 기존의 방법들은 학습 문서들을 전통적인 정보검색에서의 색인과 같이 단지 불용어(stopword)를 제거한 단어들로 표현하였다. 하지만, 문서 범주화에서는 학습 문서를 문서 중심으로 표현하기 보다는 문서가 가지는 범주 중심으로 표현해야 한다. 일반적으로 문서에서 범주의 성질을 잘 나타내는 단어들을 범주 자질이라고 생각한다. 이러한 범주 자질로 학습 문서들을 표현하면 k-nearest neighbor 방법에서 보다 좋은 성능을 보일 것으로 기대된다.



<그림 1> 제안한 문서 범주화의 흐름도

<그림 1>은 본 논문에서 제안하는 k-nearest neighbor 방법을 이용한 문서 범주화의 흐름도이다. 먼저 학습 문서에서 범주 자질의 성격을 갖는 단어들을 추출한다. 이들 자질로 학습 문서를 표현하여 새로운 문서와 문서 유사도를 계산한다. 이때, 문서 표현 방법은 벡터 공간 모델(vector space model)을 이용하고 벡터간의 유사도는 코사인 유사도(cosine similarity measurement)로 계산한다. 학습 문서에서 새로운 문서와 유사도가 큰 k 개의 문서들을 추출한다. k 개의 문서들에 나타난 범주들을 새로운 문서에 할당할 범주들의 후보로 보고 범주 순서화를 통하여 문서에 문서 당 n 개의 범주를 할당하거나 임계치에 의해서 범주들을 할당한다.

우선 제안하는 방법을 설명하기 전에 본 논문에서 실험을 위하여 사용하는 테스트 집합(test collection) Reuter-21578 에 대해서 간단히 설명하고 이 테스트 집합을 근거로 제안한 방법을 설명하겠다.

2.1 테스트 집합

본 논문에서 사용할 테스트 집합 Reuter-21578 은 1987

년의 Reuter 신문 기사 전문 21,578 건과 135 개 경제 부분 주제어로 구성되어 있다. 원래 Reuter-22173 이란 이름으로 공개되었으나, 중복되는 문서를 제거하고 철자법 오류와 잘못된 범주 할당을 수정하였다. 그리고 또한 문서를 다루기 쉽게 하기 위해서 SGML (Standard Generalized Mark-up Language)로 문서를 표현하였다.

Reuter-21578 로 실험한 많은 연구들이 있었으나, 자신들의 실험에 맞게 학습 문서 집합(training document set)과 실험 문서 집합(test document set)으로 구분하였다. 대표적인 예로 베지언 확률 모델에서 사용한 Lewis division[6]과 규칙을 기반한 방법에서 사용한 Apté division[4]이 있다. Lewis division 은 1987 년 8 월 8 일 이전의 기사 13,635 개를 학습 문서 집합으로, 그 이후의 기사 6,188 개를 실험 문서 집합으로 구분하였다. 이 division 에서는 범주들이 학습 문서 집합에 나타날 확률과 실험 문서 집합에 나타날 확률이 비슷하여 확률 모델을 테스트하기에 적합하다. Lewis 는 문서에 범주가 할당하지 않을 경우도 고려하였기 때문에 이 division 의 문서들 중에는 상당 부분이 범주를 가지지 않는다. Apté 는 Lewis division 에서 범주를 가지지 않는 문서를 제거하여 9,603 개의 학습 집합과 3,299 개의 실험 집합으로 구분하였다.

본 연구에서는 Reuter-21578 의 Apté division 에서 범주가 할당되어 있지 않는 문서가 있어서 제거하였고, 또한 본문이 없이 제목만이 있는 문서들을 제거하였다. 이 결과로 7,110 개 문서들로 구성된 학습 문서 집합과 2,756 개 문서들로 구성된 실험 문서 집합을 테스트 집합으로 구축하였다. <표 1>은 본 논문의 테스트 집합에 대한 분석 결과이다

<표 1> 본 논문의 테스트 집합 분석

	훈련 집합	실험 집합
문서 수	7,110	2,756
문서 당 평균 단어 수	75.8	70.6
할당된 범주 종류	115	92
문서 당 평균 범주 수	1.25	1.24
50 문서 이상에 할당된 범주 종류	23	9
20 문서 이상에 할당된 범주 종류	41	23

2.2 자질 추출

일반적으로 문서 범주화에서 범주에 대한 자질로써 학습 문서 집합에서 범주와 많이 공기(Co-occurrence)하는 단어들을 선택한다. 범주들을 문서 단위로 할당하기 때문에 단어와 범주가 같은 문서에서 나타나는 경우를 공기한다고 한다. 베지언 확률 모델에서 단어와 구(phrase) 단위의 자질을 실험한 결과에 의하면 단어가 구보다는 좋은 성능을 보였다[6]. 물론 단어와 구를 동시에 사용한다면 보다 좋은 결과를 보일 것이다. 본 논문에서는 자질의 단위를 단어로 한정하였다.

범주에 대한 자질을 선택하기 위해서 문서에 나타나는 단어들을 다음과 같은 순서에 의해서 제한하였다.

1. 97% 이상의 정확률을 보이는 Brill 의 규칙을 기반한 태거(rule-based tagger)[8]로 단어들을 태깅하여 기능단어(functional word)들을 제거하고 숫자와 기호로 구성된 단어들을 제거한다.
2. 나머지 단어에서 불용어를 제거한다.
3. Porter 어간 추출기(Porter stemmer)를 이용하여 어간을 추출한다.
4. 추출한 어간에서 불용어를 제거한다.
5. 학습 문서 집합에서 한 문서에서만 나타나는 단어들을 제거한다.

위와 같이 범주 자질이 될 수 없는 단어들을 제거한 후, 범주 자질인 단어들을 추출하기 위해서 기대 상호 정보 척도(expected mutual information measure)를 이용한다. 단어 W_i와 범주 C_j에 대한 기대 상호 정보 $I(W_i, C_j)$는 식(1)과 같다[5][6].

$$I(W_i,C_j) = \sum_{a=0,b=0,1} \sum P(W_i = a, C_j = b) \log_2 \frac{P(W_i = a, C_j = b)}{P(W_i = a) \times P(C_j = b)} \quad \text{----------(1)}$$

식 (1)에서 $P(W_i=1)$은 문서에 단어 W_i 가 나타날 확률이고 $P(W_i=0)$은 문서에 단어 W_i 가 나타나지 않을 확률이다. 마찬가지로 $P(W_i=1, C_j=1)$은 문서에 범주 C_j가 할당된 경우, 단어 W_i가 그 문서에 같이 나타날 확률이다. 단어와 범주 간의 기대 상호 정보 $I(W_i, C_j)$는 범주 C_j가 단어 W_i와 함께 공기하던지 아니면 상호 전혀 다르게 공기하던지 할 경우에 높은 값을 가진다. 반대로 단어 W_i가 범주 종류에 무관하게 출현한다면 기대 상호 정보 $I(W_i, C_j)$는 낮은 값을 가진다. 그러므로, 기대 상호 정보 측도를 이용하면 단어들이 여러 범주 자질로 판단되지 않아서, 범주 자질이 상호 배제적인 성격을 가진다.

본 논문에서는 학습 문서를 그 문서가 가지는 범주를 구별할 수 있는 단어만으로 표현하기 위해서 상호 기대 정보 값이 일정한 임계치를 넘는 단어-범주 공기의 단어만으로 학습 문서를 색인한다. 그러므로, k-nearest neighbor 방법에서 범주를 새로이 할당할 문서에 가장 가까운 k 개의 문서를 검색할 경우, 범주의 자질로 사용되지 않는 단어로 인해 검색되는 문서들의 수를 줄여 보다 좋은 범주 순서화 및 범주 할당이 가능하게 할 수 있다.

2.3 문서-문서 유사도 계산

일반적으로 k-nearest neighbor 방법을 기반한 문서 범주화는 두 단계로 이루어진다. 첫 번째 단계에서는 새로이 범주를 할당할 문서와 가장 문서-문서 유사도가 높은 k 개의 학습 문서를 추출한다. 두 번째 단계에서는 추출된 k 개의 문서들이 가지는 범주들을 이용하여 새로운 문서가 가질 수 있는 범주의 가능성을 계산하여 범주들을 할당한다. 이 절에서는 우선 첫 번째 단계를 설명하고, 다음 절에서 두 번째 단계의 방법을 언급하겠다.

본 논문에서 학습 문서를 표현하는 방법은 전통적인 정보검색의 색인에서 많이 사용하는 벡터 공간 모델을 이용한다. 문서에 대한 단어 가중치 계산 방법은 용어 빈도수(term frequency)와 문서 역빈도(inverse document frequency)를 변형시킨 INQUERY 식의 단어 가중치를 이용한다. [4]에서 제목에 나타나는 단어의 빈도수를 본문에서 나타나는 빈도수와 달리 2 배로 하여 보다 나은 결과를 보였다. 본 논문에서도 제목에 나타나는 단어의 빈도를 2 로 하여 제목에 나타나는 단어의 가중치를 높였다. 문서 d 에 나타나는 단어 W_i의 가중치 $W_{d,Wi}$는 식 (2)와 같다[2].

$$W_{d,W_i} = \frac{\log(tf_{d,W_i}+0.5)}{\log(\max_{tf}+1.0)} \times \frac{\log\left(\frac{N}{n}\right)}{\log(N)} \quad \text{----------(2)}$$

where

tf_{d,W_i} : 문서 d에서 W_i의 빈도수,

$\max_{tf}$: 문서 d에서 가장 많이 나타나는 단어의 빈도수

N : 학습 문서의 수

n : W_i가 나타나는 문서의 수

범주를 새로이 할당할 문서 D_x와 학습 문서 D_{ij} 간의 유사도를 계산하기 위해 전통적인 벡터 공간 모델에서 사용하는 코사인 유사도 식 (3)을 이용한다. 식 (3)에 의하여 문서 D_x에 가장 유사도가 높은 k 개의 학습 문서들을 추출한다.

$$sim(D_x, D_{ij}) = \frac{\sum_{t_i \in (D_x \cap D_{ij})} W_{D_x,t_i} \times W_{D_{ij},t_i}}{\|W_{D_x}\|^2 \times \|W_{D_{ij}}\|^2} \quad \text{----------(3)}$$

where

t_i : D_x와 D_{ij}에서 같이 발생하는 단어

$$\|W_{D_x}\|^2 = \sqrt{W_{D_x,1}^2 + W_{D_x,2}^2 + W_{D_x,3}^2 + \ldots}$$

$$\|W_{D_{ij}}\|^2 = \sqrt{W_{D_{ij},1}^2 + W_{D_{ij},2}^2 + W_{D_{ij},3}^2 + \ldots}$$

2.4 범주 순서화와 범주 할당

새로이 범주를 할당할 문서와 유사도가 높은 k 개의 학습 문서를 추출한 후, 이들 k 개의 문서들이 가지는 범주들을 이용하여 새로운 문서에 대한 범주 할당 적합성을 계산하여 순서화한다. 이때 사용되는 범주 순서화 방법으로 k 개 문서들에서 나타나는 범주들의 빈도 순으로 하는 방법과 범주가 포함되어 있는 문서의 유사도를 합산하는 방법이 있다. 빈도순으로 할 경우, 동점인 범주들이 많이 발생하기 때문에 문서 유사도를 합산하는 방법이 효과적이다[9]. 본 논문에서도 새로운 문서에 범주를 할당할 가능성 계산하기 위해서 문서 유사도를 합산한다.

k-nearest neighbor 방법은 전통적인 분류 방법의 후위 확률(posterior probability)에 근사적이기 때문에 새로운 문서 D_x에 범주 C_k를 할당할 확률 $P(C_k \mid D_x)$를 식 (4)와 같이 표현한다[12]. 본 논문에서 식 (4)를 문서에 대한 범주 유사도라고 한다.

$$P(C_k \mid D_x) \approx \sum_{D_j \in \{\text{k top-ranking documents}\}} sim(D_j, D_x) \times P(C_k \mid D_j) \quad \text{--------(4)}$$

일반적으로 식 (4)에서 $P(C_k \mid D_j) = 1$ 로 보아 계산을 한다. 본 논문에서는 한 문서가 2 개 이상의 범주들을 가질 경우에, 범주 자질을 이용하여 문서에서 각 범주에 대한 중요도를 다르게 하는 전략을 채택하여 식 (5)와 같이 사용한다.

$$P(C_k \mid D_j) \approx \frac{D_j\text{에서 } C_k\text{의 자질인 단어 수}}{D_j\text{에 나타나는 단어 수}} \quad \text{----------(5)}$$

본 논문에서 식 (5)의 경우와 $P(C_k \mid D_j) = 1$ 로 하는 경우를 구별하기 위해서 식 (5)의 형식을 이용한 범주 유사도를 비례 범주 유사도라고 하고 $P(C_k \mid D_j) = 1$ 을 이용할 경우를 균일 범주 유사도라고 한다.

식 (4)의 계산에 의하여 범주를 할당할 문서 D_x와의 유사도에 따라 범주들을 순서화할 수 있다. 이러한 경우, 적합한 범주들만을 문서 D_x에 할당하기 위해서 다

음과 3 가지 방법론들을 주로 사용한다[6].

1. 문서 당 k 범주 할당 : 각 문서에 상위 k 개의 범주들만 할당하는 방법이다. k 의 선택은 훈련 문서에서 문서 당 범주들의 비율을 고려하면 도움이 될 것이다.
2. 임계치에 의한 할당 : 사용자가 정의한 임계치를 넘는 범주 유사도를 가진 모든 범주들을 문서에 할당하는 방법이다.
3. 비례 할당(proportional assignment) : 먼저 각 범주에 대해 실험 문서들을 범주 유사도로 내림차순으로 정렬한다. 그리고, 학습 문서들이 그 범주를 가지는 비율에 대해 사용자가 정의한 비례 상수(proportional constant) 배를 한 만큼의 우선 순위에 있는 실험 집합의 문서들에 그 범주를 할당한다. 예를 들면, 학습 문서에서 범주 C_k를 가지는 문서들의 비율이 10%이고 비례 상수가 0.5 라면 범주 유사도가 높은 2%의 실험 문서들에 범주 C_k를 할당한다.

서론에서도 언급하였지만, 1 과 2 의 방법은 범주를 할당할 문서의 특성을 고려하지 않는다는 단점을 가지고 3 의 경우는 항상 일괄 처리(batch processing)에서만 가능하다는 단점을 가진다. 하지만, 본 논문에서는 범주 자질이 k-nearest neighbor 문서 범주화의 성능에 미치는 영향을 고려하므로, 위와 같은 단점에도 불구하고 1 과 2 의 방법을 채택하여 실험한다. 이 실험 결과를 바탕으로 향후 연구에서는 범주 할당 문제를 최적화할 수 있는 방법에 대하여 고려할 것이다.

3. 평가 방법

k-nearest neighbor 문서 범주화에서 범주와 문서간의 유사도 계산에 사용되는 단어들이 달라지는 것에 따른 성능 차이를 평가하기 위해서, 다음과 같이 단어와 범주 자질을 4 가지로 분류하여 실험 및 평가한다. 1 과 2 는 단어들의 집합으로 학습 문서들을 색인한 경우이고 3 과 4 는 범주 자질만으로 학습 문서들을 색인한 것이다.

1. Not Stemming Word Set : 학습 문서에 나타나는 단어들 중에서 태깅 후, 기능단어와 숫자, 특수 문자, 불용어를 제거한다. 그리고, 학습 문서들에서 단어의 문서 빈도가 1 인 단어들을 제거한다.
2. Stemming Word Set : Not Stemming Word Set 에서 Porter 어간 추출기로 추출한 단어의 어간들로 학습 문서를 표현한다.
3. Feature Set(1) : Stemming Word Set 의 단어들과 범주들 간의 기대 상호 정보 값이 0.0002 이하인 단어-범주 공기들을 학습 문서들에서 제거한 후, 색인을 한다.
4. Feature Set(2) : Stemming Word Set 의 단어들과 범주들 간의 기대 상호 정보 값이 0.0005 이하인 단어-범주 공기들을 학습 문서들에서 제거한 후, 색인을 한다.

<표 2>는 위의 4 가지 단어 및 자질 집합을 이용할 경우, 학습 문서 집합에서 유일한 단어 수와 단어-범주의 공기 수를 나타내고 있다.

<표 2> 단어 및 자질 집합에 대한 분석

구분	Not Stemming Word Set	Stemming Word Set	Feature Set(1)	Feature Set(2)
단어 수	15,154	11,906	10,313	6,648
단어-범주 공기 수		91,359	44,269	18,820

<표 2>에서 Feature Set(1)과 (2)는 Stemming Word Set 에서 범주 자질을 추출하기 때문에 비교대상이 아닌 Not Stemming Word Set 에서의 단어-범주 공기 수는 표시하지 않았다.

본 논문에서 기존의 범주 유사도와 달리 식 (4)에서 $P(C_k|D_j)$를 식 (5)로 계산하는 비례 범주 유사도를 평가하기 위해, 식 (5)를 계산할 수 있는 Feature Set(1)과 (2)에서 균일 범주 유사도와 비교 평가한다. Not Stemming Word Set 과 Stemming Word Set 에서는 자질 추출을 하지 않았기 때문에 균일 범주 유사도만을 이용하여 실험한다.

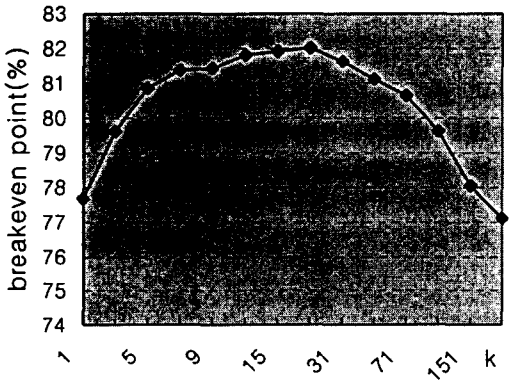
평가 방법으로는 문서 범주화에서 많이 사용하는 마이크로 평균 정확률(micro-average precision)과 마이크로 평균 재현율(micro-average recall)을 이용한다. 마이크로 평균이라는 의미는 k 개의 범주들과 d 개의 실험 문서들이 있을 경우, n=k×d 개의 범주 할당이 있을 수 있는데, 이들을 모두 하나의 단위로 생각하여 재현율(recall)과 정확률(precision)을 계산하는 방법이다. 또한 정확률과 재현율을 같이 평가할 수 있는 방법으로 breakeven point 라는 평가 방법을 사용한다. Breakeven point 는 재현율과 정확률이 같게 되는 경우 중에서 가장 높은 값을 의미한다.

4. 실험 및 평가

본 논문에서는 3 장에서 설명한 4 종류의 단어 및 자질 집합에서 균일 범주 유사도를 이용하여 실험을 하고 비

례 범주 유사도를 이용하여 2 종류의 자질 집합에서 실험한다. 각 집합이 k-nearest neighbor 문서 범주화에 미치는 영향을 마이크로 평균 정확률/재현율과 breakeven point 로 비교 평가하였다.

각 단어 및 자질 집합에서 k-nearest neighbor 방법의 smoothing parameter k 를 1, 3, 5, 7, 9, 11, 21, 31, 51, 71, 101, 151, 201 로 다르게 실험을 하였다. <그림 2>는 본 실험에서 가장 좋은 성능을 내는 Feature Set(2)에서 균일 범주 유사도를 이용한 경우, k 에 따른 breakeven point 의 변화를 보여주고 있다. <그림 2>에서 알 수 있듯이, 테스트 집합 Reuter-21578 에서는 k 의 변화가 문서 범주화의 성능에 크게 주지 않음을 알 수 있다. 일반적인 k-nearest neighbor 경우에 k 에 따라 성능 차이가 많은데 비하여 본 실험에서는 성능 차이가 4~5% 내외인 것으로 보아 Reuter-21578 에서 성능 향상이 상당히 어렵다는 것을 짐작할 수 있다.



<그림 2> k 에 따른 breakeven point 의 변화

<표 3>은 각 단어와 자질 집합에서 가장 좋은 결과를 보이는 k 와 breakeven point 를 나타내고 있다.

<표 3> 각 Set 과 범주 유사도에서 최상의 결과

Set 과 범주 유사도	Breakeven Point (%)	k
Not Stemming Word Set , 균일 범주 유사도	80.38	21
Stemming Word Set , 균일 범주 유사도	79.94	21
Feature Set(1), 균일 범주 유사도	81.05	21
Feature Set(1), 비례 범주 유사도	80.93	15
Feature Set(2), 균일 범주 유사도	82.01	21
Feature Set(2), 비례 범주 유사도	81.49	11

<표 3>에서 보듯이, 가장 좋은 결과를 보이는 Set 은 범주 자질 추출에 많은 노력을 들인 Feature Set(2)이다. 또한 균일 범주 유사도가 비례 범주 유사도보다 더 좋은 성능을 보이며 어간 추출을 하지 않은 경우가 더 좋은 결과를 나타내고 있다. 영어권의 경우, 문서 검색에서 어간 추출을 하지 않는 것이 일반적이다. 본 실험에서도 자질 추출을 Stemming Word Set 에서 하지 않고 Not Stemming Word Set 에서 했으면 보다 좋은 결과를 보였을 것이라고 기대한다.

Breakeven point 인 지점에서의 범주 유사도를 사용자 임계치로 하여 범주 할당에 사용한다면 82.01%의 마이크로 평균 재현율과 정확률을 가진다. 하지만, 본 실험에서 확인 문서 집합(validation document set)과 실험 문서 집합으로 구분하지 않고, 단지 확인 문서 집합의 역할을 하는 실험 문서 집합만으로 실험하였다. 그러므로, 실제로 breakeven point 의 범주 유사도를 임계치로 사용할 경우에 본 실험과 같은 결과를 가져 올 것인가에 대해서는 의문이다. 하지만, 일반적으로 확인 문서 집합의 성능과 실제 실험 문서 집합의 성능이 유사하기 때문에 임계치를 이용한 범주 할당의 성능을 82.01%로 보아도 무관할 것이다.

<표 4>는 균일 범주 유사도로 문서 당 n 개의 범주를 할당할 경우, 가장 좋은 결과를 보이는 k 와 마이크로 평균 정확률과 재현율이다.

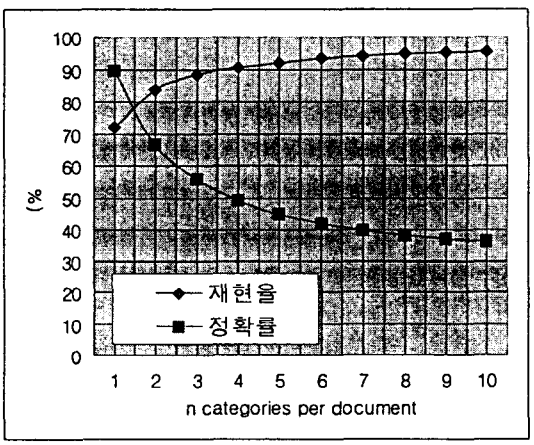
<표 4> 문서 당 n 개의 범주 할당 결과

Set , k	n	정확률(%)	재현율(%)
Not Stemming Word Set, k=31	1	88.28	70.93
	2	61.65	83.67
	3	49.14	88.08
Stemming Word Set, k=31	1	87.66	70.44
	2	60.89	83.53
	3	48.15	87.81
Feature Set(1), k=15	1	88.57	71.17
	2	63.90	83.50
	3	53.45	87.99
Feature Set(2), k=15	1	89.59	71.98
	2	66.44	83.70
	3	55.68	88.19

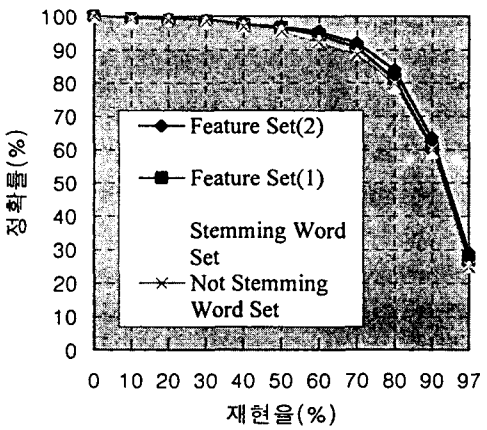
<표 4>의 결과에서 Word Set 의 경우는 가장 좋은 breakeven point 를 보이는 k=21 보다 문서 당 n 개의 범주 할당에서 좋은 결과를 보이는 k 값이 커지는 경향이 있고, 반대로 Feature Set 인 경우에서는 k 값이 적어지는 경향이 있다.

<그림 3>은 k=15 로 한 균일 범주 유사도로 Feature

Set(2)에서 n 개의 범주들을 문서에 할당할 경우, n 에 따른 정확률과 재현율의 변화를 보이고 있다. n =1 에서 상당히 높은 89.59%의 정확률을 보이고 있다. 실험 집합 자체에서 n=1 일 경우의 재현율이 (1/1.24[실험 집합에서의 문서 당 평균 할당 범주 수])×100 = 80.64%인 것을 감안하면, 본 실험에서 n=1 인 조건하에서는 71.98%의 재현율은 실제로 89%이다. 즉, 만약 문서 당 하나의 범주만이 할당하는 문제라면 재현율과 정확률이 약 90% 정도임을 알 수 있다. 최적 범주 할당을 위해서는 기본적으로 각 문서 당에 하나의 범주를 할당하고, 이에 벗어나는 적합한 범주 28% 정도를 임계치를 이용한 방법이나 기타 방법으로 찾는 것이 효율적일 것이다. 향후에 본 실험을 토대로 보다 정확한 범주 할당이 가능하도록 할 수 있을 것이다.



<그림 3> 할당 범주 수에 따른 정확률과 재현율



<그림 4> 각 Set 에 대한 결과

본 실험의 목적인 자질 추출이 가지는 성능 향상의 효과는 각 Set 에서 가장 좋은 결과들을 재현율 11 point 에 대한 정확률로 표현한 <그림 4>에서 쉽게 파악할 수 있다. <그림 4>에서 각 Set 에 대한 결과를 구별하기가 어렵듯이, 본 논문에서 제안한 자질 추출 후의 색인이 k-nearest neighbor 문서 범주화에서 성능 향상이 1~5% 정도로 미약하다는 것을 알 수 있다.

본 논문에서 자질 추출을 위해 사용한 기대 상호 정보 측도는 범주와 단어 간의 관련성이 큰가를 평가하기도 하지만, 반대로 그 범주와 상관이 없는 단어에 대해서도 상당히 높은 값을 준다. 그러므로, 기대 상호 정보 측도는 범주에 대하여 단어가 차별성을 가지는가에 대한 평가 기준으로서 역할을 한 것으로 보인다. 기대 상호 정보 측도를 이용한 자질 추출이 학습 문서에서 비차별성을 가진 단어들을 제거하는 효과로 시스템의 성능에서는 breakeven point 2%의 증가 효과를 보였지만, 본 논문에서 제안한 자질 추출의 역할을 충분히 하지 못 하기 때문에 비례 범주 유사도가 균일 범주 유사도보다 나쁜 결과를 보였다. 앞으로 자질 측정에 대한 새로운 방법을 고려하여 재실험을 해 보아야 할 것이다.

5. 결론

문서 범주화에서 문제가 되는 최적 범주 할당을 어떻게 할 것인가를 고려하기 위해서, 문서 당 하나의 범주를 할당할 경우에 좋은 결과를 보이는 방법을 선택하는 것이 중요하다. 본 논문에서는 이러한 결과를 보인 k-nearest neighbor 문서 범주화에서 성능 향상을 위해 단어와 범주 자질의 종류에 제한을 하는 방법을 제안하였고 실험하였다.

본 논문에서 수행한 자질 추출이 자질 추출의 효과는 크게 보이지 않지만, 자질 추출을 위해 사용한 상호 기대 정보 측도가 범주를 차별화하지 않는 단어를 제거하는 효과를 가져 breakeven point 가 약 2% 정도 향상하였다. 또한 breakeven point 는 82%이라는 상당히 좋은 결과를 보였다.

실험에서 문서마다 하나의 범주만을 할당하는 문제라고 가정하면, 제안한 문서 범주화 시스템은 약 90%의 마이크로 평균 재현율과 정확률을 가진다. 하지만 일반적으로 문서 범주화에서 문서는 하나 이상의 범주들을 가지므로 각 문서의 내용에 따라 적절하게 범주들을 할당할 필요가 있다. 본 논문의 연구 내용을 토대로 향후 연구에서는 이러한 문제를 해결할 것이다.

참고 문헌

1. 권오욱, "확률벡터와 메타 범주를 이용한 최적 문

서 범주화 모델," 석사학위 논문, 한국과학기술원, 1995.

2. Amit Singhal, Gerard Salton, Mandar Mitra, and Chris Buckley, "Document Length Normalization," Information Processing & Management, Vol. 32, No.5, pp.619-633, 1996.
3. Brij Masand, Gordon Linoff and David Waltz, "Classifying News Stories using Memory Based Reasoning," In Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval (SIGIR'92), pp. 59-65, 1992.
4. Chidanand Apté and Fred Damerau, "Automated Learning of Decision Rules for Text Categorization," ACM Transactions on Information Systems, Vol. 12, No. 3, ,July 1994, pp. 233-251.
5. C. J. Van Rijsbergen, "A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval," Journal of Documentation, Vol. 33, No. 2, June 1977, pp. 106-119.
6. David D. Lewis, "Representation and Learning in Information Retrieval," PhD thesis, Department of Computer Science; Univ. of Massachusetts; Amherst, MA 01003, 1992.
7. David D. Lewis and Marc Ringuette, "A comparison of two learning algorithms for text categorization," In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93, 1994.
8. Eric Brill, "A Simple Rule-Based Part of Speech Tagger," In Proceedings of the 3rd Conference on Applied Natural Language Processing, pp. 152-155, 1992.
9. Leah Larkey and W. Bruce Croft, "Combining Classifiers in Text Categorization," In Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR '96), pp. 289-297, 1996
10. Makoto Iwayama and Takenobu Tokunaga, "Cluster-Based Text Categorization: A Comparison of Category Search Strategies," In Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR'95), pp. 273-280, 1995.
11. Philip J. Hayes, Laura E. Knecht, and Monica J. Cellio, "A News Story Categorization System," In Proceedings of the 2nd Conference on Applied Natural Language Processing, pp. 9-17, 1988.
12. Yiming Yang, "Expert Network: Effective and Efficient Learning from Human Decision in Text Categorization and Retrieval," In Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval (SIGIR'94), pp. 13-22, 1994.