점진적으로 계산되는 분류정보와 링크정보를 이용한 하이퍼텍스트 문서 분류 모델

오효정, 임정묵, 이만호, 맹성현

충남대학교 컴퓨터과학과

{dol, jmlim, mhlee, shmyaeng}@cs.chungnam.ac.kr

A Hypertext Categorization Model Exploiting Link and Incrementally Available Category Information

Hyo-Jung Oh, Jeong-Mook Lim, Mann-Ho Lee, Sung-Hyon Myaeng
Department of Computer Science, Chungnam National University

요 약

본 논문은 하이퍼텍스트가 갖는 중요한 특성인 링크 정보를 활용한 문서 분류 모델을 제안한다. 하이퍼링크는 문서간의 관계를 나타내는 유용한 정보로서 링크를 통해 연결된 두 문서는 내용적으로 관련이 있어 검색에 도움을 준다는 것은 이미 밝혀진 바 있다. 본 논문에서는 이러한 과거 연구를 바탕으로 새로운 문서 분류 모델을 제안하는데, 이 모델의 주안점은 대상 문서와 링크로 연결된 이웃 문서의 내용 및 범주를 분석하여 대상 문서 벡터를 조정하고, 이를 근거로 문서의 범주를 결정한다. 이웃 문서에 포함된 용어를 반영함으로써 대상 문서의 내용을 확장 해석하고, 이웃 문서의 가용 분류 정보가 있는 경우 이를 참조함으로써 정확도 향상을 기한다. 이 모델은 이웃한 문서의 범주가 미리 할당되어 있지 않은 경우 용어 기반 분류 방법으로 가용 범주를 할당하고, 이렇게 할당된 분류 정보가 다시 새로운 문서의 범주를 결정할 때 사용됨으로써, 문서 집합 전체의 분류가 점진적으로 이루어지며 그 정확도를 더해 나가는 효과를 가져올 수 있다. 이러한 접근 방법은 일반 웹 환경에 적용할 수 있는데, 특히 하이퍼텍스트를 주제별로 분류하여 관리하는 검색 엔진의 경우 매일 쏟아져 나오는 새로운 문서와 기존 문서간의 링크를 활용함으로써 전체 시스템의 점진적인 분류에 매우 유용하다. 제안된 모델을 검증하기 위하여 Reuter-21578과 계몽사(ETRI-Kyemong) 자료를 대상으로 실험한 결과 18.5%의 성능 향상을 얻었다

1. 서론

최근 하이퍼텍스트를 대상으로 한 다양한 응용들이 붐을 일고 있다. 특히 하이퍼텍스트가 갖는 특성을 기존 모델에 접목시키려는 시도가 늘어나고 있다. 그 중에서도 하이퍼링크는 문서간의 관계를 나타내는 유용한 정보로서 링크를 통해 연결된 두 문서는 내용적으로 관련이 있어 검색에 도움을 준다는 것은 이미 밝혀진바 있다[11, 20].

이와 병행하여 하루에도 수만 건씩 쏟아져 나오는 하이퍼텍스트를 주제별로 분류하여 관리하기 위한 분류 모델에 관한 연구도 활발히 진행중이다. 분류 결과는 비단 문서를 관리하는 측면뿐 아니라 검색의 효율을 높인다거나 단어의 의미 중의성(ambiguity) 해소를 위해서도 사용된다[1]. 특히 검색 결과를 필터링(filtering) 하는데 문서의 분류 정보를 활용함으로써 사용자에게 보다 정확한 정보를 제공하려는 연구도 시도되고 있다[11, 18]. 이러한 현실을 볼 때 하이퍼텍스트의 특성을 고려하여 문서를 정확히 분류하는 분류기의 필요성은 매우 절실하다.

본 논문에서는 이러한 과거 연구를 바탕으로 새로운 문서 분류 모델을 제안하는데, 이 모델의 주안점은 대상 문서와 링크로 연결된 이웃 문서의 내용 및 범주를 분석하여 대상 문서 벡터를 조정하고, 이를 근거로 대상 문서가 어느 범주에 해당하는지를 결정한다. 이웃 문서에 포함된 용어를 반영함으로써 대상 문서의 내용을 확장 해석하고, 이웃 문서의 가용 분류 정보가 있는 경우 이를 참조함으로써 정확도 향상을 기한다.

개선된 방법의 주요한 특징은 다음과 같다.

- 한 문서에 링크로 연결된 문서의 분류 정보를 참조하여 이에 해당하는 범주의 가중치(weight)를 조절한다.
- 링크로 연결된 문서의 내용을 대표하는 용어를 반영하여 대상 문서 벡터를 조정한다.

논문의 구성은 다음과 같다. 2절에서는 문서 분류와 관련된 이전 연구에 대해 살펴보고, 3절에서는 본 논문에서 제안한 링크 기반 모델에 대해 설명하며, 4절에서는 이를 검증하기 위한 실험 및 결과에 대해 기술한다. 마지막으로 5절에서 결론을 맺도록 한다.

2. 관련연구

많은 양의 문서를 관리하고 이를 효율적으로 검색하기 위한 문서 분류 모델에 관한 연구는 이미 오래 전부터 계속되어 왔다. 그 중 대표적인 모델로는 크게 규칙 기반 모델(Rule-based Model)과 연역적 학습 모델(Inductive Learning Model), 검색을 활용한 모델로 나뉘어 진다.

먼저 규칙 기반 모델은 학습 문서들에서 나타나는 범주간의 구별된 규칙을 전문가가 찾아주거나 학습을 통해 추출된 규칙을 이용하여 문서를 분류하는 모델이다[6].

연역적 학습 모델로는 학습 문서에서 자질을 추출하여 이를 확률적인 접근방법으로 사용한 베이지언(Bayesian) 모델[5, 7, 8, 12, 13, 21]과 트리 구조로 표현하여 자질의 유무로 범주를 결정하는 결정 트리(Decision Tree) 모델[8], 학습 문서를 통해 생성된 양성 자질(positive feature)과 음성 자질(negative feature)을 벡터 공간으로 표현하고 이들 차이를 극명하게 하는 벡터(hyperplane)인 지원 벡터(support vector)를 찾는 SVM(Support Vector Machine)이 있다[19, 21].

이와는 달리 정보검색 관점에서 분류할 대상문서를 질의로 보고 이와 유사한 문서를 찾는 방법인 최근린법(K-nearest Neighbor)[13, 14, 21]과 적합성 피드백(relevance feedback)을 기초로 이를 분류에 응용한 Roccio 모델이 있다[9].

최근에는 이들 방법들을 비교하여 그 특성을 알아내거나[8, 13, 21], 각 방법의 장점을 복합하여 사용함으로써 성능 향상을 꾀하는 연구가 계속되고 있다[14]. 또한 하이퍼텍스트를 대상으로 MRF(Marcov Random Field)기법을 통해 하이퍼링크를 활용하거나[17, 18], 문서내의 구조적인 정보를 활용하는 연구[3]도 시도되고 있다.

본 논문은 분류하려는 대상 문서 뿐만 아니라 대상 문서와 링크로 연결된 문서들도 분류해야 하는 경우가 발생하므로 분류 속도가 가장 빠른 베이지언(Bayesian) 모델을 사용하였다.

3. 링크 기반 분류 모델

3.1. 링크 기반 분류 시스템

본 논문에서 제안하는 링크 기반 분류 모델은 하이퍼텍스트가 갖는 링크를 활용한다는 점에서 기존의 일반 문서 분류 모델과 차이가 있다. 일반 문서 분류 모델의 경우 문서에 출현하는 용어만을 사용하여 분류하는 반면 링크 기반 분류 모델은 하이퍼텍스트내의 링크 정보를 활용하여 대상 문서와 연결된 문서를 참조함으로써 분류의 정확도를 향상시키는 모델이다. 링크 기반 분류 과정은 기존 용어 기반 분류 과정을 사용하면서 문서 집합의 링크를 분석하고 활용하는 단계가 더 필요하다. 이는 문서 집합 내에 링크를 분리하여 저장하고 이를 관리하는 디지털 도서관의 환경에서 보다 효율적이다[4]. 그림 1은 링크 기반 분류 과정을 나타낸다.

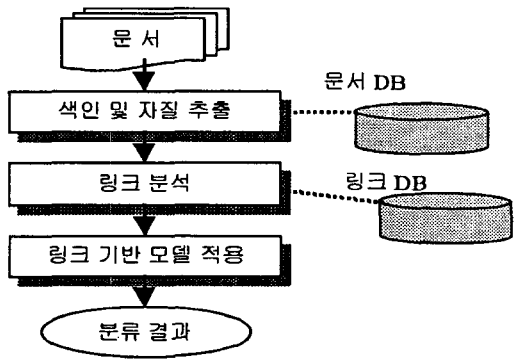

그림 1. 링크 기반 분류 시스템의 흐름도

3.2. 색인 및 자질 추출

이 단계는 문서 집합을 색인하고 분류에 필요한 자질을 추출하는 단계로 모든 분류에 공통으로 들어가는 단계이다. 일반적으로 베이지언(Baysien) 모델의 경우 단어보다는 구(phrase)나 단어 클러스터(word cluster)를 사용한 경우의 성능이 높지만[12], 본 논문에서는 단어를 대상으로 자질을 추출하였다.

자질을 추출하기 위해 기대 상호 정보 척도(EMIM: Expected Mutual Information Measure)를 이용하였으며 기대 상호 정보 척도(EMIM)가 임계치(threshold) 이상인 용어만을 추출하여 범주의 중심 벡터(centroid)를 표현한다. 범주에 대한 자질의 기대 상호 정보 척도를 구하는 식은 다음과 같다[7].

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

기대 상호 정보 척도란 용어가 주어진 범주와 얼마나 함께 나타나는 가에 대한 용어와 범주 간의 기대 상호 정보를 구하는 식이다. 식 (1)에서 $P(W_i=1, C_j=1)$은 해당문서에 단어 W_i가 출현하고 범주 C_j가 할당될 확률을 의미하고, $P(W_i=1)$는 문서에 단어 W_i가 출현할 확률, $P(C_j=1)$는 문서에 범주 C_j가 할당될 확률을 의미한다. 단어 W_i가 주어진 범주 C_j와 높은 빈도로 함께 나타나거나 전혀 다르게 나타나는 경우 기대 상호 정보 척도 값은 높다. 반대로 단어 W_i가 범주 C_j와 무관하게 출현한다면 낮은 값을 갖게 된다. 그러므로 기대 상호 정보 척도를 이용하여 범주를 나타내는 자질이 될 수 없는 단어를 제거 할 수 있다.

3.3. 베이지언 분류 모델

본 논문에서 사용한 분류 방법은 단순 베이지언(Naïve Bayesian) 모델을 이용하였다[5, 7, 12, 13]. 베이지언(Bayesian) 모델은 대상 문서가 각 범주에 속할 확률을 구해 가장 큰 확률 값을 갖는 범주에 그 문서를 할당하는 방법이다. 즉, 문서 d는 $P(c|d)$ 값이 최대가 되는 범주 c에 할당된다. 이를 수식으로 표현하면 다음과 같다[7, 12].

$$\begin{aligned} \underset{c}{Max}[P(c \mid d)] &= \underset{c}{Max}\left[\frac{P(c)P(d \mid c)}{P(d)}\right] \\ &= \underset{c}{Max}\left[P(c)\prod_{i=1}^{T} P(t_i \mid c)^{N(t_i|d)}\right] \end{aligned} \quad (2)$$

$$\begin{aligned} P(c)\prod_{i=1}^{T} P(t_i \mid c)^{N(t_i|d)} \\ \propto \quad \frac{\log P(c)}{n} + \sum_{i=1}^{T} P(t_i \mid d)\log\left(\frac{P(t_i \mid c)}{P(t_i \mid d)}\right) \end{aligned} \quad (3)$$

n = 문서의 길이

식 (2)에서 $N(t_i|d)$는 문서 d에서 용어 t_i가 출현하는 횟수(tf: term frequency)를 의미하고 T는 전체 문서 집합내의 용어의 수를 나타낸다. 일반적으로 범주 c에 용어 t_i가 많이 나타나고 문서 d에 용어 t_i의 빈도가 높으면 문서 d가 범주 c에 속할 확률이 높다. 그러나 식 (2)를 보면, 용어 t_i가 문서 d에 많이 출현할수록 즉, $N(t_i|d)$가 커질수록 오히려 $P(c|d)$ 값이 작아진다. 이러한 문제를 해결하기 위해 문서가 각 범주에 할당될 확률 값을 구하는 식을 식(3)과 같이 변형한다[15, 16].

식 (3)의 $P(c)$는 전체 학습문서 집합에서 해당 범주가 나타날 확률을 의미하고 $P(t_i|c)$는 해당 범주에서 용어 t_i가 출현할 확률, $P(t_i|d)$는 대상 문서에서 용어 t_i가 출현할 확률을 의미한다. 문서간의 차이를 나타내기 위해 Kulback-Leiber Divergence 값을 사용하였고, 각각의 범주에 대한 KL Divergence 값을 표현하기 위해 $P(t_i|c)$를 $P(t_i|d)$로 나눠주었다[16]. 각각의 확률 값은 다음과 같이 구할 수 있다[10].

$$c = vec(w_1, w_2, ..., w_T)$$

$$w_i = P(t_i \mid c) = \begin{cases} \text{최소확률값} & iff \quad N(t_i \mid c) = 0 \\ \dfrac{N(t_i \mid c) + 0.5}{Total\_N(c) + 0.5 \times T(c)} & iff \quad N(t_i \mid c) \neq 0 \end{cases}$$

$Total\_N(c)$ = 범주 c에 나타난 전체 용어 출현 빈도
$T(c)$ = 범주 c에 나타난 용어의 수

범주 c는 W_i의 벡터로 표현되는데, W_i는 범주 c에서 i번째 용어 t_i의 가중치(weight)를 의미하는 것으로 식 (3)의 $P(t_i|c)$에 해당한다. 이때 범주 c에서 용어 t_i가 한번도 출현하지 않은 경우 $P(t_i|c)$의 값이 0이 되므로 최소 확률값을 할당한다.

마찬가지로 문서 d는 용어 t_i의 가중치 W_i를 갖는 벡터로 표현되며 W_i는 식 (3)의 $P(t_i|d)$에 해당한다.

$$d = (w_1, w_2, ..., w_T)$$

$$w_i = P(t_i \mid d) = \frac{N(t_i \mid d) + 0.5}{Total\_N(d) + T(d)} \quad (4)$$

$Total\_N(d)$ = 문서 d에 나타난 전체 용어 빈도
$T(d)$ = 문서 d에 나타난 용어의 수

3.4. 링크 기반 분류 모델

3.1에서도 언급했듯이 하이퍼텍스트의 가장 큰 특징 중 하나가 링크를 갖는다는 점이다. 또한 두 문서가 링크로 연결되어 있을 경우, 이 두 문서는 서로 내용적으로 연관이 있다고 가정할 수 있다[11, 18, 20]. 본 논문에서는 이러한 가정하에, 기존의 베이지언 모델에 링크 정보를 활용하여 이웃 문서의 분류 정보를 반영할 수 있도록 식(2)와 식(3)을 다음과 같이 수정하였다. 식 (5)와 식 (6)에 나타난 $Neighbor(c)$는 대상 문서 d와 이웃한 문서들의 분류 정보를 반영한다.

$$\underset{c}{Max}[P(c \mid d)] = \underset{c}{Max}\left[P(c)\prod_{i=1}^{T} P(t_i \mid c)^{N(t_i|d)} \times Neighbor(c)\right] \quad (5)$$

$$P(c)\prod_{i=1}^{T} P(t_i \mid c)^{N(t_i,d)} \times Neighbor(c)$$

$$\propto \frac{\log P(c)}{n} + \sum_{i=1}^{T} P(t_i \mid d)\log\left(\frac{P(t_i \mid c)}{P(t_i \mid d)}\right) + \log(Neighbor(c))$$

n = 문서의 길이 (6)

또한, 대상문서와 이웃한 문서내의 용어를 참조하여 공통된 용어의 가중치를 높여줌으로써 대상문서의 내용을 확장 해석하였다. 즉, 식 (4)의 $d=(w_1, w_2, ..., w_T)$가 다음과 같이 수정되었다.

$$d = (w_1', w_2', ..., w_T')$$
$$w_i' = w_i + (Neighbor\_w_i \times ref) \quad (7)$$
$$Neighbor\_w_i = \text{이웃문서의 용어 } t_i\text{의가중치}$$

식 (7)의 *ref*는 링크의 반영비율을 의미하는 것으로 이웃 문서의 용어의 가중치를 반영하는 정도를 나타낸다. 이는 문서 집합의 성격에 따라 달라지며 실험에 의해 결정된다.

이웃 문서의 분류 정보를 반영하는 알고리즘은 대상문서 d와 링크로 연결된 문서로 이루어진 집합을 생성하고, 이 집합내의 분류 정보를 통해 *Neighbor(c)*를 구한다. 이때 이웃 문서의 범주가 미리 할당되어 있는 경우에는 이를 완전 신뢰하여 반영하고, 미리 할당되어 있지 않은 경우에는 용어 기반 분류를 이용하여 가용 범주(available category)를 할당한 뒤 이를 부분 신뢰하여 반영한다. 이와 같이 링크 기반 분류 모델은 대상문서와 이웃한 문서의 분류 정보가 없는 경우 이의 가용 범주를 추정한 후, 이를 다시 대상문서의 범주를 결정할 때 활용한다는 점에서 이웃 문서의 분류 정보가 없는 일반 웹 환경에도 적용될 수 있다는 장점이 있다. 또한 이전의 대상 문서에 할당된 분류 정보가 현재의 대상 문서의 범주 결정에 사용됨으로써 전체 문서 집합의 분류가 점진적으로 이루어지고 그 정확도가 높아진다. 상세한 분류 알고리즘은 다음과 같다.

[단계 1] 대상문서(d)와 링크로 연결된 문서(d')들의 집합 N 생성

[단계 2] 집합 N에 분류 정보 $cat(d')$와 신뢰도 $confidence(d')$ 할당
- 이웃 문서의 분류 정보가 있는 경우
  $cat(d')$는 문서 d'에 할당된 범주를 사용하고 $confidence(d')$값은 완전 신뢰도 값을 할당한다.
- 이웃 문서의 분류 정보가 없는 경우
  $cat(d')$는 용어 기반 분류(식(3))를 이용한 가용 범주를 할당하고, $confidence(d')$값은 부분 신뢰도 값을 이용한다.

[단계 3] 이웃 문서의 분류 정보 반영

$$Neighbor(c) = \frac{link\_cnt(c)}{Total\_link\_cnt} \times link\_avg\_conf$$

$link\_cnt(c)$=범주 c에 해당하는 링크의 수
$Total\_link\_cnt$=대상 문서 d와 연결된 링크의 수
$link\_avg\_conf$=링크의 평균 신뢰도

링크 기반 분류 모델은 [단계 2]를 통해 이웃 문서의 분류 정보에 대한 신뢰도를 달리 한다. 즉 문서내의 용어만을 이용해 할당한 분류 정보는 부분 신뢰하고, 미리 할당된 분류 정보나 링크 정보를 활용한 경우의 분류 정보는 완전 신뢰한다. 부분 신뢰의 정도는 용어 기반 분류기의 성능에 따라 결정된다. 링크 기반 분류 모델은 현재 할당된 분류 정보가 이후의 문서의 범주를 결정할 때 다시 활용되기 때문에 분류되는 문서의 순서에 따라 성능이 달라질 수 있다. 그러므로 실험 문서 집합의 링크 연결성을 분석하여 링크가 많은 문서 즉 주위 문서에 영향을 많이 주는 문서부터 미리 할당해 나간다면 가용 범주를 할당하는 경우가 줄어들기 때문에 보다 빠르고 정확한 분류가 이루어 질 수 있다.

4. 실험 및 평가

4.1. 실험 방법

본 연구에서 제안한 링크 기반 검색의 효율을 검증하기 위해 다음 2가지 실험을 실시하였다.

[실험 1] Reuter-21578과 계몽사(ETRI-Kyemong) 집합을 대상으로 기존 용어 기반 분류
[실험 2] 계몽사(ETRI-Kyemong) 집합을 대상으로 링크 기반 분류

[실험 1]은 본 논문에서 비교 대상으로 삼은 용어 기반 분류 모델의 객관성을 입증하기 위한 실험이고, [실험 2]는 본 논문에서 제안한 링크 기반 분류의 성능을 알아보기 위한 실험이다.

Reuter-21578 데이터는 이미 기존 연구[5, 7, 8, 12, 21]에서 많이 사용한 실험 데이터 집합으로 총 21, 578건의 문서와 135개의 분류 정보를 갖고 있다. 그러나 Reuter-21578 집합은 21,578개의 문서 중 하나 이상의 분류 정보가 할당된 문서의 수는 11,367개이고, 이 중 70%에 해당하는 문서가 10개의 범주에 해당하며, 10개 이하의 문서가 할당된 범주도 전체 33%로 분류 정보가 매우 불균형적이다[21]. 그러므로 본 시스템에서는 이 중 10,000개의 문서와 87개의 분류 정보를

사용하였으며 용어의 수는 불용어와 빈도가 1인 용어를 제거한 25,574개이다. 계몽사(ETRI-Kyemong) 데이터는 23,113건의 문서와 77개의 분류 정보, 182,844개의 링크를 갖고 있다. 23,113개의 문서 중 분류 정보가 할당된 문서는 21,525건이고 용어의 수는 빈도가 1인 용어를 제거한 49,578개이다. 실험 데이터에 대한 색인은 충남대학교 색인기를 사용하였으며[2], 자질 추출을 위한 기대 상호 정보 척도의 임계치는 0.0005를 사용하였다.

평가 방법으로는 재현도(Recall)와 정확도(Precision)를 함께 이용하여 성능을 나타내는 F-score를 사용하였다. F-score를 계산하는 방법은 다음과 같다[21].

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

여기서 β는 재현도(R)와 정확도(P)의 비중을 선택할 수 있게 하는 변수로 β>1이면 정확도의 비중을, β<1이면 재현도의 비중을 높게 두는 의미이다. 본 논문에서는 β=1로 하여 재현도와 정확도의 비중을 같게 두었다.

4.2. 실험 결과

4.2.1. 실험 1

앞에서도 언급했듯이 분류에 관한 이전 연구 중, Reuter-21578을 대상으로 한 연구들이 많이 있다[5, 7, 8, 12, 21]. [실험 1]은 본 논문의 결과에 대한 객관성을 입증하기 위해 실시한 실험으로 결과는 표 1과 같다.

표 1을 보면 Reuter-21578 집합의 경우 정확도에 비해 재현도가 낮게 나타난다. 그 이유는 실험 데이터에 할당된 분류 정보가 평균 1.1개로 중복 할당되어 있지만 본 논문에서 구현한 용어 기반 분류 모델은 각 문서 당 하나의 범주만 할당하기 때문이다. 반면 계몽사(ETRI_Kyemong) 집합은 중복 할당된 문서가 없으므로 재현도와 정확도가 같다.

표 1. Reuter-21578과 계몽사 집합에 대한 분류 결과

문서 집합	학습 문서	실험 문서	재현도	정확도	F-score
Reuter-21578	2,316	2,903	79.66	87.56	83.61
	7,097		80.92	89.01	84.95
계몽사 집합	4,309	3,611	73.08		
	17,225		78.67		

단위: 문서 갯수, %

분류기의 성능은 실험 데이터에 따라 달라지기 때문에 이전 연구 성능과 직접적인 비교는 할 수 없지만 Reuter-21578을 대상으로 한 실험 결과 본 논문에서 비교 대상으로 삼은 용어 기반 분류의 성능이 떨어지지 않음을 알 수 있다. 계몽사(ETRI_Kyemong) 집합에 대한 용어 기반 분류 결과(*F-score=78.67*)를 [실험 2]의 비교 기준(baseline)으로 사용한다. 또한 이를 가용 범주에 대한 부분 신뢰도로 이용한다.

4.2.2. 실험 2

[실험 2]는 본 논문에서 제안하는 링크 기반 분류의 성능을 알아보기 위한 실험으로 용어 기반 분류와 비교하였다. 이를 위해 링크 정보를 가지고 있는 계몽사(ETRI_Kyemong) 집합을 대상으로 다음 3가지 실험을 실시하였다.

[실험 2-1] 이웃 문서의 분류 정보만을 활용한 분류
[실험 2-2] 링크 기반 분류에 영향을 미치는 요인에 따른 분류
[실험 2-3] 이웃 문서의 분류 정보를 아는 정도에 따른 링크 기반 분류

[실험2-1]은 문서 집합에서 링크의 가용성 즉 신뢰도를 알아보기 위한 실험으로 문서내의 용어를 사용하지 않고 링크만을 이용하여 분류하였다. 즉 링크로 연결된 문서의 분류 정보만을 종합하여 대상문서의 범주를 결정하였다.

표 2. 이웃 문서의 분류 정보만을 활용한 분류 결과

분류 방법	재현도	정확도	F-score
$\sigma = 0.0$	63.21	66.85	65.03
$\sigma = 0.3$	61.64	81.64	71.64

$\sigma = sim(d_i, d_j)$

결과를 분석해보면 문서내의 모든 링크를 사용할 때(*65.03*)보다 링크의 시작(source) 문서와 종착(destination) 문서의 유사도(similarity)가 임계치 이상인 링크를 사용할 때(*71.64*)의 성능이 좋음을 알 수 있다. 이는 문서내의 링크가 모두 의미상으로 관계가 있어 생성된 것이 아니라 내용과는 관련 없이 만들어지는 링크가 있음을 의미한다. 그러므로 문서 집합 내의 링크 중 적합한 링크만을 사용하는 것이 성능 향상에 도움이 된다.

이러한 특성은 문서 집합의 특성에 따라 달라지며 본 논문에서는 유사도의 임계치로 0.3을 사용하였다. 표 2를 통해 용어를 사용하지 않고 링크만을 사용한 경우의 성능(*71.64*)이 용어 기반 분류의 성능(*73.08*)에 비해 크게 떨어지지 않는 점을 보아 본 논문에서 가정한 가설이 타당함을 입증한다

표 2를 보면 재현도와 정확도가 다른 것을 알 수 있다. 그 이유는 계몽사(ETRI-Kyemong) 문서 집합 내에 링크가 하나도 없는 문서가 있기 때문이다. 즉 문

서내의 용어를 전혀 사용하지 않고 오직 링크를 통해 이웃 문서의 분류 정보만을 활용하여 분류하기 때문에 링크가 없는 문서에는 어떤 범주도 할당되지 않는다.

[실험 2-2]는 링크 기반 분류에 영향을 주는 요인의 최적치를 찾기 위한 실험으로 성능에 영향을 미치는 다양한 변수를 조합해서 실험하였다. 링크 기반 분류의 성능에 영향을 미치는 요인으로는 다음과 같은 것들이 있으며 이에 따른 실험 결과가 표 3에 나와 있다.

- 링크로 연결된 이웃 문서 용어의 사용 여부
- 링크로 연결된 이웃 문서의 분류 정보 활용 여부
- 링크 신뢰도 정도

표 3. 다양한 요인에 따른 링크 기반 분류 결과

분류 방법			F-score
용어 기반 분류(Baseline)			78.67
링크 기반 분류	연결된 문서의 용어만 사용	σ = 0.0	76.30
		σ = 0.3	79.92
	연결된 문서의 범주만 사용	σ = 0.0	84.67
		σ = 0.3	88.14
	연결된 문서의 용어와 범주 사용 (w/σ = 0.3)	Random	88.97
		최다링크우선	89.27
	연결된 문서의 용어와 범주 사용 (w/σ = 0.0)	최다링크우선	83.26 *(-6.01)*

$\sigma = sim(d_i, d_j)$

결과를 분석해보면 문서내의 용어만을 사용한 경우(*78.67*)보다 신뢰할 만한 링크를 통해 문서를 확장 해석한 경우(*79.92*)의 성능이 다소 높음을 알 수 있다. 그러나 대상문서와 링크로 연결된 이웃 문서의 유사도가 임계치 이하인 문서의 용어까지 사용하게 되면 오히려 성능이 떨어지는데, 그 이유는 신뢰도가 떨어지는 링크로 연결된 문서에 대상 문서와는 상관없는 용어(noise)가 많이 포함되어 있기 때문이다.

표 3은 이웃 문서의 분류 정보를 사용한 경우(*89.27*)의 성능이 그렇지 않은 경우(*78.67*)에 비해 월등하다는 것을 보여준다. 그러나 이웃 문서의 용어 사용과 마찬가지로 링크로 연결된 문서와의 유사도가 적은 링크를 사용하면 성능 향상율이 저하된다. 그러므로 링크의 신뢰도는 전체 시스템의 성능에 매우 큰 영향을 끼친다.

직관적으로 링크가 많은 문서 즉 정보를 많이 갖고 있는 문서부터 범주를 할당하는 것이 성능 향상이 도움이 된다고 볼 수 있다. 그러나 실험결과 실험 집합을 구성할 때 링크 연결성을 분석하여 링크가 많이 있는 문서부터 분류하는 경우(*89.27*)가 그렇지 않은 경우(*88.97*)에 비해 다소 높긴 하지만 그 차이가 매우 적게 나타난 것으로 보아 분류 순서가 성능에 큰 영향을 미지친 않는다고 판단된다,

실험 결과를 종합해보면 신뢰할 만한 링크를 사용하여 이웃 문서의 용어를 통해 대상 문서를 보다 정확히 표현하고, 이웃 문서의 분류 정보를 반영하는 경우가 링크 기반 분류의 최적치 임을 알 수 있다. 반면 링크의 신뢰도를 판단하기 위해서는 링크의 시작 문서와 종착 문서 간의 유사도를 비교해야 하는 오버헤드(overhead)가 있다. 그러나 실험 결과 링크의 신뢰도를 판단하지 않는 경우(*83.26*), 최적의 경우(*89.27*)보다 6.01% 성능이 떨어지는 것을 볼 때 이는 꼭 필요한 요소이다. 유사도 계산을 위한 오버헤드를 줄이기 위해 처음 실험 문서 집합의 링크를 분석할 때 연결된 문서간의 유사도를 미리 계산하여 링크베이스를 구축한다면 보다 나은 효율을 얻을 수 있을 것이다[4].

[실험 2-3]은 대상 문서내의 용어만을 사용하는 기존 용어 기반 분류 방법과 본 논문에서 제안한 링크 기반 분류를 비교한 실험으로 본 논문의 궁극적인 목적이다. 일반적으로 분류 시스템의 성능은 학습 문서의 양에 따라 달라진다. 특히 링크 기반 분류는 이웃 문서의 분류 정보를 아는 정도에 따라 성능이 달라지므로 이를 변화하여 실험하였다. 실험 변수로 사용한 학습 문서의 양은 전체 문서 집합의 20%(4,309개)와 80%(17,225개)를 사용하였고, 이웃 문서의 분류 정보를 아는 정도로는 전혀 알지 못하는 경우에서부터 20%만 아는 경우, 50%, 80%, 모두 아는 경우까지 점차 증가시켜 실험해 보았다.

표 4. 용어기반 분류와 링크 기반 분류의 성능 비교

T-level(%)	용어기반	K-level(%)	링크 기반
20	73.08	0	80.62
		20	82.60
		50	84.10
		80	85.80
		100	86.60
80	78.67	0	83.37
		20	85.40
		50	87.38
		80	88.48
		100	89.27

T-level : 학습 문서의 양
K-level: 이웃 문서의 분류 정보를 아는 정도

표 4를 분석하면 용어 기반 분류보다 링크 기반 분류의 성능이 모든 경우에 있어 월등함을 알 수 있다. 또한 학습에 사용한 문서의 양이 많을수록 성능이 높고, 링크 기반 분류의 경우 이웃 문서의 분류 정보를 많이 알수록 성능이 향상되었다. 비록 학습 문서의 양이 적더라도 대상 문서와 이웃한 문서의 분류 정보

를 참조한 링크 기반 분류의 경우(*80.62*)가 많은 문서를 학습한 경우의 용어 기반 분류(*78.67*)보다 성능이 더 좋다. 이는 분류기를 만들 때 사람의 힘을 덜 필요로 하고도 좋은 성능의 분류기를 생성할 수 있음을 의미한다. 현재 많은 검색 엔진이 하이퍼텍스트를 주제별로 관리 하고 있다. 이러한 현실에서 본 논문에서 제안한 링크 기반 분류 모델은 매일 같이 쏟아져 나오는 새로운 문서와 주제별로 관리된 기존 문서간의 링크를 활용함으로써 전체 시스템의 점진적인 분류에 매우 유용하다. 특히 이웃 문서의 분류 정보를 전혀 모르는 경우에도 용어 기반 분류에 비해 성능이 매우 높게 나타나는데, 이는 이웃 문서의 분류 정보가 미리 할당되지 않았으면 가용 범주를 할당함으로써 전체 실험 집합을 점진적으로 분류해 나가기 때문이다. 이를 통해 링크로 연결된 문서의 분류 정보가 미리 정해 있지 않은 일반 웹 환경에서도 좋은 효과가 기대된다.

5. 결론 및 향후 연구 방향

본 논문은 하이퍼텍스트의 중요한 특성인 링크를 이용하여 대상 문서를 확장 해석하고 이웃 문서의 분류 정보를 활용하는 링크 기반 분류 모델을 제안하였다. 실험 결과 문서 내의 용어만을 사용하는 기존의 용어 기반 분류 모델에 비해 링크 기반 분류 모델이 18.5%의 성능 향상을 얻을 수 있었다.

또한 이웃 문서의 분류 정보를 전혀 모르는 경우를 대비함으로써 일반 웹 환경에 적용 가능성을 보였고, 적은 문서를 학습하고도 이웃 문서의 분류 정보 활용을 통해 높은 성능을 보여줌으로써 분류기 생성을 보다 자동화할 수 있음을 보였다.

향후 연구 방향으로는 하이퍼텍스트의 또 다른 중요한 특성인 구조정보를 반영할 수 있도록 제안된 링크 기반 분류 모델을 개선하고 이와 유사한 연구에 대한 비교 실험을 통해 그 효과를 검증할 예정이다. 또한 링크의 분류 정보를 정보검색에 활용하여 그 효율을 높이는 연구도 수행할 예정이다.

참고 문헌

[1] 이호, "단어 의미 중의성 해결을 위한 분류 정보 모형", *고려대학교 박사학위 논문*, 1999.

[2] 장동현, 맹성현, "효율적인 색인어 추출을 위한 복합명사 분석방법", *제 8회 한글 및 한국어 정보처리학술대회*, 1996.

[3] 정성화, 이종혁, "문서 구조 정보에 기반한 웹 페이지 범주화 모델", *제 10회 한글 및 한국어 정보처리학술대회*, , 1998.

[4] 조은일, 임정묵, 오효정, 이만호, 맹성현, "CORBA와 JAVA를 사용한 에이전트 기반 디지털 도서관 프로토타입 구현", *한국정보과학회 춘계 학술대회*, 1999.

[5] Andrew McCallum and Kamal Nigram, "A comparison of Event Models for Naïve Bayes Text Classification", *AAAI '98 Workshop on Learning for Text Categorization*, 1998.

[6] Chidanand Apté, Fred Damerau, and Sholom M. Weis, "Towards language independent automated learning of text categorization models", *Proc. Of the 17th annual international ACM-SIGIR*, 1994.

[7] David D. Lewis, "Representation and Learning in Information Retrieval", *Ph.D thesis, Dep. Of Computer Science, Univ. of Massachusetts*, 1992.

[8] David D. Lewis and Marc Ringuette, " A comparison of Two Learning Algorithms for Text categorization", *Proc. Of the 3rd Annual Symposium on Document Analysis and Information Retreival*, 1994.

[9] David D. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka, "Training Algorithms for Linear Text Classifier", *Proc. Of the 19th annual international ACM-SIGIR '96*, 1996.

[10] Hang Li and Kenji Yamanishi, "Document Classification Using a Finite Mixture Model", *The Association for Computational Linguistics, ACL '97*, 1997.

[11] Jeong-Mook Lim, Hyo-Jung Oh, Sung-Hyon Myaeng, and Mann-Ho Lee, M. H., "Improving Efficiency with Document Category Information in Link-based Retrieval", *Proc. Of the International Workshop on IRAL'99*, 1999. To appear

[12] L. Douglas Baker and Andrew K. Maccallum, "Distributional Clustering of Words for Text Classification", *Proc. Of the 21th annual international ACM-SIGIR*, 1998.

[13] Leah S. Larkey. "Automatic Essay Grading Using Text Categorization Techiniques", *Proc. Of the 21th annual international ACM-SIGIR*, 1998.

[14] Leah S. Larkey and W. Bruce Croft, "Combining Classifiers in Text Categorization", *Proc. Of the 19th annual international ACM-SIGIR '96*, 1996

[15] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Sean Slattery, "Learning to Extract Knowledge from the World Wide Web. *Proc. Of the International Workshop on AAAI '98*, 1998.

[16] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Sean Slattery, "Learning to Extract Symbolic Knowledge from the World Wide Web", *Internal Report, School of Computer Science, CMU*, CMU-CS-98-122, September 1, 1998

[17] Soumen Chakrabarti, Byron Dom, and Piotr Indyk, "Enhanced hypertext categorization using hyperlinks", *Proc. of International Conference on SIGMOD '98*, 1998

[18] Soumen Charkrabarti, Martin Van den Berg, and Byron Dom, "Focused Crawling: A New Approach to Topic-spefiic Web Resource Discovery", *Proc. Of*

International Conference, WWW '99, 1999.

[19] Thorsten Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", *Proc. Of European Conference on Machine Learning, ECML'98,* 1998.

[20] Won-Kyun Joo and Sung-Hyoun Myaeng, "Improving Retrieval Effectiveness with Link Information", *Proc. Of the International Workshop on IRAL'98,* 1998.

[21] Yiming Yang and Xin Liu, "A re-examination of text categorization methods", *Proc. Of the 22th annual international ACM-SIGIR '99,* 1999.