k-NN 분류 알고리즘과 객체 기반 시소러스를 이용한 자동 문서 분류

(Automatic Document Classification Based on k-NN Classifier and Object-Based Thesaurus)

방 선 이 [†] 양 재 동 [††] 양 형 정 [†††]
(Sun-Iee Bang) (Jae-Dong Yang) (Hyung-Jeong Yang)

요 약 기존의 통계적인 기법과 기계학습 기법 등을 이용한 자동 문서 분류는 주로 문서 벡터만으로 분류기를 학습하여 분류를 행하기 때문에 특정 범주로 문서를 분류하는데 명확치 않은 경우가 빈번히 발생하여 일정 수준 이상의 정확도를 얻는 데에는 한계를 보이고 있다. 이러한 문제를 해결하기 위해 본 논문에서는 기존 문서 분류 알고리즘에 범주 간의 관련성을 반영하여 분류를 시행하는 방법을 제안한다. 이 방법은 간단한 알고리즘에 비해 좋은 성능을 보이고 있는 k-NN 분류 알고리즘을 이용하여 일차적인 문서 분류를 수행한 후 특정 범주로 분류하기가 명확치 않을 경우, 객체 기반 시소러스에서 제공되는 범주들 간의 일반화 관계, 집성화 관계, 연관화 관계 그리고 인스턴스 관계를 이용하여 문서가 할당될 범주를 결정함으로써 자동 문서 분류의 정확도를 향상시킬 수 있다. 본 논문에서 제안된 방법으로 실험한 결과 k-NN 분류 알고리즘의 분류 결과에 비해 재현율은 유지되면서 최고 13.86% 까지 정확률이 향상되었다.

키워드 : 자동 문서 분류, 최근접 이웃 분류, 객체 기반 시소러스

Abstract Numerous statistical and machine learning techniques have been studied for automatic text classification. However, because they train the classifiers using only feature vectors of documents, ambiguity between two possible categories significantly degrades precision of classification. To remedy the drawback, we propose a new method which incorporates relationship information of categories into extant classifiers. In this paper, we first perform the document classification using the k-NN classifier which is generally known for relatively good performance in spite of its simplicity. We employ the relationship information from an object-based thesaurus to reduce the ambiguity. By referencing various relationships in the thesaurus corresponding to the structured categories, the precision of k-NN classification is drastically improved, removing the ambiguity. Experiment result shows that this method achieves the precision up to 13.86% over the k-NN classification, preserving its recall.

Key words : Document Classification, Nearest Neighbor Classification, Object-Based Thesaurus

1. 서 론

최근 인터넷의 보급으로 온라인상에서 획득할 수 있는 문서 정보의 양이 급격하게 증가함에 따라 방대한 문서 데이터로부터 유용한 정보를 효과적으로 획득할 수 있는 자동 문서 분류에 대한 필요성이 급격히 부각되고 있다. 자동 문서 분류란 문서의 내용을 파악하여 미리 정의되어 있는 범주 중 하나로 문서를 자동 할당하는 것이다. 기존의 문서 분류는 전문가를 통한 수작업으로 이루어지는 수동 분류 방식으로 정확성은 높은 반면에 많은 시간과 노력, 비용이 들게 되는 문제점을 안고 있으며, 웹과 같이 새로운 문서 데이터가 끊임없이 생성되는 환경에서는 동적인 분류가 이루어지기 어려운 단점이 있다. 예를 들어, 의학 관련 기사들을 1000만 건 이상 수록하고 있는 MEDLINE의 경우, 매주 7000~8000건의 기사가 추가되는데 이들을 수동 분류 방식으로 주기적인 분류를 수행하는 것은 많은 비용이 요구된

· 본 연구는 한국과학재단 지역대학우수과학자 지원연구(R05-2003-000-11986-0) 지원으로 수행되었음
† 학생회원 : 전북대학교 컴퓨터통계정보학과
slbang@chonbuk.ac.kr
†† 비 회 원 : 전북대학교 전자정보공학부 교수
jdyang@chonbuk.ac.kr
††† 비 회 원 : 카네기멜론대학 컴퓨터과학과 연구원
hjyang@cs.cmu.edu
논문접수 : 2004년 3월 11일
심사완료 : 2004년 7월 9일

다[1,2].

이에 자동 문서 분류 알고리즘에 대한 연구가 활발히 진행되었다. 대표적인 알고리즘으로는 베이지안 확률분류(bayesian classifier)[3], 결정 트리(decision tree)[4], 최근접 이웃분류(k-nearest neighbor classification)[5], 규칙 학습(rule learning algorithm) [6], 신경망(neural networks)[7], 퍼지 개념을 이용한 알고리즘[8] 등이 있으며, 최근에는 SVM(support vector machine)[9]을 이용한 문서 분류 방법이 제안되었다[10].

이 자동 문서 분류 알고리즘들은 대부분 문서에 표현되는 단어로부터 문서 벡터를 생성하고 벡터화된 훈련 문서들을 예제로 사용하여 분류기를 학습함으로써 관련 범주를 할당한다. 그러나 문서 벡터만으로 분류기를 학습하여 분류를 행하는 방법은 특정 범주로 문서를 분류하는데 애매모호한 경우가 빈번히 발생하기 때문에 일정 수준 이상의 정확도를 얻는 데에는 한계를 보이고 있다[11,12]. 예를 들어, 'LCD', 'CRT', 'PDP' 및 이들과 관련된 단어들이 많이 나타나는 디스플레이 장치 전반에 대해 설명하는 문서가 주어졌을 때 단순히 문서 벡터만을 가지고 최근접 이웃 분류 알고리즘에 의해 단일 범주 분류를 수행한다면 근접한 훈련 문서들이 속하는 범주인 LCD, CRT, PDP 범주 중 하나에 할당될 것이다. 그러나 주어진 문서와 이 범주들의 속할 정도가 비슷할 경우 모호성이 발생하여 정확성이 떨어지게 된다. 이 때 범주 간의 관련성을 이용한다면 후보범주 중 보다 정확한 범주에 이 문서를 할당할 수 있을 것이다.

본 논문에서는 최근접 이웃(k-NN)[5,13-15] 분류 알고리즘을 이용한 자동 문서 분류 방법에 객체 기반 시소러스로부터 제공되는 범주들 간의 관련성을 이용하여 정확도를 향상시키는 새로운 문서 분류 방법을 제안한다. k-NN 분류 알고리즘이 사용된 이유는 분류하고자 하는 문서와 근접한 훈련 문서를 이용하는 예제 기반 방법으로서 단순한 알고리즘에 비해 비교적 좋은 성능을 보이는 것으로 평가되기 때문이다[10,14]. 본 제안 방법은 먼저 k-NN 분류 알고리즘에 의해 문서 분류를 수행하고 분류 범주가 명확하지 않을 경우, 객체 기반 시소러스에서 제공되는 범주들 간의 일반화 관계, 집성화 관계, 연관화 관계 그리고 인스턴스 관계를 이용하여 문서가 할당될 범주를 결정함으로써 자동 문서 분류의 정확도를 향상시킬 수 있다. 본 방법으로 실험한 결과 k-NN 분류 알고리즘 분류 결과의 재현율을 유지하면서 정확률은 최고 13.86% 까지 성능이 향상되었다.

본 논문의 구성은 다음과 같다. 2장 선행 연구에서는 자동 문서 분류의 개관과 k-NN 분류 알고리즘을 이용한 문서 분류에 대해서 살펴보고, 3장에서는 객체 기반 시소러스를 사용하여 k-NN 분류 알고리즘을 이용한 문서 분류에서 생기는 모호성을 해결하는 방법을 기술한다. 4장에서는 실험 결과를 보이고 마지막으로 5장에서는 결론 및 향후 과제를 제시한다.

2. 선행 연구

2.1 자동 문서 분류의 개관

자동 문서 분류는 일반적으로 자질추출과정과 분류과정으로 나눌 수가 있다[13,6,10]. 자질추출과정은 전처리 과정과 자원축소과정을 거쳐 문서에 출현하는 단어들을 바탕으로 문서벡터를 형성한다. 전처리 과정은 문서로부터 태그와 불용어를 제거하고 형태소 분석 및 어간화 작업을 통해 특정 용어들을 추출한다. 전처리 과정을 거쳐서 추출된 자질을 벡터형식으로 표현하기에는 문서 표현은 차원이 크기 때문에 단어의 문서에 대한 출현 빈도수 등을 고려하여 차원을 축소하는 과정을 거친다. 차원축소과정을 거쳐 추출된 단어는 문서를 어느 정도 대표하는지에 대한 가중치와 함께 문서를 벡터 형식으로 표현하는데 사용된다. 단어에 대한 가중치를 계산하는 방법은 이진 자질(boolean weighting), 단어 빈도(tf: word frequency weighting), 역문서 빈도(idf:inverse document frequency weight), 단어 빈도와 역문서 빈도의 곱(tfidf weighting)등이 있다[10]. 본 논문에서는 실험적으로 가장 높은 성능을 보이는 tfidf 기법을 이용하여 문서벡터를 형성한다.

문서의 수를 N이라 하면 문서 집합 $D=d_1, d_2, \cdots, d_N$에 대해 추출된 단어집합을 $W=t_1, t_2, \cdots, t_{N_w}$라 하자. tfidf 기법을 이용하여 문서벡터를 생성하면 다음과 같다.

$$tfidf_{ik} = f_{ik} \times \log(\frac{N}{n_i})$$

f_{ik} : 문서 k에 대한 단어 i의 빈도수

n_i : 전체 문서 집합에서 단어 i가 나온 문서의 수

tfidf 가중치 a_{ik}는 코사인정규화(cosine normalization)를 이용하여 $tfidf_{ik}$를 [0,1]의 값으로 정규화하여 나타낸다[10].

$$a_{ik} = \frac{tfidf_{ik}}{\sqrt{\sum_{s=1}^{N_w} tfidf_{sk}^2}}$$

문서 $d_k \in D$에 대한 문서벡터는 다음과 같다.

$$word(d_k) = \{t_1/a_{1k}, t_2/a_{2k}, \cdots, t_{N_w}/a_{N_wk}\}.$$

문서벡터를 기반으로 분류 알고리즘을 적용하여 문서를 분류하는 과정은 훈련 문서를 이용하여 분류기를 학습하는 학습과정과 실험 문서에 대해 분류기를 통해 범주를 할당하는 할당과정으로 이루어진다[10,16]. 분류과정은 크게 훈련 문서로부터 규칙을 생성해 내는 규칙

기반 알고리즘[6]과 훈련 문서로부터 추출한 자질과 범주의 확률적 모형에 입각한 베이지언 모델(bayesian model)[3], 트리 구조로 표현하여 자질의 유무로 범주를 결정하는 결정 트리(decision tree)[4], 훈련 문서로부터 음성 자질과 양성 자질을 벡터공간으로 표현해 최적 분리 경계면을 제공하는 SVM[9] 등과 같은 연역적 학습 알고리즘, k-NN 분류 알고리즘과 같이 입력문서와 가장 유사한 문서를 찾는 예제 기반 알고리즘[5] 등이 적용되고 있다[10,16,17]. 국내에서도 최근 들어 자동 문서 분류에 대한 관심이 높아지면서 이러한 기존 알고리즘들을 바탕으로 한국어 문장 구조의 특징을 고려한 분류 등의 연구가 수행되고 있다[15,18]. 다음 절에서는 분류 과정에서 적용될 수 있는 분류 알고리즘들 중 본 논문에서 채택한 k-NN 분류 알고리즘[5,14]에 대해 보다 자세히 살펴보도록 한다. k-NN 분류 알고리즘은 지금까지 개발된 분류 알고리즘들 중 가장 간단하면서도 비교적 좋은 성능을 보이는 것으로 평가되고 있다.

2.2 k-NN 분류 알고리즘을 이용한 문서 분류

k-NN 분류 알고리즘은 전문가에 의해 이미 분류되어 있는 훈련 문서들로부터 단어들의 출현 빈도 정보를 추출하여 문서 벡터를 형성하고 새로운 문서 d에 대해 훈련 문서 벡터와의 유사도를 따져 k개의 가장 가까운 문서를 정한 뒤, 이를 이용하여 d가 속할 범주를 예측한다[13,14,15]. 즉, k개의 이웃하는 문서들이 할당되는 범주들을 고려해서 이들과 가장 관련이 높은 범주로 d를 할당한다.

k-NN 분류 알고리즘을 이용하여 문서 분류를 하기 위해서는 다음과 같이 몇 가지 정의가 필요하다.

정의 1. 전체 범주가 $C=\{c_1, c_2, \cdots, c_m\}$이고 전체 문서 집합이 D일 때, 새로운 문서 d에 대해 정해지는 k개의 최근접 이웃 문서 $k-NN(d)$는 다음과 같다.

$$k-NN(d) = \{ d_j^{k-nn} \in D \mid j = 1, 2, \cdots, k \}$$

정의 2. 범주 c_i, $i=1,2,\cdots,m$에 대한 $d^{k-nn} \in k-NN(d)$의 할당여부 $c_i(d^{k-nn})$는 다음과 같이 정의된다.

d^{k-nn}가 범주 c_i, $1 \le i \le m$에 속하면 $c_i(d^{k-nn})=1$이고, 그 외의 경우, $c_i(d^{k-nn})=0$ 이다.

[정의 2]를 이용하여 각 범주 c_i, $i=1,2,\cdots,m$에 문서 d가 속할 정도 w_i를 구해보면 다음과 같다.

$$w_i = \sum_{j=1}^{k} c_i(d_j^{k-nn}), \quad i=1,2,\cdots,m$$

$c_i(d)$는 문서 d가 속할 정도 w_i와 함께 $c_i(d, w_i)$로 나타낼 수 있다.

k-NN 분류 알고리즘에서는 단순히 w_i가 가장 높은 범주에 문서를 할당한다[15]. 그러나 본 논문에서는 범주 간 관련성을 이용하여 문서가 할당될 범주의 모호성 해결방안을 제안하므로 k-NN 분류 알고리즘에 의해 문서 d가 속할 정도가 미리 정해 놓은 임계치 α 이상으로 비슷한 범주들을 후보범주라 하고 d의 최종 범주를 후보범주에서 선택하는 변경된 k-NN 분류 알고리즘을 사용한다.

정의 3. α를 임계치라고 할 때, 문서가 속할 정도 $w_i \ge \alpha$에 의한 문서 $d \in D$의 후보범주 $C_\alpha(d)$는 다음과 같이 정의된다.

$$C_\alpha(d) = \{ c_i(d, w_i) \mid c_i \in C, \ w_i \ge \alpha, \ 1 \le i \le m \}$$

$C_\alpha(d)$의 차수는 $|C_\alpha(d)|$이고, $w_i \ge \alpha$가 중요하지 않을 때는 $C_\alpha(d)$를 $C(d)$로 표기하도록 한다.

$|C(d)|=1$인 경우는 문서 d가 $c_i \in C(d)$범주에 속하게 되고, $|C(d)|>1$인 경우 k-NN 분류 알고리즘에서는 단순히 후보범주들 중 속할 정도가 가장 높은 범주에 문서를 할당하게 된다. 그러나 이는 비슷한 속할 정도를 가지고 있는 다른 후보범주와 속할 정도가 임계값 보다 작지만 어느 정도 관련이 있는 다른 범주들을 모두 무시하는 단점을 가지고 있다. 이들 범주들 사이의 관련성은 후보범주들 중 d가 최종 할당될 범주를 결정하는데 중요한 단서로 사용될 수 있다.

예를 들어, k=11, α=4이고 범주 $c \in C=\{c_1, c_2, \cdots, c_8\}$일 때, 문서 $d \in D$의 $k-NN(d)$가 다음과 같다고 하자.

$$k-NN(d)=\{128, 68, \cdots, 23, 9\}$$

$d^{k-nn} \in k-NN(d)$에 대해 $c(d_j^{k-nn})$, $j = 1, 2, \cdots, 11$를 구해보면, 그림 1과 같이$\{c_1(128)=1, c_4(68)=1, \cdots, c_4(23)=1, c_5(9)=1\}$이고, $c_i(d, w_i)$, $i=1,2,\cdots,8$는 $c_1(d,4)$, $c_2(d,0)$, $c_3(d,0)$, $c_4(d,5)$, $\cdots$, $c_8(d,0)$이다.

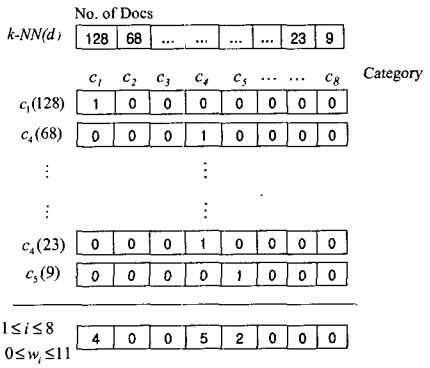

그림 1 k-NN 분류 알고리즘을 이용한 문서 분류의 예

α=4에 대해 후보범주는 $C(d)=\{c_1, c_4\}$이므로 k-NN 분류 알고리즘에서는 속할 정도가 가장 높은 c_4의 범주

에 문서 d를 할당한다. 그러나 만약, 실제 문서가 c_1의 범주에 속하는 경우라면 이 방식은 속할 정도가 4인 c_1 범주와 후보범주로는 설정되지 않았지만 속할 정도가 임계값 보다 조금 낮은 c_5범주를 모두 무시한 분류로 정확률이 떨어지는 결과를 낳는다. 이러한 문제점은 주어진 문서가 c_1과 c_4의 범주에 속할 정도가 같은 경우라도 그대로 남게 된다. 즉, k-NN 분류 알고리즘은 c_5 범주를 고려치 않고, 어느 범주에 문서를 할당하게 될지 모호한 상태에서 c_1과 c_4범주 모두에 문서를 할당함으로써 정확성을 떨어뜨리게 된다. 다음 절에서는 k-NN 분류 알고리즘의 정확도를 향상시키기 위해 본 논문에서 채택한 범주들 간의 관련성을 제공하는 객체 기반 시소러스에 대해 설명한다.

2.3 객체 기반 시소러스[19]

객체 기반 시소러스는 기존의 시소러스에 객체 지향 패러다임을 적용한 시소러스로 객체 관점과 관계 관점에 따라 의미적 관점을 제공한다. 즉, 모든 시소러스 개념은 객체로 간주되며, 일반적인 의미의 개념 객체는 보다 구체적인 의미의 개념 객체나 인스턴스 객체를 하위 객체로 가진다. 여기서 인스턴스 객체는 개념 객체의 실제 예의 의미를 가진 객체로서 실세계의 많은 개념들이 인스턴스로 분류될 수 있다. 따라서 하나의 개념 객체는 자신이 소유한 모든 인스턴스들을 대표하게 된다. 특히, 시소러스 개념들은 그 응용 도메인에 따라 인스턴스 또는 개념 객체로 그 단위가 결정된다. 예를 들어, 그림 2의 객체 기반 시소러스 구조에서 TFT-LCD 객체는 LCD 개념 객체에 대해 구체적인 개념을 지닌 인스턴스 객체이다.

관계 관점에서 본 객체 기반 시소러스는 기존의 시소러스에서 BT(Broader Term)/NT(Narrower Term)관계로 표현되었던 시맨틱이 일반화(generalization)와 클래스화(specialization)로 재정의되고, 막연한 관련성만을 표현했던 RT(Related Term)관계는 집성화(composite) 관계나 연관화(association) 관계로 세분되어 보다 구체적인 시맨틱으로 정의된다. 따라서 하나의 최상위 개념은 수직적으로 일반화/클래스화 관계에 의해 객체 계층을 형성하고, 집성화, 연관화 관계는 수평적인 관계 계층을 기술하는 속성으로서 하위 계층의 객체들에게 상속되어 개념들 간에 존재하는 시맨틱을 보다 구조적이고 체계적으로 표현할 수 있는 도구로 이용된다. 이러한 의미적 관계에 대한 관계 정보는 관련정도를 부여하여 나타낸다. 개념 간의 관련정도는 도메인에 따라 각각 다르게 정의될 수 있는데, 본 논문에서는 일반화/클래스화 관계에 0.9, 집성화 관계에 0.8, 연관화 관계에 0.7의 묵시적 관련정도를 할당한다.

3. k-NN 분류 알고리즘과 시소러스를 이용한 문서 분류

3.1 범주의 구조

본 논문에서는 먼저 k-NN 분류 알고리즘을 이용하여 문서를 분류한 후 모호성이 발생할 경우 객체 기반 시소러스에서 제공되는 범주 간의 관련성을 이용하여 모호성을 해결함으로써 정확성을 향상 시킨다. 시소러스 구조를 반영하기 위해 전체 범주 집합 C는 시소러스 Th의 레벨 수를 n이라고 할 때 다음과 같이 정의한다.

정의 4. $C = \{c_{i_1 i_2 \cdots i_l} \mid i_l \neq 0 \in I^+, l = 1,2,\cdots,n\}$에 대해 c_{i_1}은 i_1번째의 최상위 개념이고, $c_{i_1 i_2 \cdots i_l}$은 $2 \leq l \leq n$인 $l-1$단계에서의 $c_{i_1 i_2 \cdots i_{l-1}}$번째 개념의 i_l번째 하위 개

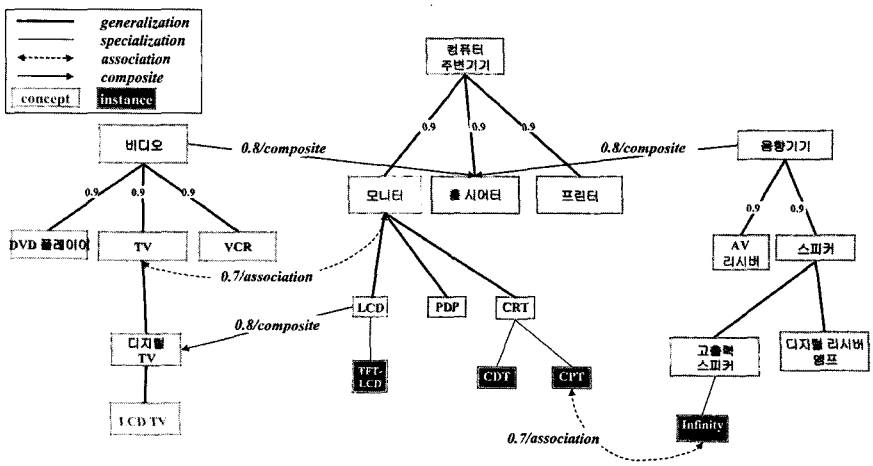

그림 2 객체 기반 시소러스

념이다. c_{i_1}가 최상위 개념임을 강조하기 위해서는 $c_{i_1}^{top}$로 정의한다. 시소러스의 개념 $c_{i_1 i_2 \cdots i_l} \in C$에 대한 인스턴스는 $I(c_{i_1 i_2 \cdots i_l})$로 정의한다.

시소러스 구조에 대응되는 범주 구조는 그림 3과 같이 도식화할 수 있다.

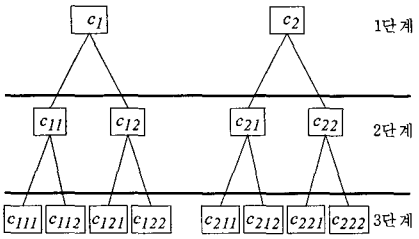

그림 3 시소러스 구조에 대응되는 범주 구조의 한 예

시소러스 구조를 분류상 범주 구조로 반영함에 있어 본 논문에서는 객체 기반 시소러스에서 사용되는 개념과 범주를 동일한 용어로 사용한다. 그러나 객체 기반 시소러스의 인스턴스들은 그 수가 너무 많고 지나치게 구체적이므로 범주와 대응된다는 것은 비현실적이다. 따라서 범주 c의 인스턴스 집합 $\{I(c) | c \in C\}$는 범주 c를 특성화하는 세부 범주 용어 사전(local dictionary)으로 사용한다. 예를 들어, 다음과 같은 구조의 범주를 살펴보자.

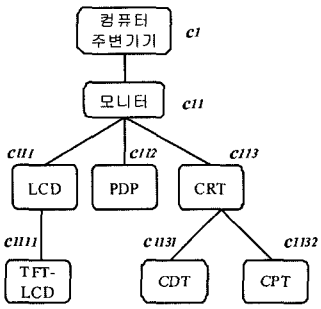

그림 4 범주 구조의 예

인스턴스 집합 $\{I(c_{111})\}=\{c_{1111}\}$과 $\{I(c_{113})\}=\{c_{1131}, c_{1132}\}$는 범주로 사용되는 대신 세부 범주 용어 사전으로 사용되어 문서 d의 특징 벡터와 일치할 경우 각각 w_{111}과 w_{113}를 증가시킬 수 있다. 세부 범주 용어 사전을 사용하는 이차 분류는 3.2.3절에서 다루기로 한다.

객체 기반 시소러스에서 개념은 수직적으로 일반화 관계를 가진다. 따라서 최상위 개념에서 파생되는 하위 개념들은 일반화 관계를 통해 일반화 계층을 형성한다. 이러한 시소러스의 계층적인 구조에 따른 범주 구조에 의해 후보범주 $c \in C(d)$는 문서 d가 $c_{i_1 i_2 \cdots i_l}$에 속하면, 시소러스의 일반화계층에 따라 상위 개념인 $c_{i_1 i_2 \cdots i_{l-1}}$에도 속하게 된다. 이를 다음과 같이 정의할 수 있다.

정의 5. 문서 d에 대해 $c_{i_1 i_2 \cdots i_l}(d)=1$이면 $c_{i_1 i_2 \cdots i_{l-1}}(d)=1$, $1 \le l \le n$ 이다.

예를 들어, LCD모니터의 상위 개념, 최상위 개념으로 모니터와 컴퓨터 부품이 존재한다면, 문서 d가 LCD모니터에 속할 때, 이 문서는 모니터와 컴퓨터 부품이라는 범주에도 속하게 된다.

위의 특징을 일반화하면 [명제 1]과 같다.

명제 1. 문서 d에 대해 $c_{i_1 i_2 \cdots i_l}(d)=1$이면, $c_{i_1 i_2 \cdots i_s}(d)=1$, $s=1,2,\cdots,l$, $1 \le l \le n$ 이다.

증명. [I] [정의 5]에 의해 $c_{i_1 i_2}(d)=1$이면 s=1,2에 대해 $c_{i_1}(d)=1$이고 $c_{i_1 i_2}(d)=1$이다.

[II] $c_{i_1 i_2 \cdots i_{l-1}}(d)=1$이면 $c_{i_1 i_2 \cdots i_s}(d)=1$, $s'=1,2,\cdots,l-1$라고 가정하자.

[정의 5]에 의해 $c_{i_1 i_2 \cdots i_l}(d)=1$이면, $c_{i_1 i_2 \cdots i_{l-1}}(d)=1$이므로 수학적 귀납법에 의해 [명제 1]이 증명된다.

[명제 1]에 따라 상위 범주에 속할 정도의 특징을 일반화 하면 [명제 2]와 같다.

명제 2. $d \in D$와 $c_{i_1 i_2 \cdots i_l} \in C$, $1 \le l \le n$ 에 대해 $c_{i_1 i_2 \cdots i_l}(d, w_{i_1 i_2 \cdots i_l})$, $w_{i_1 i_2 \cdots i_l} > 0$ 이면, $w_{i_1 i_2 \cdots i_s} \ge w_{i_1 i_2 \cdots i_{s'}}$ $s=1,2,\cdots,l-1$ 이다.

증명. $d \in D$가 $c_{i_1 i_2 \cdots i_l} \in C$, $1 \le l \le n$에 속할 정도 $w_{i_1 i_2 \cdots i_l} = \sum_{j=1}^{k} c_{i_1 i_2 \cdots i_l}(d_j^{k-nn})$에 대해 [명제 1]을 적용하면 $c_{i_1 i_2 \cdots i_l}(d_j^{k-nn})=1$ 일 때 $c_{i_1 i_2 \cdots i_s}(d_j^{k-nn})=1$ 이므로 $w_{i_1 i_2 \cdots i_s} \ge w_{i_1 i_2 \cdots i_{s'}}$ $s=1,2,\cdots,l-1$, $1 \le l \le n$ 이다.

후보범주 집합의 특성을 일반화하면 [명제 3]과 같다.

명제 3. $\alpha > 0$ 에 대해 $c_{i_1 i_2 \cdots i_l}(d, w) \in C_\alpha(d)$, $w \ge \alpha$, $1 \le l \le n$ 이면, $c_{i_1 i_2 \cdots i_s} \in C_\alpha(d)$, $1 \le s \le l$ 이다.

증명. [정의 3]에 의해 $c_{i_1 i_2 \cdots i_l} \in C_\alpha(d)$는 $c_{i_1 i_2 \cdots i_l}(d, w_{i_1 i_2 \cdots i_l})$, $w_{i_1 i_2 \cdots i_l} \ge \alpha$로 나타낼 수 있다. [명제 2]에 의해, $c_{i_1 i_2 \cdots i_s}(d, w')$에 대해 $w' \ge w_{i_1 i_2 \cdots i_l}$이므로 $c_{i_1 i_2 \cdots i_s} \in C_\alpha(d)$ $s=1,2,\cdots,l$가 된다.

만약, $c_{i_1 i_2 \cdots i_s} \in C(d)$일 때, $\forall s' > s$, $1 \le s, s' \le n$ 에 대해 $c_{i_1 i_2 \cdots i_s} \notin C(d)$이면 $c_{i_1 i_2 \cdots i_s}$를 c_{i_1}의 가장 구체적인 s단계의 범주 또는 단순히 가장 구체적인 범주라고 한다.

3.2 시소러스를 이용한 k-NN 분류 알고리즘 범주 할당의 모호성 제거

이번 절에서는 k-NN 분류 알고리즘을 이용한 문서의 분류에서 발생하는 모호성을 객체 기반 시소러스를 이용하여 해결함을 보인다.

정의 6. 범주 $c_{i_1 i_2 \cdots i_l} \in C$에 대한 상위 범주 $Sup(c_{i_1 i_2 \cdots i_l})$를 다음과 같이 정의한다.

$Sup(c_{i_1 i_2 \cdots i_l}) = \{ c_{i_1 i_2 \cdots i_s} \mid 1 \le s \le l-1 \}$,

$Sup(c_{i_1}^{top}) = c_{i_1}^{top}$ 이다.

정의 7. 후보범주 집합 $C(d)$에 대해 축약된 후보범주 집합을 $C_R(d)$라 하고 다음과 같이 정의한다.

$C_R(d) = C(d) - \bigcup_{l=1}^{n} Sup(c_{i_1 i_2 \cdots i_l})$, 여기서 $c_{i_1 i_2 \cdots i_l} \in C(d)$

$C_R(d)$로 부터 문서 d의 최종 범주를 선택하는 과정에서 고려되는 경우는 $|C_R(d)|$를 축약된 후보 범주 집합의 수라 할 때 $C_R(d)$의 원소 $c_{i_1 i_2 \cdots i_l}$에 대해

- $|C_R(d)| = 1$
- $|C_R(d)| \ge 2$ 이면서 $c_{i_1}^{top}$이 동일한 경우
- $|C_R(d)| \ge 2$ 이면서 $c_{i_1}^{top}$이 상이한 경우

로 나눌 수 있다.

$|C_R(d)| = 1$인 경우는 최하위 레벨에 해당하는 개념이 하나인 경우로 문서가 명확하게 최하위 레벨의 개념 범주에 속하고, $|C_R(d)| \ge 2$인 경우는 최하위 레벨에 해당하는 개념이 두 개 이상인 경우로 후보범주에 대해 시소러스의 집성화, 연관화 관계나 일반화 관계를 이용하여 해당 문서에 대해 명확한 범주를 정한다.

3.2.1 축약된 후보범주가 하나인 경우

$|C_R(d)| = 1$인 경우에는 문서 d가 속할 범주를 명확하게 결정할 수 있으므로 문서 d는 $C_R(d)$집합 내 범주에 속한다. 예를 들어, LCD 모니터와 관련된 내용을 담고 있는 아래 문서 d를 살펴보자.

> 2003년은 **FPD(Flat Panel Display)**시장에서 우리나라의 힘과 위상을 보여준 해였다. 전자 업계에서 여러 부문이 1위를 차지했고, 그중에서도 특히 **TFT LCD** 부문의 성장은 다른 어떤 나라도 넘보지 못할 정도였다. 올 한해도 **PDP/유기EL/CRT**등 모든 디스플레이 분야에서 … 이번에 살펴볼 제품은 **LCD 모니터** 중에서도 최근 사용자들이 가장 관심에 두고 있는 제품이다. 모델명은 **싱크마스터 매직 CX710P**로 고양이와 탁구공이 등장하는 TV 광고속의 그 제품이다.요즘의 CX710P에 대한 질문을 요약해 보면, '**싱크마스터 177X**와의 비교'/'응답속도 문제'/'정말 OSD 버튼이 없는가?'/'177X와 CX710P의 패널차이'. … 그리고 '고양이와 탁구공 나오는 **LCD모니터** 어디서 팔아요?(먼가요?)'등 일반 사용자와 초보 사양자의 다양한 질문들이 오가고 있다. 간단히 이 질문들에 대답을 해 본다면, 177X는 **TN방식**의 패널이고 CX710P는 PVA방식의 패널을 채용했다. …

위의 문서는 시소러스가 그림 5와 같이 주어졌을 때, k=9, α=4인 k-NN 분류 알고리즘에 의한 분류의 결과 후보범주로 c_1인 컴퓨터 주변기기와 c_{11}인 모니터, c_{111}인 LCD를 얻었다. 즉, $C(d)=\{c_1, c_{11}, c_{111}\}$이고, $C_R(d)=\{c_{111}\}$이다. 따라서 문서 d는 c_{111}인 LCD 범주에 할당한다.

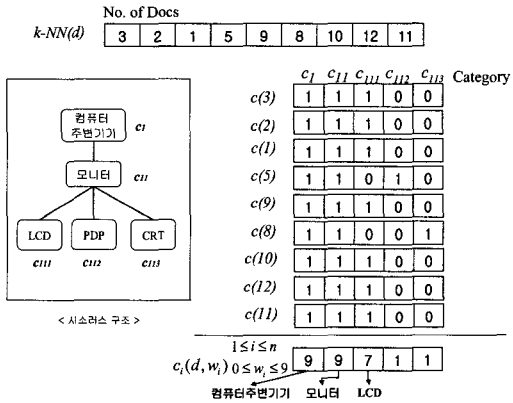

그림 5 $|C_R(d)| = 1$인 경우의 k-NN 분류 알고리즘의 분류 예

3.2.2 축약된 후보범주가 둘 이상이면서 최상위 개념이 동일한 경우

$|C_R(d)| \ge 2$인 경우, $\forall c, c' \in C_R(d)$의 최상위 개념이 동일한 경우에는 $C_R(d)$ 내 범주에 대해 객체 기반 시소러스에서 제공되는 범주 간 관련성을 이용하여 문서가 속할 명확한 범주를 정하도록 시도한다. 이를 위해 k-NN 분류 알고리즘에서 선택된 범주이기는 하나 후보범주 집합에 속하지 않는 범주 집합인 이차 후보범주 집합에 대해 다음과 같이 정의한다.

정의 8. 문서 $d \in D$에 대해 $C'(d) = C_{\alpha'}(d) - C_{\alpha}(d)$, $\alpha > \alpha' > 0$를 이차 후보범주로 정의한다. 따라서 $c_{i_1 i_2 \cdots i_l} \in C'(d)$에 대한 축약된 이차 후보범주는 $C'_R(d) = C'(d) - \bigcup_{l=1}^{n} Sup(c_{i_1 i_2 \cdots i_l})$ 이다.

$C'_R(d)$를 이용하여 $c, c' \in C_R(d)$ 중 d의 최종 범주를 선택하는 개략적인 과정은 다음과 같다.

1) 시소러스에서 각 $c'' \in C'_R(d)$에 대해 $c, c' \in C_R(d)$와 집성화나 연관화 관계를 가지고 있는지 확인한다.
2) c, c' 중 어느 범주가 c''와 집성화나 연관화 관계를 보다 많이 가지고 있는지 정량화한다.
3) 범주 간 관련성을 고려한 속할 정도 계산 방법에 의해 범주 c가 선택되었다면, 문서 d를 c 범주에 할당한다.
4) 3)의 과정에서 범주가 설정되지 못한 경우, c, c' 범주와 c, c'의 공통 상위 범주를 기준으로 상위범주들

에 문서를 할당한다.

예를 들어, 다음 CRT 모니터에서 LCD 모니터로의 전환에 대한 아래 문서 d를 살펴보자.

> CRT(음극선관, 미국의 브라운이라는 사람이 개발해서 브라운관으로도 불린다) TV가 성능면에서는 최고봉에 있다. 그럼에도 불구하고 LCD 방식의 TV가 계속 등장하고 있는 것은 CRT가 갖는 외형상의 한계 때문이다. 가장 큰 CRT TV는 38인치로 더 큰 화면은 만들 수 없다. …
> "LCD는 노트북에서 일반 데스크톱 PC, TV 등으로 수요시장이 확장성을 갖고 있다"면서 경기가 회복되면 추가로 시장이 늘어날 것"이라고 전망했다.
> …
> 싱크마스터 175W는 와이드 화면 비율을 지원, DVD, 게임 등에 최적화된 것이 특징. 또한 기존 LCD모니터가 밝기면에서 250칸델라 정도에 그쳤던 것과 달리 450칸델라를 지원할 뿐 아니라, CRT에서만 채용됐던 독자 기술 매직브라이트 기능을 채용해 밝고 생생한 화면을 보여준다.

위의 문서에 대해 k=10, α=4인 k-NN 분류 알고리즘의 분류 결과가 그림 6과 같다고 하자.

후보범주로 c_1인 컴퓨터 주변기기, c_{11}인 모니터, c_{111}인 LCD, c_{113}인 CRT가 나왔다면, $C(d)=\{c_1, c_{11}, c_{111}, c_{113}\}$이고, $C_R(d)=\{c_{111}, c_{113}\}$이다. 따라서 k-NN 분류 알고리즘의 분류 결과로는 문서 d가 $C_R(d)$집합 내 c_{111}인 LCD와 c_{113}인 CRT 중 어느 범주에 속하는지를 명확히 정할 수 없다. 그러므로 $C_R(d)$에는 속하지 않지만 $C'_R(d)$에는 속하는 Digital TV와의 집성화, 연관화 관계를 이용하여 LCD와 CRT 범주 중 명확한 범주를 설정하도록 한다.

정의 9. 범주 c_i에 대한 집성화, 연관화 관계에 대해 다음과 같이 정의한다.

$$r(c_i) = comp(c_i) \cup assoc(c_i)$$

여기서 $comp(c_i)$는 c_i와 집성화 관계에 있는 범주 집합이고, $assoc(c_i)$는 c_i와 연관화 관계에 있는 범주 집합이다.

정의 10. 범주 c_i와 $c_j \in r(c_i)$간의 집성화와 연관화 관련정도를 각각 $w_{c_ic_j}/comp$와 $w_{c_ic_j}/assoc$로 나타낸다.

정의 11. $c_i \in C_R(d)$이고 $c_j \in C'_R(d)$일 때, $r(c_i)$를 고려한 문서 d가 c_i에 속할 정도 $w^r_{c_i}(d)$는 다음과 같이 계산된다.

$$w^r_{c_i}(d) = w_{c_i} + \sum_{c_j \in comp(c_i)} w_{c_ic_j}/comp \times w_{c_j} + \sum_{c_j \in assoc(c_i)} w_{c_ic_j}/assoc \times w_{c_j}$$

$w^r_{c_i}(d)$에서 문서 d를 명기할 필요가 없을 경우에는 단순히 $w^r_{c_i}$로 표기한다.

예를 들어, 그림 6에서 $LCD, CRT \in C_R(d)$에 대해 w^r_{LCD}와 w^r_{CRT}를 구하면 다음과 같다.

$Digital\ TV \in C'_R(d)$에 대해 $w^r_{LCD} = 4 + w_{LCD,\,Digital\ TV}/comp \times w_{Digital\ TV} = 4 + 0.8 \times 2 = 5.6$, $w^r_{CRT} = 4$이다. 따라서 $w^r_{LCD}(d) > w^r_{CRT}(d)$이므로 문서 d는 LCD 범주에 속한다.

만약, 위 과정을 통해 명확한 범주를 설정하지 못한 경우 문서 d는 LCD와 CRT 범주와 LCD와 CRT의 상위 범주를 기준으로 상위 범주들을 설정하게 된다. 상

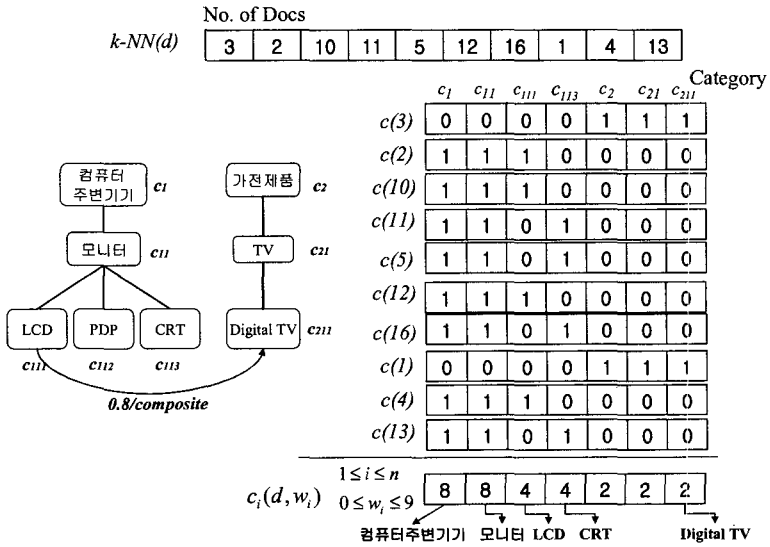

그림 6 $|C_R(d)| \geq 2$이면서 최상위 개념이 동일한 경우

위 범주를 구하는 방법은 [명제 4]와 같다.

명제 4. $|C_R(d)| \geq 2$라 하자. $\forall c_{i_1 i_2 \cdots i_{l-1} i_l} \in C_R(d)$에 대해 상위 범주는 $c_{i_1 i_2 \cdots i_{l-1}}$이고, $Sup_{direct}(\{c_{i_1 i_2 \cdots i_{l-1} i_l} \in C_R(d)\}) = c_{i_1 i_2 \cdots i_{l-1}}$로 나타낸다.

예를 들어, 그림 6에서 $C_R(d)=\{c_{111}, c_{113}\}$이고 $w^r_{LCD} = w^r_{CRT}$라면, LCD(c_{111})와 CRT(c_{113}) 범주와 이들을 일반화하는 모니터(c_{11})를 기준으로 한 상위범주들이 문서 d가 설정될 범주로 제안된다. [정의 5]에 의해 최하위 범주로 할당이 된 문서가 당연히 상위의 범주들에도 할당이 되는 경우와 후보범주들에 대해 일반화하는 범주를 구해 그 범주의 상위범주들에 문서를 할당하는 경우가 같지 않음을 유의하자. 예를 들어, 후보범주 c_{111}이 공통된 상위 범주 모니터 외에 다른 상위 범주 액정화면을 가지고 있는 경우라면 모니터에만 할당을 하고 CRT와는 관계없는 액정화면에는 할당하지 않게 되므로 이 방식은 보다 정확률을 향상시킬 수 있다.

$\forall c,c' \in C_R(d)$의 최상위 개념이 동일할 때, 문서를 분류하는 방법은 다음 알고리즘과 같다.

[알고리즘 1]

Resolve ($C_R(d), C'_R(d), d, Th$)

Begin

1. $c,c' \in C_R(d)$, $|C_R(d)| \geq 2$,
   시소러스 Th로부터 $c'' \in (\gamma(c) \cup \gamma(c')) \wedge c'' \in C'_R(d)$의 w^r_c과 $w^r_{c'}$를 계산한다.
2. 만약, $\forall c' \in C_R(d) \wedge c \neq c'$에 대해 $w^r_c > w^r_{c'}$이고 이를 만족하는 c가 하나이면 Return(c).
3. 그렇지 않으면, $C_R(d) \leftarrow \{ c,c' \mid w^r_c = w^r_{c'} \}$.
4. $Sup_{direct}(C_R(d)) \neq \emptyset$이면 d를 $c \in C_R(d)$에 각각 할당하고 Return($Sup_{direct}(C_R(d))$).
   그렇지 않으면, d를 $c \in C_R(d)$에 각각 할당한다.

End

다음 절에서는 $c,c' \in C_R(d)$의 최상위 개념이 상이한 경우를 고려하여 이 알고리즘의 4번째 단계를 재정의하기로 한다.

3.2.3 축약된 후보범주가 둘 이상이면서 최상위 개념이 상이한 경우

$|C_R(d)| \geq 2$이면서 $\forall c,c' \in C_R(d)$의 최상위 개념이 상이한 경우는 문서 중 단어들의 가중치를 구해 이를 반영함으로써 해결한다. 예를 들어, LCD 모니터의 종류 및 용도와 관련된 내용을 담고 있는 아래 문서 d를 살펴보자.

> 현재 시중에는 다양한 종류의 **액정모니터**가 출시되어 있는데, …
> **STN(Super Twisted Nematic) LCD** : **핸드폰** 액정화면으로 주로 사용되고 있으며, **노트북**, **개인휴대단말기(PDA)** 이나 **데스크탑용 액정모니터**에도 이전에 많이 쓰였습니다. 가격이 싼 편이지만, 정면에서 확인할 때에만 잘 보이는 단점이 있으며 측면에서는 화면이 어둡게 보입니다. … DSTN은 Double Super Twisted Nematic로 STN방식을 2층으로 보강한 방식입니다. …
> **TFT(thin film transistor) LCD** : 최근 노트북은 대부분 **TFT 모니터**를 사용하기 때문에 측면에서도 자연색을 확인할 수 있습니다. TFT방식은 빠른 스크롤에도 화면 떨림이 적고 색감이나 전력 소비면에서도 DSTN보다 우수합니다. **TFT-LCD**는 소형 **TV** 에서부터 대형 **벽걸이형 TV**, 차량용 **네비게이션**, 노트북, **모니터** 그리고 **디지탈 카메라**나 **캠코더** 등 다양한 제품에 접목되고 있습니다.

위의 문서에 대해 k=8, α=3인 k-NN 분류 알고리즘에 의한 분류 결과는 다음과 같다고 하자.

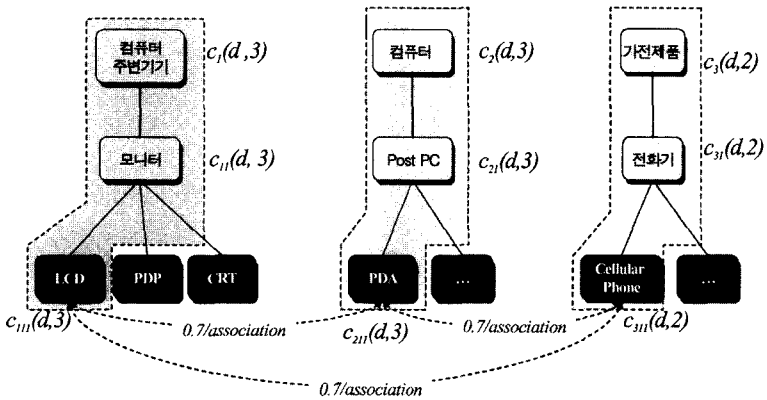

그림 7 $|C_R(d)| \geq 2$이고 $LCD, PDA \in C_R(d)$의 최상위 개념이 각각 상이한 경우

그림 7에 의해 $C_3(d)=\{c_1, c_{11}, c_{111}, c_2, c_{21}, c_{211}\}$이고 $C_R(d)=\{c_{111}, c_{211}\}$, $C_R'(d)=\{c_{311}\}$임을 알 수 있다. 여기서 $w^r_{c_{111}}$과 $w^r_{c_{211}}$로부터 d의 범주를 결정할 수 있다면 축약된 후보범주가 둘 이상이면서 최상위 개념이 동일한 경우와 같이 [알고리즘 1]을 적용할 수 있을 것이다. 그러나 $w^r_{c_{111}} = w^r_{c_{211}} = 3 + 0.7 \times 2 = 4.4$이고, $Sup_{direct}(C_R(d))$도 구할 수 없으므로 문서 d가 속할 범주를 결정할 수 없다. 이 경우 시소러스는 개념 간의 일반화와 집성화, 연관화 관계 그리고 인스턴스 관계로부터 범주들의 특성을 잘 기술하고 있기 때문에 이 정보로부터 생성되는 각 범주별 관련 단어들의 집합을 범주 결정에 이용할 수 있다. 즉, 각 후보범주의 관련 단어 집합 중 문서에 표현된 단어가 많이 포함되어 있는 범주일수록 문서가 속할 가능성이 높으므로 이 정보를 이용하여 d가 속할 범주를 예측할 수 있다. 이 방법은 [알고리즘 1]의 경우에서도 $Sup_{direct}(C_R(d))$를 구하기 이전 좀 더 구체적인 범주를 d의 최종 범주로 설정할 수 있으므로 정확률을 향상시키게 된다. 이 방법을 기술하기에 앞서 다음과 같은 정의가 필요하다.

정의 12. 범주 $c \in C_R(d)$에 대한 세부 범주 용어 사전 $ld(c)$를 다음과 같이 정의한다.

$ld(c)=\{c\} \cup ass(c) \cup comp(c) \cup sym(c) \cup I(c) \cup sym(I(c))$.

여기서 $sym(c)$는 c의 동의어집합을 말한다.

범주 $c \in C_R(d)$의 세부 범주 용어 사전은 범주 c의 새로운 특징 벡터로 볼 수 있다. 따라서 범주 c의 세부 범주 용어 사전과 $word(d)$의 공통되는 단어의 가중치 정보는 범주 c와 문서 d가 얼마나 관련 있는지를 나타낸다.

$ld(c)$와 문서 d의 $word(d)$간의 공통 단어는 다음과 같이 정의한다.

정의 13. 문서 d의 $word(d)=t_1/a_1, t_2/a_2, \cdots, t_{N_w}/a_{N_w}$에 대해 $ld(c) \sqcap word(d)=\{ t_i/a_i \mid t_i \in ld(c) \cap W\}$, 여기서 W는 D의 단어집합이다.

일단 $ld(c) \sqcap word(d)$가 구해지면, $c \in C_R(d)$의 범주 선택의 모호성은 세부 범주 용어 사전에 의해 해결될 수 있다. $t_i/a_i \in ld(c) \sqcap word(d)$의 a_i의 값은 코사인 정규화에 의해 [0,1]의 가중치를 가지고 있고, 전체 범주 C의 각 범주 c에 대해 w_c는 [0,k]의 값을 가지고 있으므로 범주결정에 있어 w_c와 a_i의 영향력을 동등하게 하기 위해 w_c을 k값으로 나누어 속할 정도를 계산한다.

정의 14. 문서 d에 대해 $t_i/a_i \in ld(c) \sqcap word(d)$의 a_i를 고려한 속할 정도 w^{ld}_c는 다음과 같이 계산된다.

$$w^{ld}_c = w_c / k + \sum a_i$$

세부 범주 용어 사전을 고려하여 [알고리즘 1]을 재정의하면 다음과 같다.

[알고리즘 2]

Resolve ($C_R(d), C_R'(d)$, d, Th, β)

Begin

1. $c, c' \in C_R(d)$, $\mid C_R(d) \mid \geq 2$, 미리 정의된 임계치 $\beta > 0$ 에 대해 시소러스 Th로부터 $c'' \in (r(c) \cup r(c')) \wedge c'' \in C_R'(d)$의 w^r_c과 $w^r_{c'}$를 계산한다.
2. 만약, $\forall c' \in C_R(d) \wedge c \neq c'$에 대해 $w^r_c > w^r_{c'}$ 이고 이를 만족하는 c가 하나이면 Return(c).
3. 그렇지 않으면, $C_R(d) \leftarrow \{c, c' \mid w^r_c = w^r_{c'}\}$.
4. $c \in C_R(d)$와 $\forall t_i/a_i \in (ld(c) \sqcap word(d))$에 대해 $w^{ld}_c = w_c/k + \sum a_i$를 계산한다.
5. $\forall c' \in C_R(d)$에 대해 $w^{ld}_c - w^{ld}_{c'} \geq \beta$ 이면 Return(c).
6. $\mid C_R(d) \mid \geq 2$이고 $Sup_{direct}(C_R(d)) \neq \emptyset$이면 d를 $c \in C_R(d)$에 각각 할당하고 Return($Sup_{direct}(C_R(d))$).

그렇지 않으면, d를 $c \in C_R(d)$에 각각 할당한다.

End

예를 들어, k=8일 때, 그림 8과 같이 세부 범주 용어 사전을 가정하자.

이 용어 사전을 기반으로 $ld(LCD)$와 $ld(PDA)$를 구하면 다음과 같다.

$ld(LCD)$ = {"LCD," "TFT-LCD Monitor," "STN LCD," "Thin Film Transistor Liquid Crystal Display," "Liquid Crystal Display," " Cellular Phone," "PDA"}.

$ld(PDA)$ = {"Cellvic," "Palm," "PDA," "Personal digital assistant," "LCD," "Cellular Phone" }.

예를 들어, $word(d)$가 다음과 같다고 가정하자. $word(d)$ ={"cellular phone"/0.19, "display"/0.33, "lcd"/0.64, "pda"/0.50, "stn lcd"/0.25, ···, "tft-lcd"/0.53, ··· }.

$c, c' \in C_R(d)$에 대해 $ld(c) \sqcap word(d)$를 구하면,

$ld(LCD) \sqcap word(d)$={"cellular phone"/0.19, "lcd"/0.64, "pda"/0.50, "stn lcd"/0.25, "tft-lcd"/0.53}이고,

$ld(PDA) \sqcap word(d)$={"cellular phone"/0.19, "lcd"/0.64, "pda"/0.50 }이다. 따라서 w^{ld}_{LCD}와 w^{ld}_{PDA}는 다음과 같다.

$t_i/a_i \in ld(LCD) \sqcap word(d)$에 대해 $w^{ld}_{LCD} = w_{LCD}/8 + \sum a_i = 3/8 + 2.11 = 2.485$.

$t_i/a_i \in$ ld(PDA) $\sqcap word(d)$에 대해 $w^{ld}_{PDA} = w_{PDA}/$

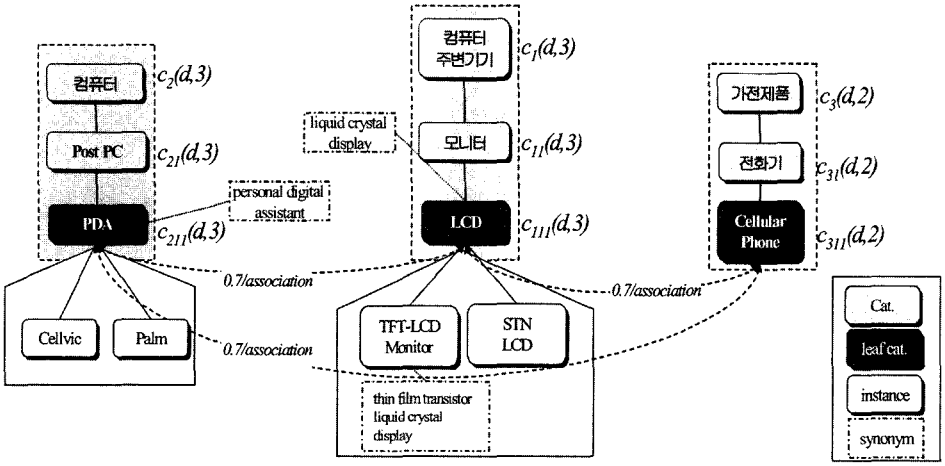

그림 8 시소러스에 의해 구해진 세부 범주 용어 사전

$8+\sum a_i=3/8 + 1.33 = 1.705$.

β=0.5에 대해 $w^{Id}_{LCD} > w^{Id}_{PDA}$이므로 LCD와 PDA 범주 중 LCD 범주에 문서를 할당한다.

이 알고리즘에서 문서 d의 $word(d)$ 대신 각 범주에 속하는 $d^{k-nn} \in k-NN(d)$의 $word(d^{k-nn})$들로부터 평균 tfidf 가중치를 구해 적용하는 방식을 고려할 수도 있을 것이다. 그러나 실험의 결과 이 방식은 k-NN 분류 알고리즘에서 이웃 문서들을 결정하는 과정에서의 오류와 범주와 무관한 각 특징벡터들이 누적됨으로 인하여 실제 실험 문서를 분류하는데 있어서 오히려 분류기의 성능을 저하시키는 결과를 보였다. 따라서 본 논문에서는 문서 d의 $word(d)$만을 고려하여 [알고리즘 2]의 4단계를 실행한다.

3.2.4 두개 이상의 축약된 후보범주에 대한 종합적 의미의 최종 범주 추론

[알고리즘 2]는 조금 수정한다면, 축약된 후보범주가 둘 이상일 때, 이들 범주를 종합적으로 의미하는 범주를 최종 범주로 추론해 낼 수 있는 잠재력을 가지고 있음을 주의해 보자. 예를 들어, 비디오, 오디오, 홈시어터에 관련된 다음의 문서를 살펴보자.

> 영화관이나 콘서트장에서 느낀 감동을 집에서도 느껴보고 싶은 마음은 누구에게나 있을 것이다. … 가전매장에 **홈시어터** 시스템을 갖추고 열띤 홍보전을 벌이면서 일반 소비자들로 …
> 방법이 가까이에 있다. **홈시어터**는 **앰프스피커케이블**대형 스크린 모두가 중요한 구성요소다. 일반적으로 **홈시어터**에는 **AV리시버**가 앰프로 사용된다. **파워앰프**와 **전방향 앰프**, **디지털리시버앰프** 그리고 튜너(tuner)를 합쳐 놓은 형태인 **AV리시버**는 음원을 선택하고 조작하는 중추적 기능을 담당한다. 이 **AV리시버**는40만원 정도의 가격이면 적당한 제품을 고를 수 있다. 디지털방식으로 영상 및 음향을 재생하기 때문에 … 얻을 수 있는 **DVD플레이어**가 **홈시어터** 구성에서 날로 중요해지고 있다. … 는 경험을 위해 **스피커**만큼 중요한 요소는 없다. **홈시어터**용 **스피커**는 음악이나 영화를 감상할 때 1페어의 **스피커**로는 부족한 저음역 출력을 목적으로 **서브우퍼**를 포함하고 있다. 이스턴전자 **스피커** 등 저가이면서 소형이지만 양질의 출력을 얻을 수 있는 **스피커**가 여러 종류 출시돼 있다. **TV**는 넓은 … **디지털TV**가 적당하다. 음향기기에 비해 가격이 높아 **홈시어터** 구성 … 조금 더 돈을 투자할 수 … 34인치 대형 **평면TV**나 **프로젝션**를 선택할 경우 더욱 극장 같은 느낌을 얻을 수 있다.

위의 문서에 대해 그림 9의 시소러스 구조에 따른 k=11, α=4인 k-NN 분류 알고리즘의 분류결과는 다음과 같다.

문서 d에 대해 $C_R(d)=\{c_1, c_2\}$와 $C'_R(d)=\{c_3\}$가 구해지고, [알고리즘 2]에 따라 w^r_{VIDEO}와 w^r_{AUDIO}를 계산하여 비디오(c_1)와 오디오(c_2) 중 홈시어터(c_3)와 관련성이 더 높은 범주에 문서 d를 할당하려고 해보자.

$HomeTheater \in C'_R(d)$에 대해 $w^r_{VIDEO} = 4 + w_{VIDEO,\,HomeTheater}/\text{comp} \times w_{HomeTheater} = 4 + 2.4 = 6.4$

$HomeTheater \in C'_R(d)$에 대해 $w^r_{AUDIO} = 4 + w_{AUDIO,\,HomeTheater}/\text{comp} \times w_{HomeTheater} = 4 + 2.4 = 6.4$

그러나 $w^r_{VIDEO} = w^r_{AUDIO}$으로 명확한 범주를 설정할 수가 없으므로 이 경우 [알고리즘 2]에서는 $C_R(d)=\{c_1, c_2\}$ 내 비디오와 오디오 범주의 세부 범주 용어 사전을 고려하여 문서 d를 할당하게 된다. 그런데 여기서 $w^r_{HomeTheater}$를 구해 본다면, $w^r_{HomeTheater} = 3 + w_{HomeTheater,\,VIDEO}/comp \times w_{VIDEO} + w_{HomeTheater,\,AUDIO}/comp \times w_{AUDIO} = 3 + 0.8 \times 4 + 0.8 \times 4 = 9.4$ 이므로 $w^r_{HomeTheater}$가 w^r_{VIDEO}

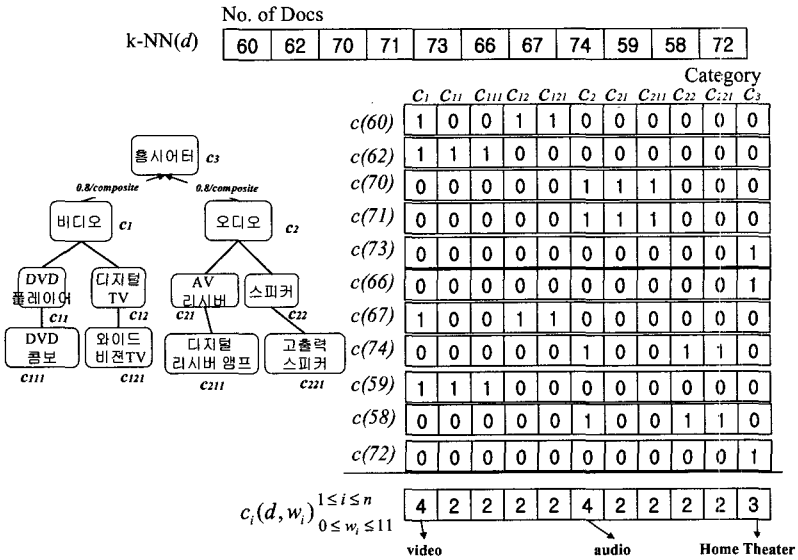

그림 9 $|C_R(d)| \ge 2$이면서 최상위 개념이 상이한 경우

나 w^r_{AUDIO}보다 현저하게 큰 것을 알 수 있다. 만약 $c_3 \in C_R(d)$였다면, 비디오와 오디오의 논리적인 결과로서 홈시어터에 문서 d를 할당함으로써 보다 지능적인 할당이 가능할 것이다. 그러나 이 방법은 $c \in C_R(d)$에 있는 모든 범주들을 무시하게 되는 문제가 발생하게 되므로 이 기법을 검증하여 기존 알고리즘을 확장하는 방법은 향후 연구 과제로 남겨놓기로 한다.

4. 실험 결과 및 고찰

본 실험에서는 전자제품에 대해 디렉토리 구조를 가지고 있는 웹 사이트들로부터 전자제품을 도메인으로 문서들을 수집하고, 이를 바탕으로 객체 기반 시소러스를 구축하였다. 실험에 쓰인 문서의 집합은 290개의 훈련 문서와 이 훈련 문서외에서 선택한 137개의 테스트 문서로 구성하였다. 객체 기반 시소러스내의 범주는 6개의 대분류 범주(Topic) - 가전제품, 컴퓨터, 컴퓨터 주변기기, 컴퓨터 부품, 음향기기, 사무기기에 대해 24개의 최하위 범주로 설정하였으며, 본 방식의 성능 향상의 효과를 보이기 위해 실험에 사용된 객체 기반 시소러스의 용어 수는 340개이다. 실험에 사용된 문서의 예는 다음과 같다.

```
<TOPIC>컴퓨터주변기기</TOPIC>
<SUBTOPIC>모니터</SUBTOPIC>
<SUBSUBTOPIC>LCD</SUBSUBTOPIC>
<TEXT>
고해상도 LCD의 대중화 선언
17인치 LCD 모니터 KDS RAD-7
특징 : 온화한 화면과 선명한 이미지를 보여준다. 디자인이 날렵해 공간을 적게 차지한다.
제원 : 화면크기 17인치, TFT 액티브 매트릭스, 최대 해상도 1028×1024/ 수직해상도 75Hz, 픽셀피치 0.264mm, 무게 6kg
LCD 모니터는 15인치가 주류다. 얇은 몸체에 선명한 화면을 보여주지만 엄청난 값 때문에 선뜻 살 수가 없었다. 그러나 지금은 많이 싸져 최고급 완전평면 모니터와 값이 비슷해졌다.LCD 모니터는 완전평면인데다 넓게 보여 15인치짜리가 17인치 일반 모니터 화면과 같다. 17인치 LCD와 19인치 모니터도 마찬가지의 관계다. LCD … 그래픽카드 연결단자 부분에 어댑터용 단자가 있다. 따라서 배선이 아주 깔끔하다. 금속 분위기의 플라스틱 몸체도 보기 좋다. 문제점은 색 온도가 지나치게 낮다는 것이다.
</TEXT>
```

문서 분류에 대한 성능 평가는 재현율(recall)과 정확률(precision), F1-Measure로 수행하였다. 재현율, 정확률, F1-Measure에 대한 식은 다음과 같다.

Category		전문가에 의한 분류	
		Correct	Incorrect
분류기에 의한 분류	Correct	a	b
	Incorrect	c	d

정확률 : $p = \frac{a}{a+b}$,

재현율 : $r = \frac{a}{a+c}$,

F1-Measure : $F_1(r,p) = \frac{2 \times p \times r}{p+r}$.

k-NN 분류 알고리즘에 의한 문서 분류는 유사도 계산 방법과 근접한 이웃의 개수인 k, 후보범주의 임계치의 변화에 따라 실험 결과가 다르게 된다[13]. 본 논문에서는 문서 벡터 간 내적을 이용하여 유사도를 계산하

표 1 k=17, 후보범주의 임계치 7에 대한 실험 결과

	실험방법	Precision	Recall	F-measure
계층 구조를 반영한 분류	k-NN	90.73	47.94	62.73
	변경된 k-NN	84.03	88.14	86.04
	변경된 k-NN+Thesaurus	89.58	93.04	91.27
최하위 범주에 대한 분류	k-NN	72.85	80.29	76.39
	변경된 k-NN	71.26	90.51	79.74
	변경된 k-NN+Thesaurus	86.71	90.51	88.57

였으며, 이웃하는 문서의 수인 k와 후보범주의 임계치를 다르게 하여 실험을 수행하였다. 실험은 크게 계층 구조를 반영한 분류와 계층 구조의 최하위 범주만을 기준으로 한 분류로 나눠 볼 수 있다. 세부적으로는 1)속할 정도가 가장 높은 범주 하나만을 선택하여 문서를 할당하는 "k-NN"과 2)k-NN에 시소러스의 계층 구조를 반영하여 문서가 최하위 범주에 할당되면 상위의 범주들과 후보범주 모두에 할당을 하는 "변경된 k-NN", 3)변경된 k-NN에 [알고리즘 2]를 적용한 "변경된 k-NN+Thesaurus"방법으로 실행하였다. 먼저, k=17이고 후보범주의 임계치 7에 대한 실험 결과를 살펴보면 다음과 같다.

표 1에서 보는 바와 같이 계층 구조를 반영한 분류에서 k-NN의 정확률은 90.7%정도로 높지만 그에 비해, 재현율은 상당히 많이 떨어진다. 이는 k-NN이 속할 정도가 가장 높은 하나의 범주에만 문서를 할당하는데, 계층적인 구조에서는 최상위에 있는 범주들이 대부분 속할 정도가 가장 높으므로 문서가 최상위 범주에만 할당되는 결과가 나오기 때문이다. 따라서 k-NN에 시소러스의 계층 구조의 특성을 반영한 변경된 k-NN이 k-NN에 비해 정확률은 다소 떨어지지만 재현율은 높음을 알 수가 있다. 변경된 k-NN에 비해 변경된 k-NN+Thesaurus의 정확률은 5.5% 정도 향상되었음을 보이는데, 이는 변경된 k-NN 에서도 계층 구조에 의해 최하위 범주에 속하면 상위 범주에도 자동 할당하게 되므로 실제적으로 최하위 범주가 잘못 할당되었다 하더라도 상위의 범주들이 제대로 할당이 된 경우는 정확률이 많이 떨어지지 않기 때문이다. 그러나 최하위 범주에 대해서만 분류를 실행해 보면, 변경된 k-NN+Thesaurus가 변경된 k-NN보다 13.86% 정도 정확률을 향상 시켰음을 알 수 있다. 변경된 k-NN+Thesaurus가 계층구조를 반영한 분류보다 최하위 범주에 대한 분류에서 더욱 향상된 결과를 보이므로 계층구조를 반영한 분류에선 문서가 상위 범주는 제대로 할당이 되고 최하위 범주만 잘못 할당된 경우가 많음을 알 수 있다.

표 2 k-NN을 이용한 문서 분류 결과

Method(k)	Precision	Recall	F-Measure
k-NN(14)	76.82%	84.67%	80.56%
k-NN(15)	75.51%	81.02%	78.17%
k-NN(16)	74.15%	79.56%	76.76%
k-NN(17)	72.85%	80.29%	76.39%
k-NN(18)	75.17%	81.75%	78.32%

최하위 범주에 대해 k를 달리하여 실행한 k-NN의 분류 결과는 표 2와 같다.

최하위 범주에 대해 k-NN 을 이용한 분류 후 객체 기반 시소러스를 이용한 문서 분류 결과는 표 3과 같다.

위의 결과에 대한 재현율과 정확률 그래프는 그림 10과 같다.

k-NN에 비해 객체 기반 시소러스를 이용한 분류의 재현율과 정확률의 향상된 결과치를 알아보면 그림 11과 같다.

그림 11은 k-NN에 비해 객체 기반 시소러스를 이용한 분류의 정확률이 최고 13.86% 까지 향상되었음을 보이고 있다. 여기서 특이한 사항은 k가 17인 경우 시소러스를 이용한 분류의 정확률이 현저히 향상되었다는 점이다. 그 이유는 각 범주 당 훈련 문서가 10~15개 정도이므로 k가 17일 때 k-NN에서 범주 할당 시 모호한

표 3 k-NN과 객체 기반 시소러스를 이용한 문서 분류 결과

Method(k, 후보범주 임계치)	Precision	Recall	F-Measure
변경된 k NN(14,6) + Thesaurus	87.77%	89.05%	88.41%
변경된 k NN(15,6) + Thesaurus	86.52%	89.05%	87.77%
변경된 k NN(16,7) + Thesaurus	84.89%	86.13%	85.51%
변경된 k NN(17,7) + Thesaurus	86.71%	90.51%	88.57%
변경된 k NN(18,7) + Thesaurus	87.05%	88.32%	87.68%

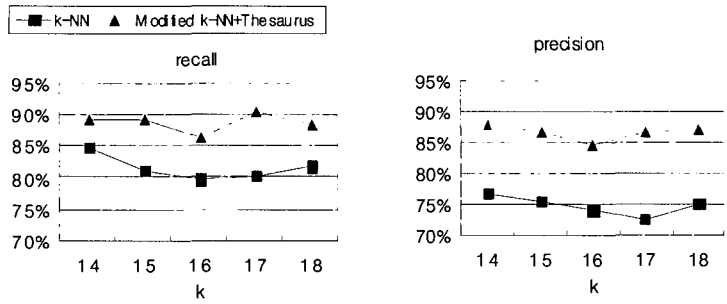

그림 10 최하위 범주에 대해 k-NN과 변경된 k-NN+Thesaurus의 비교

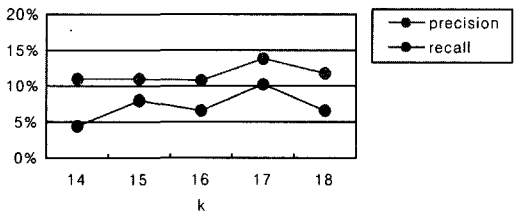

그림 11 최하위 범주에 대해 k-NN과 비교한 변경된 k-NN+Thesaurus 분류의 향상치

경우가 특히 많이 발생하였고, 객체 기반 시소러스를 이용한 분류가 이 모호성을 효과적으로 해결하였기 때문이다.

본 실험 방법에서 얻어지는 F-Measure는 사용된 시소러스 용어 개수에 따라 달라질 수 있다. 따라서 본 실험에서 사용한 340개 이상의 규모로 객체 기반 시소러스를 정교하게 구축한다면 더욱 향상된 성능을 기대할 수 있을 것이다.

5. 결론 및 향후 과제

본 논문에서는 k-NN 분류 알고리즘에 전문가의 지식을 바탕으로 한 시소러스를 접목시킴으로써 자동 문서 분류의 성능을 높이는 새로운 방법을 제안하였다. 즉, k-NN 분류 알고리즘을 이용한 문서 분류 시 특정 범주로 분류하기가 명확치 않을 경우, 객체 기반 시소러스로부터 얻어지는 범주 간의 관련성을 이용하여 모호성을 줄이는 방식을 제안하였다. 또한 실험을 통하여 본 방법이 340개의 시소러스 용어 사용만으로 기존 k-NN 분류 알고리즘에 비해 정확률을 최고 13.86%까지 향상시킴을 보였다.

분류상의 모호성은 k-NN 분류 알고리즘뿐만 아니라 다른 문서 분류 알고리즘에서도 공통적으로 발생하는 문제이므로 본 방식과 연동된다면 k-NN 분류 알고리즘 외의 다른 알고리즘들 또한 향상된 성능을 기대할 수 있을 것이다.

향후 과제로는 3.2.4절에 제시된 두개 이상의 축약된 후보범주에 대한 종합적 의미의 최종 범주를 추론하는 방법을 보완하여 알고리즘을 개선하고 나아가 본 논문에서 제안한 자동 문서 분류 방법을 시맨틱 웹 환경에 확장 적용하는 것이다. 즉, 시맨틱 웹 환경에서의 온톨로지는 시소러스에 비해 많은 표현관계를 가지고 있으므로 이를 지식베이스로 이용하여 웹 문서의 자동 문서 분류에 적용하면 기존 문서 분류의 모호성을 줄여 보다 정확한 문서 분류 결과를 얻을 수 있다.

참 고 문 헌

[1] Mehnet, R., "Federal Agency and Federal Library Reports : National Library of Medicine," Bowker Ann : Library and Book Trade Almance, second ed., pp. 110-115, 1997.

[2] Yang, Y., "An evaluation of statistical approaches to text categorization," Journal of Information Retrieval, Vol. 1, No. 1/2, pp. 67-88, 1999.

[3] Lam, W., Low, K. F. and Ho, C. Y., "Using a Bayesian network induction approach for text categorization," In Proceeding of the fifteenth International Joint Conference on Artificial Intelligence(IJCAI), Vol. 1, pp. 745-750, 1997.

[4] Diao, L., Hu, K., Lu, Y. and Shi, C., "Boosting simple decision trees with Bayesian learning for text categorization," In Proceeding of the fourth World Congress on Intelligent Control and Automation, Vol. 1, pp. 321 - 325, 2002.

[5] Soucy, P. and Mineau, G. W., "A Simple KNN Algorithm for Text Categorization," In Proceeding of the first IEEE International Conference on Data Mining(ICDM), Vol. 28, pp. 647-648, 2001.

[6] Sasaki, M. and Kita, K., "Rule-Based Text Categorization Using Hierarchical Categories," In Proceeding of the IEEE International Conference on Systems, Man and Cybernetics, Vol. 3, pp. 2827-2830, 1998.

[7] Jalam, R. and Teytaud, O., "Kernel-based text categorization," In Proceeding of the International Joint Conference on Neural Networks(IJCNN), Vol.

3, pp. 15-19, 2001.
[8] Schapire, R. E. and Singer, Y., "Text categorization with the concept of fuzzy set of informative keywords," In Proceeding of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Vol. 2, pp. 609-614, 1999.
[9] Duda, R. O. and Hart, P. E., "An algorithm for text categorization with SVM," TENCON '02. In Proceeding of the IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, Vol.1, pp. 47-50, 2002.
[10] Sebastiani, F., "Machine learning in automated text categorization," ACM Computing Surveys, Vol. 34, Issue. 1, pp. 1-47, 2002.
[11] Antonie, M. L. and Zaiane, O. R., "Text document categorization by term association," In Proceeding of the second IEEE International Conference on Data Mining(ICDM), pp. 19-26, Dec. 2002.
[12] Hiroshi, U., Takao, M. and SHIOYA, I., "Improving Text Categorization By Resolving Semantic Ambiguity," In Proceeding of the IEEE Pacific Rim Conference on Communications, Computers and Signal processing(PACRIM), pp. 796-799, 2003.
[13] Bao, Y. and Ishii, N., "Combining Multiple K-Nearest Neighbor Classifiers for Text Classification by Reducts," In Proceeding of the fifth International Conference on Discovery Science, pp. 340-347, 2002.
[14] Han, E. H., Karypis, G. and Kumar, V., "Text categorization using weight adjusted k-nearest neighbor classification," In Proceeding of the fifth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining(PAKDD), pp. 53-65, 1999.
[15] Lim, H. S., "A Comparative Evaluation of Korean Text Categorization based on kNN Learning," In Proceeding of the International Conference on Artificial Intelligence(IC-AI), pp. 755-759, 2002.
[16] 고영중, 서정연, "문서관리를 위한 자동문서범주화에 대한 이론 및 기법", 정보관리연구, 제33권, 제2호, pp. 19-32, 2002.
[17] Aas, K. and Eikvil, L., "Text Categorization : A Survey," Report No. NR 941, Norwegian Computing Center. URL http://citeseer.ist.psu.edu/aas99text.html
[18] 이경찬, 강승식, "자질 중요도 계산 기법에 의한 자동 문서 범주화", 한국정보과학회 봄 학술발표 논문집(B), 제30권, 제2호, pp. 537-539, 2003.
[19] Choi, J. H., Yang, J. D. and Lee, D. G., "An Object-Based Approach to Managing Domain Specific Thesauri: Semiautomatic Thesaurus Construction and Query-Based Browsing," International Journal of Software Engineering & Knowledge Engineering, Vol. 10, No. 4, pp. 1-27, 2002.

방 선 이
2000년 전북대학교 통계학과(학사). 2002년 전북대학교 전산통계학과(석사). 2002년~현재 전북대학교 컴퓨터통계정보학과 박사과정. 관심분야는 문서 정보검색, 인공지능, 온톨로지 등

양 재 동
1983년 서울대학교 컴퓨터공학과(학사) 1985년 한국과학기술원 전산학과(석사) 1991년 한국과학기술원 전산학과(박사) 1995년~1996년 Univer.of Florida, Visiting Scholar. 현재 전북대학교 전자정보공학부 교수. 관심분야는 멀티미디어 정보검색, 문서 정보 검색, 온톨로지, OODB, Expert System 등

양 형 정
1991년 전북대학교 전산통계학과(학사) 1993년 전북대학교 전산통계학과(석사) 1998년 전북대학교 전산통계학과(박사) 2000년 동신대학교 컴퓨터 응용학과군 전임강사. 2003년~현재 카네기멜론대학 컴퓨터과학과 포스트닥터 연구원. 관심분야는 문서 정보검색, 멀티미디어 정보검색, 온톨로지, OODB 등