工學碩士 學位論文

온톨로지를 적용한 문서 분류에서의 자질 추출

Feature Extraction with Ontology for Document Classification

指導敎授 辛 沃 根

2008年 2月

韓國海洋大學校 大學院

컴퓨터工學科

趙 希 英

목 차

표 목차

그림 목차

Feature Extraction with Ontology for Document Classification

Hee-Young Cho

Department of Computer Engineering,
Graduate School, Korea Maritime University.

Advised by Ok-Keun Shin

Abstract

With rapid development of Internet and information service techniques, a huge amount of electronic documents are steadily produced on the Web. The documents like news papers are classified by trained persons without any delay from day to day, but it is a very labor-intensive work and requires a lot of time and cost. Several studies on automatic document classification have been performed in order to lessen this burden. The studies using techniques of machine learning and natural language processing have shown successful results in the Web mining field. The performance of document classification systems is very much depending on feature sets even though there are also other many factors that can affect the performance. In this thesis, we propose methods for extracting good feature sets using ontology. Terms in documents are transformed into terms in ontology in order to reduce the size of feature sets and to compress information of the documents at the expense of some loss of the meaning. This transformation can be performed after or before general feature selection. We use only relations of synonyms and hypernyms in Korean ontology, U-WIN which has been developed by Ulsan University. We have experimented with the proposed methods on four classifiers and nine feature

selectors in order to objectively evaluate the performance of the proposed methods. The several experiments have shown that the proposed methods using ontology outperform existing feature selectors over most classifiers except a naïve Bayesian classifier and also the method applying ontology after feature selection outperforms that before feature selection over every classifiers. We have observed that the performance of feature selectors is very sensitive to classifiers, especially Rocchio classifier. In the future, we will experiment with a large scale of documents of various fields and many languages like English and Japanese to show more objective results. The ambiguation on multiple hypernyms of a term will be tackled as word sense disambiguation problem.

제 1 장 서론

인터넷의 발전과 함께 다양한 미디어와 인터넷을 통해서 자신의 의견이나 정보를 전달하고 있으며 사용자들은 원하는 정보를 찾기 위해 많은 시간과 비용을 소모하고 있다. 이와 같은 환경에서 정보검색 서비스 제공자인 기업은 사용자들이 원하는 정보를 보다 편리한 방법으로 찾을 수 있도록 다양한 형태의 서비스를 제공할 필요성이 날로 증가하고 있다. 이에 따라, 서비스 제공자들은 사용자들의 편의성과 욕구를 충족시키고자 하는 노력이 지속적으로 이루어져왔다(Voorhees, 2006). 이런 노력의 일환으로 문서분류(document classification)가 여러 연구자들에 의해서 널리 진행되고 있다(Sebastiani, 2002; Berger and Merkl, 2004).

일반적으로, 문서분류 시스템은 기계학습(machine learning) 방법을 많이 이용하는데 기계학습을 이용한 문서분류 시스템은 주어진 문서를 자질벡터(feature vector)로 표현하고, 표현된 벡터를 입력으로 하여 어떤 문서의 한 범주 혹은 여러 범주를 출력한다. 여기서 자질(feature)이란 어떤 대상을 다른 것과 구분할 수 있게 하는 고유한 특성들이 될 수 있는 것이다. 한편 용어(term)는 문서에 대한 특별한 정보를 의미하고, 때때로 용어 자체가 자질로 이용된다. 문서분류에서 문서의 특별한 정보가 될 수 있는 것은 문서를 이루고 있는 성분인 단어(word)이다. 그 문서의 저자 또는 문서의 발행처 등 또한 단어들로 이루어져 있으므로 문서분류에서는 단어가 고유 특성이 된다. 즉, 문서분류에서 용어들은 단어 그 자체이거나 여러 개의 단어가 될 수 있으며 이들이 자질들로서 사용된다.

문서분류는 이들 자질벡터를 이용하여 문서들과 범주(class)들 간의 유사도를 계산하여 어떤 범주로 분류될 것인지를 결정한다. 자질들이 많으면 성능이 더욱 좋겠지만, 자질이 많아지면 자질벡터의 크기도 커지게 된다. 벡터가 커지면 커질수록 계산량이 증가하여 실행속도가 떨어지는 단점이 있다(Yang and Pedersen, 1997; Forman, 2003). 본 논문에서는 이와 같은 이율배반적인(trade-off) 문제의 절충안을 찾기 위해 문서분류의 성능을 그다지 떨어지지 않으면서 속도를 개선할 수 있는 방법을 제안하고자 한다. 일반적으로 문서분류에 사용된 자질은 문서의 한 속성임에는 틀림없으나, 어떤 자질은 문서분류에 유용하지 않을 수 있다. 따라서 문서분류에 유용하지 않는 자질들을 선별하여 자질집합에서 제외시킴으로써 자질벡터의 크기를 줄일 수 있다. 자

질벡터의 크기를 줄이는 방법으로 널리 사용되는 방법은 자질추출(feature extraction)이다(Yang and Pedersen, 1997; Forman, 2003; Eyheramendy and Madigan, 2005; Pedersen, 1996; Pedersen et al., 1996). 자질추출은 문서의 내용과 관계없는 문자들을 제거하는 문서 전처리(document pre-processing), 문서의 내용을 전달하는 단어를 추출하는 자질생성(feature generation), 생성된 자질들 중에서 응용 분야에 적합한 자질만을 선택하는 자질선택(feature selection) 단계로 나누어진다.

본 논문은 온톨로지를 이용해서 문서분류에 적합한 두 종류의 자질추출 방법을 제안한다. 첫 번째 방법은 문서의 단어들을 온톨로지 정보에 따라 각 단어를 대표단어로 치환함으로써 자질선택 전에 자질의 후보를 줄이는 방법이고, 두 번째 방법은 자질선택 후 자질들을 온톨로지 정보에 따라 대표단어로 치환하여 자질벡터를 줄이는 방법이다. 또한 다양한 자질선택과 다양한 문서분류 방법을 이용하여 여러 환경에서 온톨로지를 적용한 자질추출 방법에 대한 성능을 평가하고 실험결과를 통해 온톨로지를 적용한 자질추출이 문서분류의 성능에 미치는 영향을 분석한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 자질추출 방법들과 온톨로지와 U-WIN(User-Word Intelligent Network)에 대하여 기술하고, 3장에서는 구현된 온톨로지를 적용한 자질추출 시스템에 대하여 기술한다. 4장에서는 본 논문에서 제안된 방법과 기존의 방법들의 성능을 평가하고 비교 분석한다. 마지막으로 5장에서는 결론 및 향후 연구를 제시한다.

제 2 장 관련 연구

본 장에서는 문서표현 방법에 대해서 개략적으로 설명하고 문서분류에 관련된 자질 추출 방법에 대하여 자세히 기술한다. 그 다음으로 본 논문에서 사용될 문서분류 방법을 간단히 소개하고, 온톨로지의 개념과 응용 사례 그리고 U-WIN에 대해서 간단히 소개한다.

2.1 문서표현

일반적으로 정보검색이나 문서분류 등의 분야에서 문서의 표현은 단어의 집합으로 표현된다. 이와 같은 모델링 방법을 벡터공간 모델링(vector space modeling)이라고 하며 모든 N개의 문서는 M개의 단어로 구성된 것으로 가정하고 이를 행렬로 표현한다. 이 행렬을 문서 행렬(document matrix) D라고 하며 $\{w_{ij}\}_{N\times M}$로 표현된다. 여기서 w_{ij}는 i번째 문서의 j번째 단어의 가중치를 의미한다. 이 절에서는 M의 크기를 정하기 위한 자질추출 방법과 w_{ij}를 추정하기 위한 가중치 추정 방법에 대해서 기술한다.

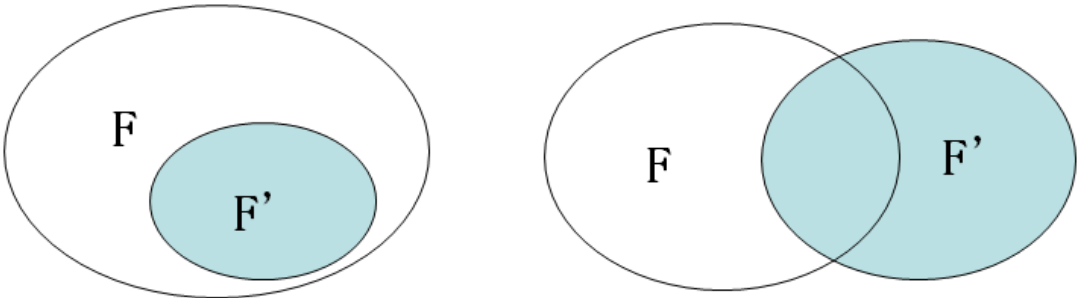
(1) 자질추출

자질이란 어떤 대상을 식별하기 좋게 하는 특성을 말한다. 자질추출은 주로 이미지 처리 분야에서 많이 사용한다. 이미지를 식별할 때 이미지 식별을 위해 좋은 특성은 어떤 것인지를 밝혀내는 작업이다. 이미지의 경우는 색과 모양 등이 자질이 될 수 있다. 이와 달리 자연어로 작성된 문서는 사람이 이해할 수 있는 글자로 이루어져 있고 이들은 단어들로 구성된다. 따라서 문서분류에서는 단어들이 자질이 될 수 있다.

문서를 자질벡터로 표현할 경우 자질은 단어(word or term)이다. 이 경우 자질의 수는 수천 개에서 수십만 개에 달한다. 벡터의 크기가 크면 클수록 기계학습에서의 학습 및 실행 시간이 길어지게 된다. 벡터의 크기를 줄이기 위해서는 단순히 문서에서 단어를 추출하는 방법 이외에 주어진 문제에서 가장 적절한 자질집합(feature set)을 선택하는 과정이 필요하다. 이 과정을 자질추출(feature extraction)이라고 한다.

그림 2.1은 자질추출과 자질선택의 차이를 나타낸 그림이다. 그림 2.1에서 보는 바와 같이 자질선택은 문서에 포함된 단어 혹은 자질 F의 부분집합 F'을 선택하는 과정

이고 자질추출은 자질 선택부분을 일부 포함한 문서 F의 의미를 그대로 전달하면서 임의의 형태로 가공된 새로운 자질집합 F'을 추출하는 과정이다.



(a) 자질선택 (b) 자질추출

그림 2.1 자질선택과 자질추출의 개념 차이

Figure 2.1 The difference between feature selection and extraction in concept

문서분류에서 자질을 추출하기 위한 일반적인 방법으로는 문서의 내용과 관계없는 문자들을 제거하는 문서 전처리(document pre-processing), 문서의 내용을 전달하는 단어를 추출하는 자질생성(feature generation), 생성된 자질들 중에서 응용 분야에 적합한 자질만을 선택하는 자질선택으로 나눌 수 있다. 문서 전처리는 불필요한 태그나 특수문자의 제거, 대소문자 통일(converting all letters to lower or upper case), 코드 정규화(Unicode normalization), 한자 정규화, 단어군집화 및 온톨로지 정규화, 형태소 분석(morphological analysis) 등이 여기에 속한다. 자질생성은 문서의 내용을 잘 표현할 수 있는 내용어(content word)를 찾기 위해 품사 태깅(part-of-speech tagging) 등을 통해 기능어에 해당하는 단어를 제거하는 방법을 통해 가능하다. 기능어에 포함되어 있는 단어(예를 들면 조사, 어미 등)는 대부분의 문서에서 모두 출현하여 문서의 분별을 떨어뜨리기 때문에 일반적으로 제거하며 이를 불용어 처리(stopword processing)라고 한다. 일반적으로는 문서의 내용을 표현하는 명사들을 내용어로서 많이 사용한다. 이후에 다양한 자질선택 방법들 중 하나를 거쳐서 문서의 내용을 가장 잘 표현하는 임의의 개수의 단어를 선택한다. 모든 문서에 대해 이와 같이 선택된 자질들을 하나의 집합으로 표현함으로써 M을 결정한다. 자질선택 방법은 2.2절에서 자세히 기술한다.

(2) 가중치 추정

가중치는 식 (2.1)과 같이 세 부분, 단어가중치(term weight, local weight) l_{ij}, 문서집합가중치(collection weight, global weight) g_{ij}, 정규화가중치(normalized weight) n_{ij}으로 나누어 생각할 수 있다(Salton and Buckley, 1988). 일반적으로 가중치로써 tf-idf 가중치(Salton and McGill 1983)를 많이 사용한다. tf-idf 가중치에서 단어가중치는 단어빈도(term frequency, tf)를 사용하고 문서집합가중치는 역문서빈도(inverse document frequency, idf)를 사용하며, 정규화는 사용하지 않는다.

$$w_{ij} = l_{ij} \times g_{ij} \times n_{ij} \tag{2.1}$$

단어가중치는 이진 가중치(Boolean weighting), 단어빈도(term frequency), 정규화된 단어빈도(normalized term frequency)로 나눌 수 있다. 이진 가중치는 가장 단순한 방법으로 자질 t_j가 문서 d_i에 출현했을 경우는 1로 표현하고 그렇지 않으면 0으로 표현하는 방법이다. 단어빈도 t_{ij}는 자질 t_j가 문서 d_i에 출현한 빈도수로 표현하는 방법이고 정규화된 단어빈도는 연구자들에 따라서 여러 가지 방법을 사용할 수 있으나 가장 대표적인 방법은 식 (2.2)와 같다.

$$l_{ij} = 0.5 + \frac{0.5 \times t_{ij}}{\max_j t_{ij}} \tag{2.2}$$

문서집합가중치는 역문서빈도(inverse document frequency, idf)와 확률역문서빈도(probabilistic inverse document frequency, p-idf)로 나눌 수 있다. 역문서빈도는 식 (2.3)과 같고, 확률역문서빈도는 식 (2.4)와 같다. 여기서 df_j은 t_{ij}가 출현한 문서수를 의미하며, 문서빈도(document frequency)라고 한다.

$$idf_{ij} = \log \frac{N}{df_j} \tag{2.3}$$

$$p-idf_{ij} = \log \frac{N - df_j}{df_j} \tag{2.4}$$

정규화가중치는 문서길이에 따른 정규화(normalization over document length)를 의미하며 주로 유클리디안 벡터길이를 이용한 코사인(cosine) 정규화를 사용한다(식 (2.5)).

$$n_{ij} = \frac{1}{\sqrt{\sum_{j=1}^{M} w_{ij}^2}} \tag{2.5}$$

2.2 자질선택

(1) 이원 분할표

대부분의 자질선택 방법들은 이원 분할표(contingency table)를 이용한다. 이원 분할표는 범주형 변수 간의 관계를 살피는데 유용하다. 이원 분할표의 각 값은 도수 또는 상대도수를 나타내며 귀무가설(null hypothesis)을 기반으로 그 값을 구하게 된다. 표 2.1는 용어 t와 범주 c_i 사이의 이원 분할표이다.

표 2.1 이원 분할표

Table 2.1 Contingency table

		문서 범주(class)		
		c_i	$\sim c_i$	
용어(term)	t	A(n11,tp)	B(n12,fp)	A+B(n1p)
	$\sim t$	C(n21,fn)	D(n22,tn)	C+D(n2p)
		A+C(np1)	B+D(np2)	N(npp)

A는 용어 t와 범주 c_i가 서로 발생한 문서의 수로서 참 긍정(true positive, tp, n11)이라고 한다. 용어 t의 범주가 제대로 속한 경우이다. B는 용어 t가 범주 c_i가 아닌 다른 범주에서 발생한 문서의 수로서 거짓 긍정(false positive, fp, n12)이라한다. 용어 t를 c_i가 아닌 다른 범주에 속한 경우이다. C는 용어 t

가 아닌 다른 용어들이 범주 c_i에서 발생한 문서의 수로서 거짓 부정(false negative, fn, n21)이라고 한다. 마지막으로 D는 N−(A+B+C)로 구하면 되고 용어 t가 아닌 용어들이 c_i가 아닌 범주에서 나타난 문서 수이다. N은 모든 문서의 수이다. 또한 m11은 용어 t와 범주 c_i가 독립적일 때, t와 c_i가 동시에 발생할 평균 빈도이며 식 (2.6)과 같다.

$$m11 = \frac{np1 \times n1p}{npp} \tag{2.6}$$

(2) 자질선택 방법

자질선택은 매우 다양한 방법이 연구되었으며, 본 논문에서 다루는 방법은 표 2.2와 같다(Yang and Pedersen, 1997; Forman, 2003; Eyheramendy and Madigan, 2005; Pedersen, 1996; Pedersen et al., 1996).

표 2.2는 용어 t와 범주 c_i사이의 관련도를 구하는 방법이다. 관련도가 높은 임의의 개수의 용어가 자질로 선택된다. 표 2.1을 이용하여 표 2.2의 식들에 대입하면 자질선택 방법들의 값을 구할 수 있다. 예를 들어 Pointwise Mutual Information(PMI)의 경우 Mutual Information(MI)인 식 (2.7)을 이용하여 식 (2.8)과 같이 구한다. 표 2.2의 A를 용어 t와 c_i가 동시에 나타난 빈도 $m11$로 나누면 MI의 값이다. MI에 로그(log)를 취하면 PMI의 값을 구할 수 있다.

$$MI = \frac{n11}{m11} \tag{2.7}$$

$$PMI = \log(\frac{n11}{m11}) \tag{2.8}$$

표 2.2 자질선택 방법

Table 2.2 Feature selection methods

자질선택 방법	수식
Document Frequency (DF)	$n11$
Bi-Normal Separation(BNS)	$\|F^{-1}(n11/np1) - F^{-1}(n12/np2)\|$ 1)
Chi-Squared (X^2)	$npp*(n11*n22 - n21*n12)^2/(n1p*np1*np2*n2p)$
Probability Ratio(PR)	$np2*n11/(np1*n12)$
Odds Ratio(Odds)	$n11*n22/(n12*n21)$
Pointwise MI(PMI)	$\log(n11/m11)$
Total MI(TMI)	$n11/npp*\log(n11/m11) + n12/npp*\log(n12/m12) + n21/npp*\log(n21/m21) + n22/npp*\log(n22/m22)$
Information Gain(IG)	$e(np1, np2) - (\frac{n1p}{npp}*e(n11, n12) + \frac{n2p}{npp}*e(n21, n22))$ 2)
t-Score	$(n11 - m11)/\sqrt{(n11)}$

2.3 문서분류

문서분류는 주어진 문서에 미리 정해진 하나 혹은 그 이상의 범주(category)를 할당하는 과정이다(Sebastiani, 2002). 하나의 문서는 여러 개의 범주에 속할 수도 있다. 또한 경우에 따라서는 미리 정해진 범주에 전혀 속하지 않을 수도 있다. 문서분류 시스템은 일반적으로 그림 2.4(고영중 & 서정연, 2002)와 같다. 이 장에서는 본 논문에서 사용되는 문서 분류기에 대해서만 간단히 기술한다. 많은 연구자들은 문서분류의 정확률을 높이기 위해서 꾸준히 노력하고 있으며 일반적으로 정확률을 높이는 연구들은 실행 시간을 떨어뜨리게 한다. 정보를 가진 자질이 많으면 많을수록 정확률은 올라가지만 벡터 크기가 커짐으로 인해 계산시간이 많이 걸리기 때문이다. 즉 정확률과 실행시간이 서로 이율배반적이다. 기존의 많은 연구들은 정확률을 높이는 데 치중하였으나 최근 문서분류가 정보검색 시스템의 한 부분으로 사용되면서 실행시간도 무시할 수 없는 중요한 연구로 등장하게 되었다. 실행시간을 고려한 연구의 대부분은 자질벡터의 크기를 줄이는 것에 초점을 맞추고 있다(Forman, 2003; Yang and Pedersen,

1) $F(\bullet)$는 정규누적분포함수(normal cumulative distribution function)이고 $F^{-1}(\bullet)$는 $F(\bullet)$의 역함수이다.

2) $e(x,y)$는 엔트로피로서 $-\frac{x}{x+y}log_2\frac{x}{x+y} - \frac{y}{x+y}log_2\frac{y}{x+y}$를 의미한다.

1997). 본 절에서는 네 가지의 문서분류기에 대하여 기술한다(Manning et al., 2007).

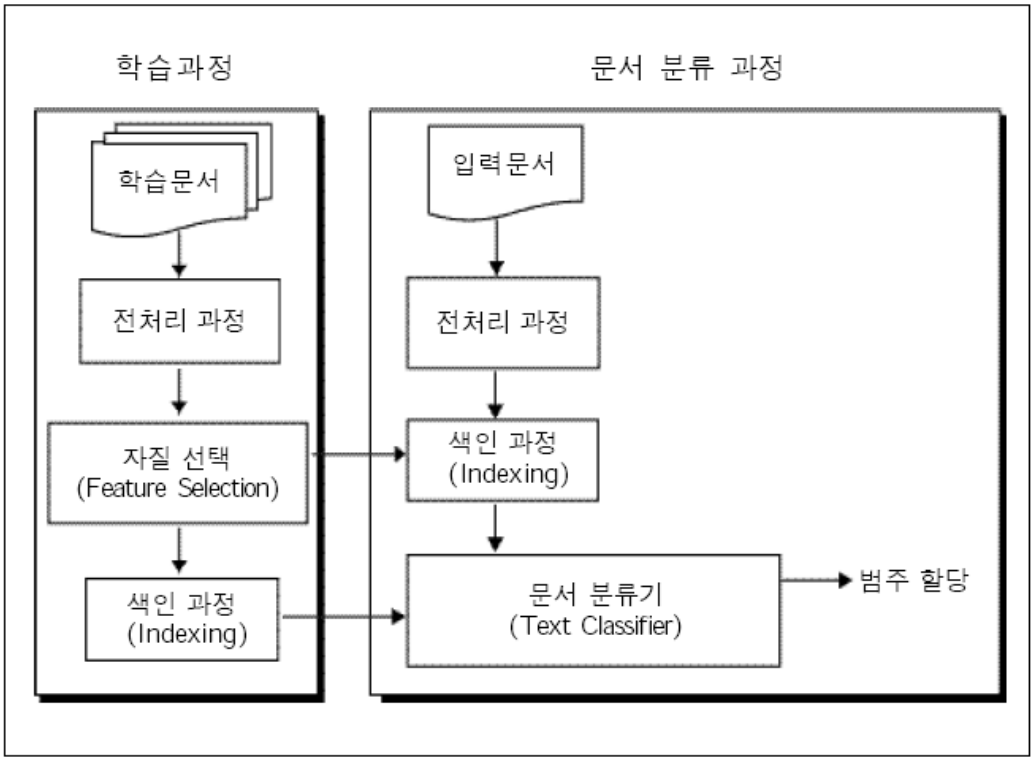


그림 2.2 일반적인 문서분류 시스템의 구성도

Figure 2.4 System configuration of a general document classification system

(1) Rocchio

Rocchio 분류기는 각 범주 c_i을 잘 표현하는 하나의 가중치 벡터 w_i를 학습하고 학습된 벡터 w_i와 주어진 용어 t_j로 이뤄진 문서 벡터 $d = <t_1, \ldots, t_M>$과의 코사인 유사도를 계산하여 그 유사도가 가장 큰 범주를 주어진 문서의 범주 c로 결정하는 방법이다(식 2.9 참조)(Joachims, 1997). 가중치 벡터 w_i를 학습하는 방법은 정보검색 분야에서 사용되는 연관 피드백(relevance feedback)에서 질의어를 수정하는 방법과 유사한 방법으로 긍정적인 영향을 주는 자질에는 가중치를 더하고 부정적인 영향을 주는 자질에는 가중치를 빼는 방법으로 가중치를 수정한다. 여기서 연관 피드백이란 검색질의에 대한 결과집합의 적합성 여부를 판단하여 검색 순위(Ranking)에 영향을 미치도록 하는 작업이다(Buckley et al., 1994).

$$c = \arg max_{c_i} \frac{\sum_{j=1}^{M} w_{ij} * t_j}{\sqrt{\sum_{j=1}^{M} w_{ij}^2} + \sqrt{\sum_{j=1}^{M} t_j}} \quad (2.9)$$

(2) Naive Bayesian

확률모형으로 주어진 문서 d에 대해서 가장 높은 확률을 가지는 범주 c를 결정하는 방법이다(McCallum and Nigam, 1998). 문서 d는 용어 벡터 $<t_1, ..., t_M>$으로 표현되고 각 용어가 서로 독립적이라고 가정하면 식 (2.10)과 같이 정의된다.

$$\begin{aligned} c &= \arg max_{c_i} \Pr(c_i|d) \quad (2.10) \\ &= \arg max_{c_i} \Pr(c_i)\Pr(d|c_i) \\ &= \arg max_{c_i} \Pr(c_i)\prod_{j=1}^{M} \Pr(t_j|c_i) \end{aligned}$$

(3) k-Nearest Neighbor(kNN)

주어진 문서 d에 가장 유사한 k개의 문서 중에서 가장 많은 범주 c를 주어진 문서 d의 범주로 간주한다. k가 1일 경우에는 주어진 문서 d의 범주 c는 학습 문서(training documents)들 중에서 d와 가장 유사한 문서가 가지는 범주가 된다. 유사도는 일반적으로 코사인 유사도(cosine similarity)를 사용한다. 이를 개념적으로 정리하면 식 (2.11)과 같이 정의된다. $\cos(d_j, d)$는 주어진 문서 d와 유사한 k개의 문서 d_j와의 코사인 값이다.

$$c = \arg max_{c_i} \sum_{d_j \in S_k} I_{c_i}(d_j) \times \cos(d_j, d) \quad (2.11)$$

여기서 S_k는 주어진 문서 d에 대한 k개의 가장 가까운 문서 집합이고, 함수 $I_{c_i}(d_j)$의 값은 문서 d_j가 범주 c_i에 속하면 1이고 그렇지 않으면 0이다. 이처럼

개념적으로 매우 간단하지만 실행 시간 동안에 주어진 문서와 모든 학습 문서들과의 유사도를 계산해야 하기 때문에 실행 속도가 매우 늦을 수 있다. 이에 따라, 유사도 계산 방법과 실행 속도를 개선하기 위해서 다양한 연구들이 진행되었다(Daelemans and van den Bosch, 2005).

(4) Support Vector Machine(SVM)

SVM은 퍼셉트론(Perceptron)과 같은 선형분류기(linear classifier)이다. 일반적인 선형분류기는 결정 경계(decision boundary)로서 직선을 사용한다. 이 경우에는 주어진 학습 데이터에 대해서 여러 개의 가능한 결정 경계를 찾을 수 있다. 그러나 SVM은 이러한 결정 경계를 초월평면(hyperplane)으로 결정하여 최적의 결정 경계를 정하는 방법이다(Joachims, 1998). 대부분의 초월평면이 선형적으로 분리할 수 없기 때문에 비선형적(non-linearly)인 자질공간을 선형적인 자질공간으로 사상하기 위한 특별한 함수를 이용한다. 이 함수를 커널 함수(kernel function)이라고 한다. 이 함수는 자질공간의 특성에 따라서 서로 다른 함수가 사용될 수 있다. 기본적인 이진 분류함수(binary classification function)는 식 (2.12)과 같이 매우 단순하나, w_j와 b학습 방법은 좀 더 복잡하다(Joachims, 2002).

$$c = \arg max_{c_i \in \{+1,-1\}} \sum_{j=1}^{M} sign(w_j \times t_j - b) \qquad (2.12)$$

2.4 온톨로지

온톨로지(ontology)의 일반적인 의미는, 우주 안에 어떤 종류의 실체들이 존재하는가에 관한 연구 또는 관심을 말한다. 어원은 '실재'라는 의미의 그리스어 'onto'와 '논문 또는 강연' 등의 의미를 갖는 'logia'의 합성어로부터 유래되었으며, 제일 원리 또는 사물의 본질에 관한 연구를 추구하는 학문이다. 정보기술에서의 온톨로지는, 전자상거래와 같이 지식의 어떤 특정 영역 내에 있는 실체 및 상호작용의 작업 모델을 의미한다. 이 용어에 대한 일반적인 해석으로 "온톨로지란 어떤 관심 분야를 개념화하기 위해 명시적으로 정형화한 명세서"라는 (Gruber, 1993)의 정의를 가장 많이 인용하고 있다.

가장 먼저 온톨로지 개념을 적용한 컴퓨터 분야는 지식표현과 활용을 연구하는 인공지능 분야이다. 특히 인공지능 분야의 에이전트 분야는 이미 1990년대 초부터 분산된 환경에서 에이전트들이 상호작용을 통해 의미 있는 문제를 해결하기 위해서는 서로 공유할 수 있는 기본 지식기반이 필요하다는 것을 인식하면서 온톨로지를 사용하기 시작하였다[3]. 이에 따라 일종의 온톨로지라고 할 수 있는 개념계층도(concept hierarchy) 등을 이용했으며, 지식과 정보를 교환하기 위한 질의어(KQML)(Finin, et al., 1997)와 지식교환형식(KIF)(Genesereth and Fikes, 1992) 등을 정의했다(Lee, 1998). 특히 미국방연구성(DARPA)의 DAML-OIL(DARPA Agent Markup Language - Ontology Inference Layer)은 대표적인 온톨로지 표현 언어 및 형식으로 받아들여지고 있다(Connolly, et al., 2001).

또 다른 대표적인 분야인 정보검색에서는 용어모음이나 동의어사전과 같은 적은 범위의 형태만으로도 불필요한 오류를 방지할 수 있고 검색효율을 높일 수 있다(정재헌, 1995). 예를 들어 사용자가 잘못 기재한 '문서불류'라는 키워드는 온톨로지를 이용해 '문서분류'로 바로잡을 수 있고, 'document classification', '도큐먼트 클래시피케이션', '클러스터링', 'clustering' 등과 같은 유사 또는 관련어를 이용해 더욱 풍부한 검색서비스를 제공할 수 있게 된다.

언어 공학에서의 온톨로지는 어휘, 용어, 어휘목록, 사전, 전문분야사전 등과 같은 어휘집합을 기반으로 하여 시소러스, 어휘망, 어휘분류 등을 포함하는 개념과 관계와 속성들이 내부적으로 형성된 상위의 지식구조 체계라 할 수 있으며 그림 2.2와 같이 온톨로지에 대한 인식범위를 설정할 수 있다(최호섭 외 4명, 2006). 온톨로지는 그림 2.3과 같이 나무구조로 표현된다.

3) http://www.etnews.co.kr/news/detail.html?id=200306230148

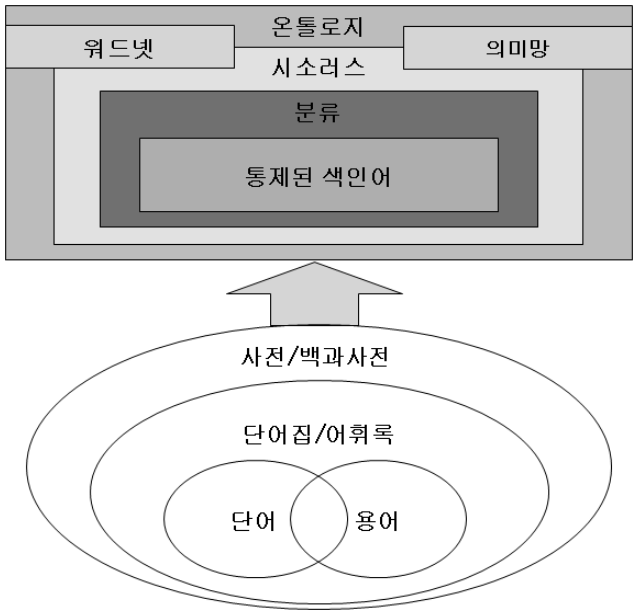


그림 2.3 온톨로지에 대한 적용 범위

Figure 2.3 The coverage of ontology

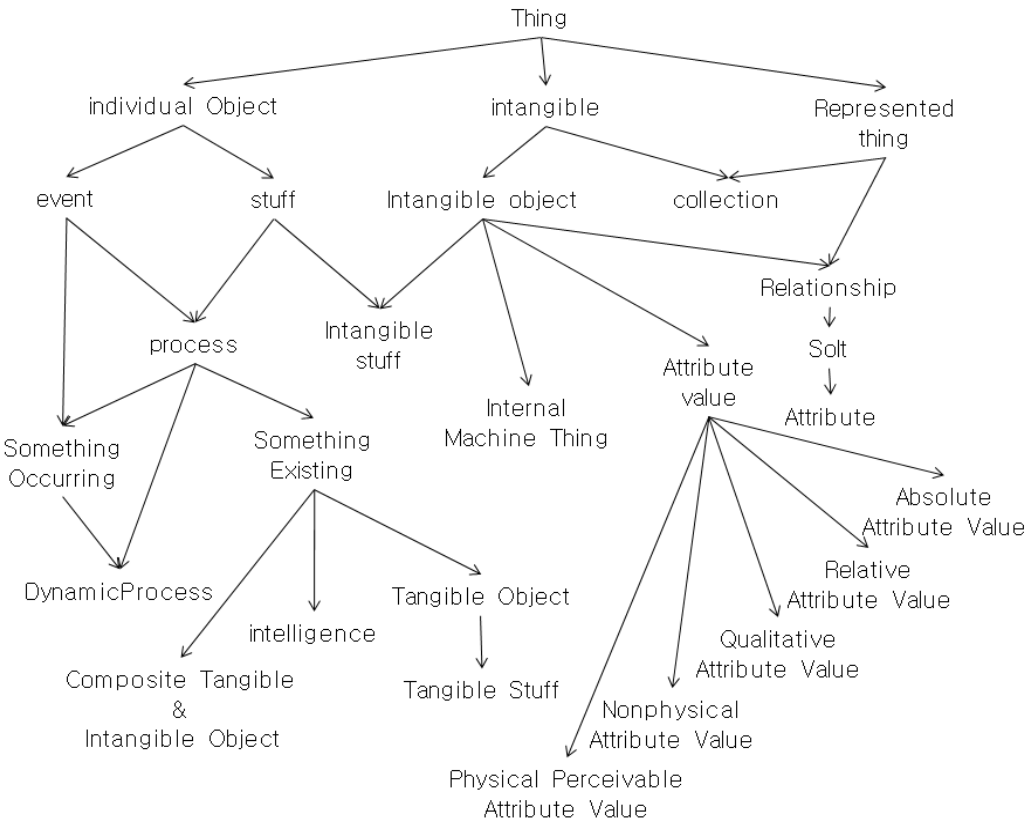


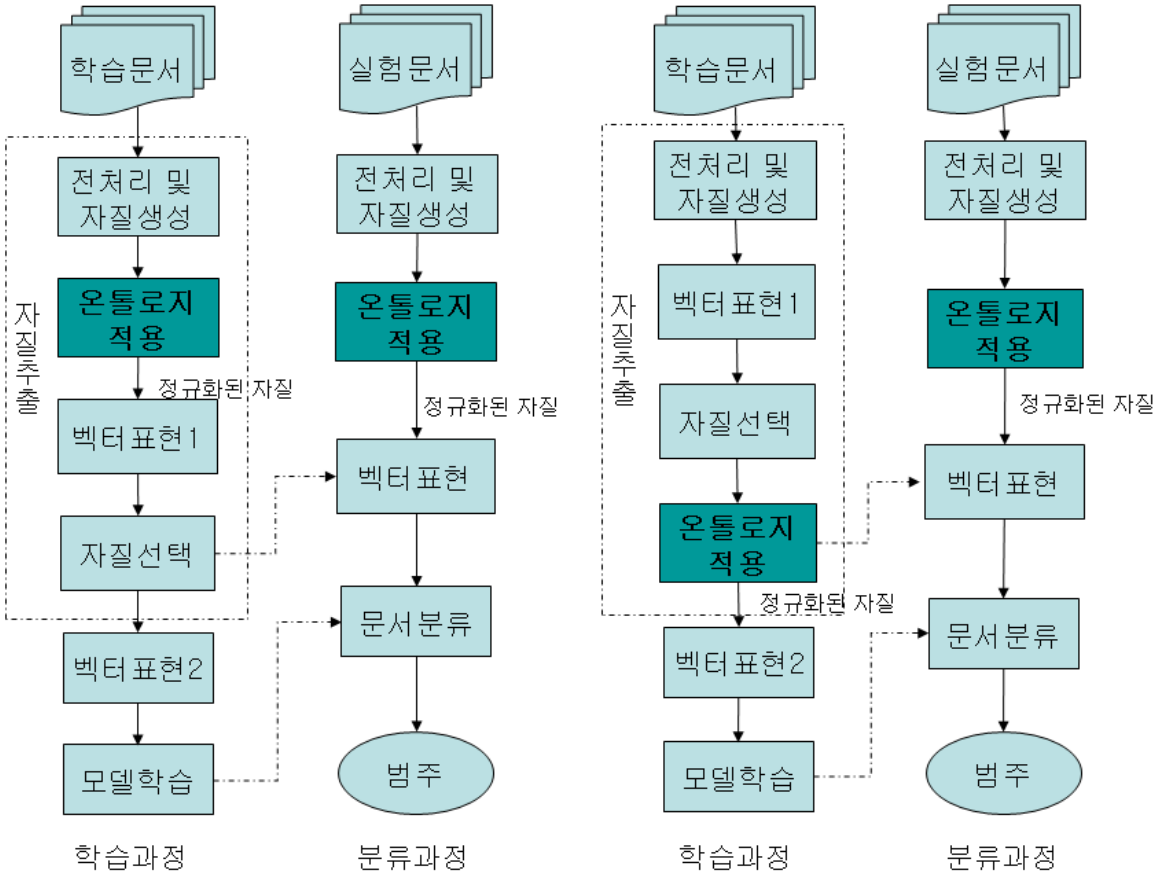
그림 2.3 온톨로지에 대한 데이터 구조의 예

Figure 2.3 An example of a data structure on ontology

본 논문에서는 대규모 어휘 데이터베이스이자 의미망인 U-WIN(User-Word Intelligent Network)을 실험에 사용하였다. U-WIN은 울산대 한국어처리연구실에서 2002년부터 지금까지 꾸준히 개발 중인 어휘망이며, 한국어의 공통적이고 개별적인 속성을 바탕으로 한국인의 보편적인 인지 체계와 개념 관계를 파악하여 이를 어휘의 의미적·개념적 연결 구조로 형성한 어휘망이다(최호섭 외 4명, 2006). U-WIN은 언어 교육용시스템, 자동어휘학습시스템, 복합명사 자동 생성 및 뜻풀이 생성 기술, 전문분야별 개념 체계 자동 생성 기술, 정보검색 등 다양한 기술에 활용되고 있다(최호섭 외 4명, 2006). 구축 대상은 한국어 어휘 전체로 명사, 동사, 형용사가 핵심적인 대상이다. 부수적인 대상은 부사, 관형사, 감탄사, 조사, 수사, 의존명사 등이며, 기타 정보적 대상으로는 북한어, 방언, 옛말, 전문용어, 고유명사, 어근, 어미 등이다. 본 논문은 U-WIN 어휘 표기 방법 중 동의어, 유의어, 상위어의 의미 관계를 이용한다.

제 3 장 온톨로지를 적용한 자질추출

본 논문은 자질추출 과정에 온톨로지를 적용하는 방법을 제안하며 제안된 방법의 성능을 평가하기 위해 2장에서 기술한 분류기와 자질선택기를 이용한 문서분류 시스템을 이용한다.



(a) 방법 1 (b) 방법 2

그림 3.1 제안된 문서분류 시스템의 구성도

Figure 3.1 Our system configuration for document classification

지도 학습을 기반으로 한 문서분류 시스템은 학습과정을 거치기 때문에 본 논문의 시스템은 그림 3.1과 같이 학습과정과 분류과정으로 나뉜다. 학습과정은 확률기반의 문서분류 시스템이 학습 데이터를 기반으로 각 범주를 잘 나타내는 자질을 학습하는 과정이고, 분류과정은 실험 데이터를 학습된 결과를 이용해 문서를 분류하여 범주를

결정하는 단계이다. 학습과정의 전처리 과정부터 자질선택 과정까지가 자질추출 과정이다. 본 논문은 Perl 언어를 사용하여 시스템을 구현하였으며, 2장에서 기술한 분류기와 자질선택기는 CPAN(The Comprehensive Perl Archive Network)[4]에서 제공하는 Perl 모듈을 사용하였다.

온톨로지를 적용하는 방법은 그림 3.1과 같이 두 가지 방법을 제안한다. 첫 번째 방법으로 자질선택 이전에 전처리의 마지막 단계에서 온톨로지를 이용하여 단어들을 치환한다(그림 3.1 (a) 참조). 이 방법은 개념적으로 같은 단어(자질)를 자질추출 전에 온톨로지를 적용하여 정규화함으로써 양질의 자질벡터를 자질선택 과정에 제공하는 예를 들면 용어 '대규모집적회로'를 자질추출 전에 온토로지의 상위어인 '집적회로'로 치환하는 방법으로 정규화한다. 이렇게 함으로써 문서의 의미는 거의 손상시키지 않고 양질의 자질을 추출하고자 하는 방법이다. 두 번째 방법은 자질선택 이후에 선택된 자질들을 온톨로지를 이용하여 추출된 단어들을 치환한다(그림 3.1 (b) 참조). 이 방법은 자질선택 방법에 의해서 양질의 자질을 먼저 선택하고 선택된 자질을 개념적으로 정규화하는 방법이다. 이렇게 함으로써 첫 번째 방법보다 자질 수를 더 줄일 뿐 아니라 불필요하게 많은 단어에 대해서 온톨로지를 접근하는 문제를 해결할 수 있다.

이하에서는 그림 3.1의 시스템의 구성요소인 전처리 및 자질 생성, 온톨로지 적용, 벡터표현, 문서분류에 대해서 자세히 기술하고자 한다.

4) http://www.cpan.org

3.1 전처리 및 자질생성

본 논문에서 자질벡터 생성과정은 그림 3.2에서 보는 바와 같이 내용어 추출과 온톨로지 적용 과정으로 나눈다. 이 과정은 학습 문서는 물론 실험 문서에도 그대로 적용되는 과정이다. 온톨로지 적용은 3.2절에서 자세히 기술하고 이 절에서는 내용어 추출에 대해서 기술한다.

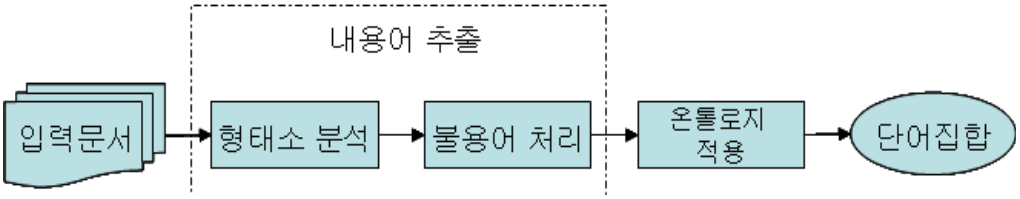


그림 3.2 온톨로지를 적용한 전처리

Figure 3.2 Preprocessing applying ontology

(1) 내용어 추출

문서의 내용이나 특징을 잘 반영하는 단어가 내용어이며, 주로 명사에 해당하는 단어들이다. 본 논문에서는 주어진 문서에 대해 형태소 분석기(morphological analyzer)를 사용하여 각 단어의 품사를 결정한다. 그 후 문서의 특성을 반영하지 못하는 불용어인 접속사, 어미, 대명사, 조사 등을 제거한다. 이 과정이 전처리 및 자질생성 과정이며 이를 내용어 추출 과정이라고도 한다. 그림 3.3은 내용어 추출 과정의 예를 보여주고 있다. 그림 3.3에서 문장 “쾌적한 환경을 제공하기 위한 조명은 보건 위생 및 작업 능률 등에 커다란 영향을 미친다.”에 대한 자질인 내용어는 {‘쾌적’, ‘환경’, ‘제공’, ‘조명’, ‘보건’, ‘위생’, ‘작업’, ‘능률’, ‘영향’}이다.

에너지절약을 위한 조도기준 설정에 관한 연구
쾌적한 환경을 제공하기 위한 조명은 보건 위생 및 작업 능률 등에 커다란 영향을 미친다. 그러나 적절하지 못한 조명은 전술한 내용을 약화시키며 에너지 낭비를 초래하게 된다. 그러므로 적절한 조도기준이 필요하다. 본 연구에서는 세계 각국의 조도기준 자료를 수집 분석하고 우리나라 사람을 대상으로 한 실험결과를 적용하여 우리나라에 적합한 조도기준을 설정하였다. 본 연구의 결과는 조도설계 및 관리의 효율화 및 작업능률의 극대화를 이룰 수 있으며, 조명 에너지 관리의 효율화를 통한 에너지 절감의 효과를 얻을 것이라 기대된다.

(a) 내용어 추출 전

에너지절약 조도기준 설정 연구
쾌적 환경 제공 조명 보건 위생 작업 능률 영향 적절 조명 전술 내용 약화 에너지 낭비 초래 적절 조도기준 필요 연구 세계 각국 조도기준 자료 수집 분석 우리나라 사람 대상 실험결과 적용 우리나라 적합 조도기준 설정 연구 결과 조도설계 관리 효율 작업능률 극대 조명 에너지 관리 효율 에너지 절감 효과 기대

(b) 내용어 추출 후

그림 3.3 전처리 및 자질생성의 예

Figure 3.3 An example of preprocessing and feature construction

3.2 온톨로지 적용

일반적으로 말은 여러 가지 상황에 따라서 같은 개체에 대해서 다양한 단어로 표현한다. 예를 들어, '아버지'라는 개체에 대해서 상황에 따라 '아빠', '부친', '가친' 등의 단어로 표현된다. 이들은 개념적으로 '아버지'라는 단어로 통일해서 사용할 수 있다. 또한 집단을 표현하는 단어(상위어)는 다른 여러 단어(하위어)의 의미를 내포할 수 있다. 예를 들어, 단어 '가축'에는 단어 '개', '소', '돼지', '닭' 등과 같은 단어의 의미를

포함하고 있다. 이와 같이 단어들 사이에는 서로 복잡한 관계를 지니고 있다. 이를 잘 표현하는 것이 온톨로지이다. 본 논문에서는 온톨로지에서 동의어, 유의어, 상위어 관계를 이용해서 자질집합을 정규화한다. 즉 문서분류에서 불필요한 자료의 중복을 피하기 위해서 이들 관계에 속하는 단어를 하나의 단어로 정규화함으로써 자료의 중복을 줄일 수 있다. 이에 따라 문서의 단어를 상위어 또는 동의어, 유의어로 치환하여 용어의 수를 줄이고, 선택되는 자질의 질을 높이고자 한다.

본 논문의 시스템에서는 U-WIN의 동의어, 유의어 의미 관계 구조 및 상위어 관계 구조의 온톨로지를 이용하여 동의어, 유의어, 상위어 집합의 단어들을 한 단계 상위의 대표 단어로 치환한다. 어느 수준의 상위어를 사용하느냐는 또 다른 문제를 야기시킬 수 있다. 예를 들어 그림 3.4에서 단어 '에코'와 '자외선'의 공통 상위어(common hypernym)은 '파동'인데 이들 두 단어는 파동이라는 공통의 속성을 지니기는 하지만 문서를 분류함에 있어서는 매우 다른 속성을 지닌 단어이다. 따라서 단어 '에코'를 포함하는 문서와 단어 '자외선'을 포함하는 문서의 분야가 서로 구별되어야 할 것이다. 이와 같은 문제를 완화시키기 위해서 본 논문에서는 단말 단어의 바로 위의 상위어를 이용하는데 상위어의 단어가 여러 개일 경우는 랜덤하게 하나로 고정한다. 예를 들면 그림 3.4에서 자질에 해당하는 단어로서 '에코'가 나타나면 '음파'와 '장치'중 하나를 포기하게 된다.

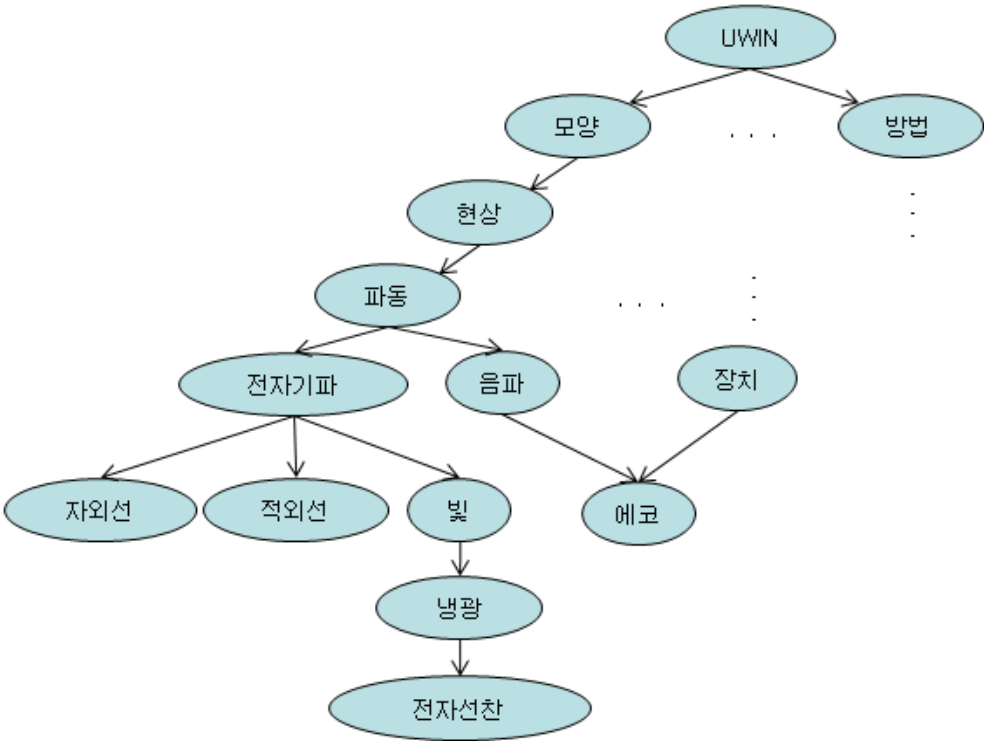


그림 3.4 U-WIN 온톨로지의 일부

Figure 3.4 A part of a hierarchy ontology, U-WIN

표 3.1은 본 논문에서 사용한 상위어 관계의 구체적인 예를 보이고 있으며, 왼쪽은 하위어들이고 오른쪽은 상위어들이다. 즉 문서에서 하위어에 속한 단어들이 출현하면 오른쪽의 상위어로 치환하여 문서를 정규화하였다. 그림 3.1에서 볼 수 있듯이 온톨로지 적용 과정은 자질선택 전 혹은 후에 적용될 수 있으며, 이 두 과정은 모두 같은 기능을 가지고 있다.

표 3.1 U-WIN 온톨로지의 일부

Table 3.1 A part of ontology U-WIN

상위어 관계		유의어/동의어 관계	
하위어	상위어	하위어	상위어
세포모양체계계산기	컴퓨터	라디오텔레폰	무선전화기
자동자료처리장치	컴퓨터	라디오폰	무선전화기
반복구조형계산기	컴퓨터	무선방위측정국	무선나침국
활동계산기	컴퓨터	무선방위탐지국	무선나침국
대규모집적회로	집적회로	반송대조방식	되돌림비교방식
거대규모집적회로	집적회로	반송조합방식	되돌림비교방식
초고밀도집적회로	집적회로	부호해독기	해독기
쌍극성집적회로	집적회로	디코더	해독기
하이브리드아이시	집적회로	정전차폐	자기차폐
진단프로그램	프로그램	정전기차폐	자기차폐
처리프로그램	프로그램	가려막기	자기차폐

3.3 자질선택

자질선택 과정은 전처리 및 자질생성 과정의 결과를 입력으로 하여 문서분류에 적합한 자질을 선택하여 새로운 자질벡터를 생성하는 과정이다. 2.4절에서 설명했듯이 N개의 모든 문서는 M개의 단어로 구성되었다고 가정하여 이를 하나의 행렬로 표현하는데 이때 M은 자질선택을 통해 결정할 수 있다. 자질선택 과정은 범주 정보를 가장 잘 표현하는 용어를 선택하는 과정이다. 본 논문에서는 2장에서 설명한 자질선택 기법들을 이용하여 추출된 자질로 실험데이터의 문서를 표현하고, 분류기를 이용하여 범주를 결정하고 성능을 평가한다.

3.4 벡터표현

벡터표현은 주어진 문서를 하나의 벡터로 표현하는 과정이다. 자질선택 과정에서 선택된 자질을 용어로 사용하기 위해서 문서에서의 단어의 순서는 문서의 내용을 표현함에 있어 영향을 끼치지 않는다고 가정하면 문서는 단순히 단어들의 집합이 된다. 선택된 각각의 자질들이 문서 내에서 나타난 빈도 등을 각 자질의 가중치로 하여 각

자질들과 문서에 대한 벡터를 만들 수 있다. 학습 문서들은 N개의 문서와 M개의 용어를 이용한 문서 행렬 D는 $\{w_{ij}\}_{N\times M}$로 표현된다. 이 행렬은 Perl 모듈[5]에서 제공되는 KnowledgeSet이라는 자료 구조를 이용한다. 그림 3.4는 문서벡터 표현의 예이다. 단어 가중치로서 단어 출현빈도만 사용하였을 때 문서에 대한 벡터표현의 예를 나타낸다. 본 논문에서 자질의 가중치는 2장에서 설명한 가중치 중 단어의 빈도(term frequency)와 역문서 빈도(inverse document frequency)만 사용하고 정규화하지 않았다.

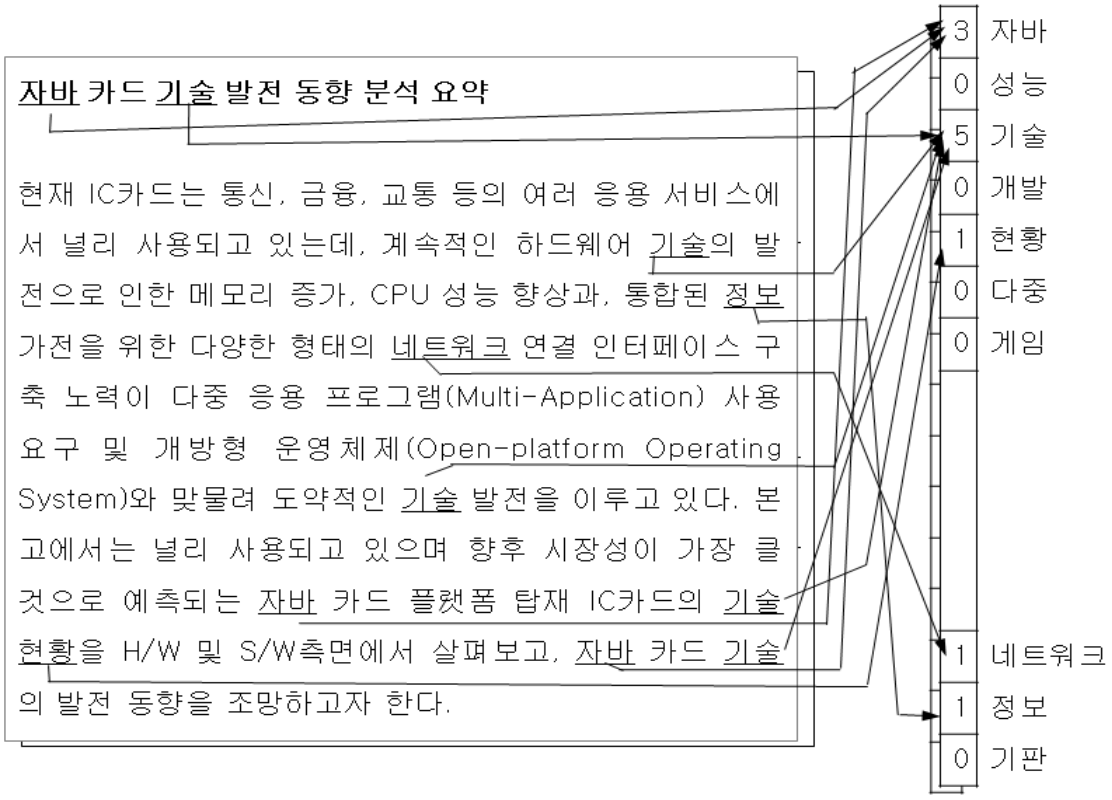


그림 3.4 문서에 대한 벡터표현의 예

Figure 3.4 An example of a vector representation of a document

3.5 문서분류

앞 절에서 구해진 자질벡터를 바탕으로 문서와 범주간의 유사도를 계산하여 문서에 알맞은 범주를 결정하는 과정이다. 본 논문에서는 2.4절에서 기술한 방법들을 이용하

5) AI::Categorizer

며, Perl 모듈 AI::Categorizer에서 제공되는 것을 사용하여 온톨로지를 적용한 자질 추출과 적용하지 않은 자질추출의 성능을 평가한다. 성능 평가는 각각의 경우의 문서 분류 시스템의 성능을 평가함으로써 온톨로지가 자질추출에 어떻게 영향을 끼치는지를 분석한다.

제 4 장 실험 및 평가

이 장에서는 실험 환경과 평가 방법을 설명하고 실험 결과를 통해서 온톨로지를 적용한 자질추출 방법의 성능을 분석하고자 한다.

4.1 실험 환경

실험 대상 문서는 한국과학기술정보연구원(KISTI)에서 제공한 것으로 전기공학, 전자공학, 통신공학, 컴퓨터/정보공학 관련 논문들의 제목과 요약을 각 주제별로 분류한 문서들이다. 전체 문서 분류와 개수는 표 4.1과 같다.

표 4.1 실험 데이터의 상세 분류와 통계 자료

Table 4.1 Test classes and their statistics

주제 분류코드	주제	문서개수	
		학습	실험
G01	전기공학	29	6
G03	전자공학	128	22
G04	통신공학	103	28
G05	컴퓨터공학/정보공학	975	242
Total		1536	

실험에서 사용하는 분류기와 자질선택기는 CPAN에서 제공하는 모듈인 AI::Categorizer에서 제공되는 것을 사용하였다. 모듈에서 제공하지 않는 자질선택기(Odds, PR, BNS, IG, PMI, TMI, t-Score)는 본 논문에서 구현하여 사용하였다. 실험에서 사용하는 온톨로지는 KISTI의 문서인 전기/전자/통신/컴퓨터/정보분야의 내용을 기반으로 작성된 울산대학교 한국어처리연구실의 온톨로지 U-WIN의 일부인 유의어, 동의어, 상위어 관계를 이용하고 상위어 관계는 단말 단어의 한 단계 위에 있는 상위어만 적용한다.

4.2 평가 방법

성능 평가에 가장 일반적으로 사용되는 방법은 정확률(accuracy)과 오류율(error rate)이다. 정확률은 전체 평가에서 정답의 비율을 구하는 것이고 오류율은 전체 평가에서 오답의 비율이다. 정확률은 표 4.2의 이원 분할표를 이용해서 구할 수 있으며, 구체적인 정의는 식 (4.1)과 같다. 본 논문의 모든 평가는 정확률만을 이용한다.

표 4.2 성능 평가를 위한 이원 분할표

Table 4.2 Contingency table for performance analysis

		정답	
		맞음	아님
시스템	맞음	a	b
	아님	c	d

정확률(accuracy) $A = \frac{a+d}{a+b+c+d}$ (4.1)

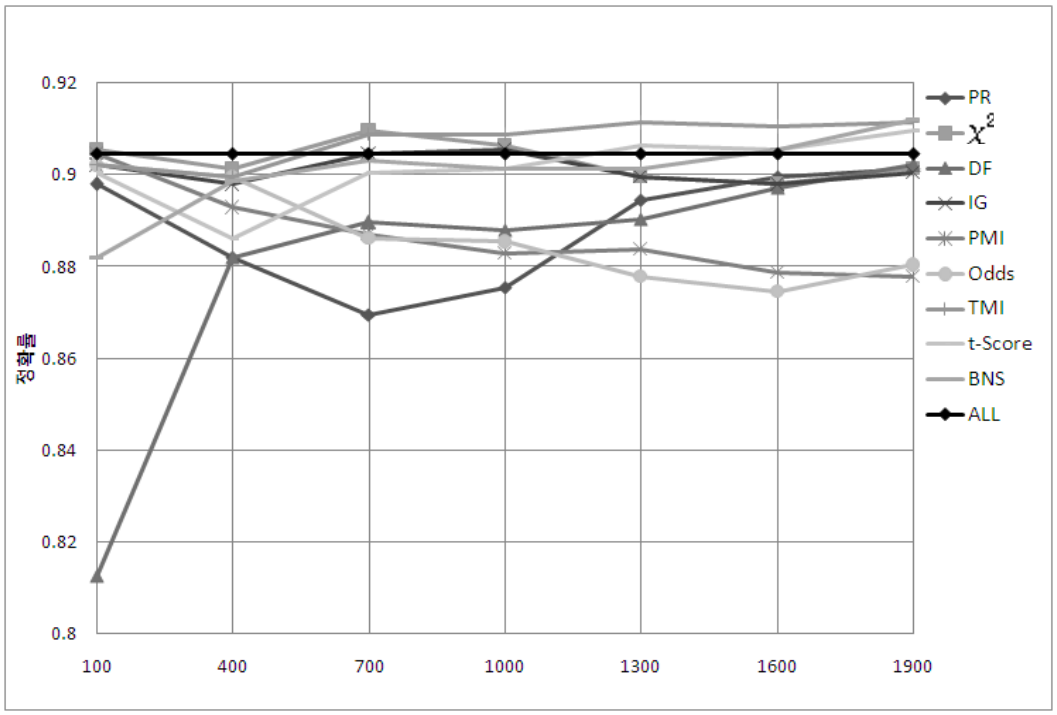
4.3 성능 평가 및 분석

본 논문은 크게 자질선택과 온톨로지 적용에 따른 네 가지의 경우로 나누어 실험하였다. 첫째는 자질을 선택하지 않고 온톨로지도 적용하지 않는 방법(Raw), 둘째는 자질을 선택하고 온톨로지를 적용하지 않는 방법(Feat), 셋째는 자질을 선택하기 전에 온톨로지를 적용하는 방법(beforFeatOnto), 넷째는 자질을 선택한 뒤 온톨로지를 적용하는 방법(afterFeatOnto)이다. 이는 크게 온톨로지의 비적용(Raw, Feat)과 적용(beforFeatOnto, afterFeatOnto)으로 구분된다. 이하에서는 온톨로지의 비적용(기존 시스템)과 적용(제안된 방법)으로 나누어 기술하고자 한다.

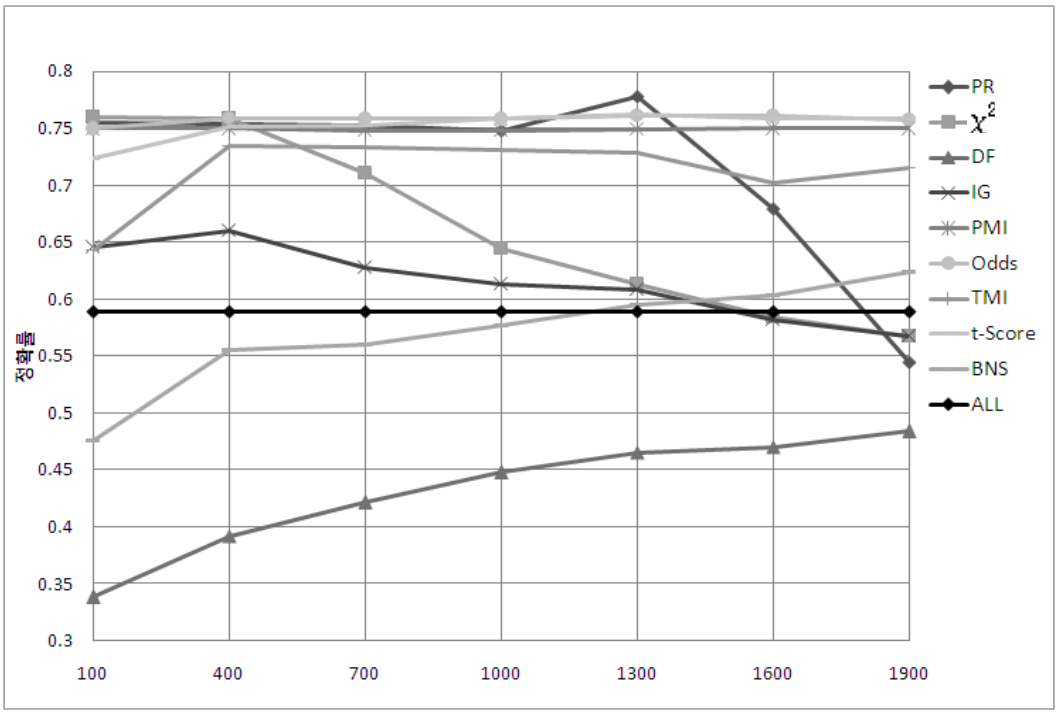
(1) 온톨로지의 비적용

그림 4.1는 온톨로지를 적용하지 않을 경우, 각 자질선택 방법에 대한 성능 변화를 나타낸 정확률 그래프이다. 그림 4.1에서 ALL은 온톨로지를 적용하지 않고 자질도 선택하지 않을 경우의 성능 변화를 보이는 그래프이다. 분류기와 자질 선택 방법에 따라 조금씩 다른 결과를 보이지만 많은 경우는 자질선택 방법을 사용하는 것이 좋은

성능을 보였다. 특히 kNN의 경우는 대부분의 자질선택 방법(Feat)이 자질을 선택하지 않을 경우(Raw)보다 좋은 성능을 보였다. 또한 자질벡터의 크기가 클수록 일반적으로 좋지 않은 성능을 보이고 있으며 평균적으로 보았을 때 400 ~ 700개 정도에서 좋은 결과를 보이고 있었다. 이와는 반대로 Naive Baysian의 경우는 대부분의 자질선택 방법(Feat)이 자질을 선택하지 않을 경우(Raw)보다 좋지 않은 성능을 보였다. 모든 단어들을 자질로 했을 경우 성능이 비교적 더 높았다. 이는 확률 기반의 학습 모델로써 학습 모델의 경우 학습 대상이 많으면 많을수록 좋은 값이 나오게 되는 특성 때문이다. Rocchio의 경우도 많은 자질선택 방법이 자질을 선택하지 않을 경우보다 좋은 성능을 보였으며 자질선택 방법에 따라 다소 성능의 변화가 심하게 나타나고 있었다. SVM의 경우는 자질벡터의 크기에 관계없이 비슷한 성능을 보였으며, 자질선택 방법이나 자질벡터의 크기에 크게 영향을 받지 않았다. 이는 자질선택 방법을 사용하여 자질벡터의 크기를 크게 줄여도 성능에는 큰 영향을 주지 않는다는 사실을 간접적으로 말하고 있다. 그림 4.1의 결과를 종합하면 자질선택 방법이 분류기에 매우 의존적임을 알 수 있었다. 즉 Naive Baysian을 사용할 경우에는 실행 속도의 문제가 존재하지 않는 한 자질선택 방법을 사용하지 않는 것이 적합하며, SVM과 kNN을 사용할 경우에는 가능하면 자질선택 방법을 사용하는 것이 매우 효과적이다. Rocchio의 경우에도 자질선택 방법을 사용하는 것이 효과적이기는 하지만 자질선택 방법에 따라 성능에 영향을 주므로 자질선택 방법을 선택함에 있어서 신중해야 할 것이다.



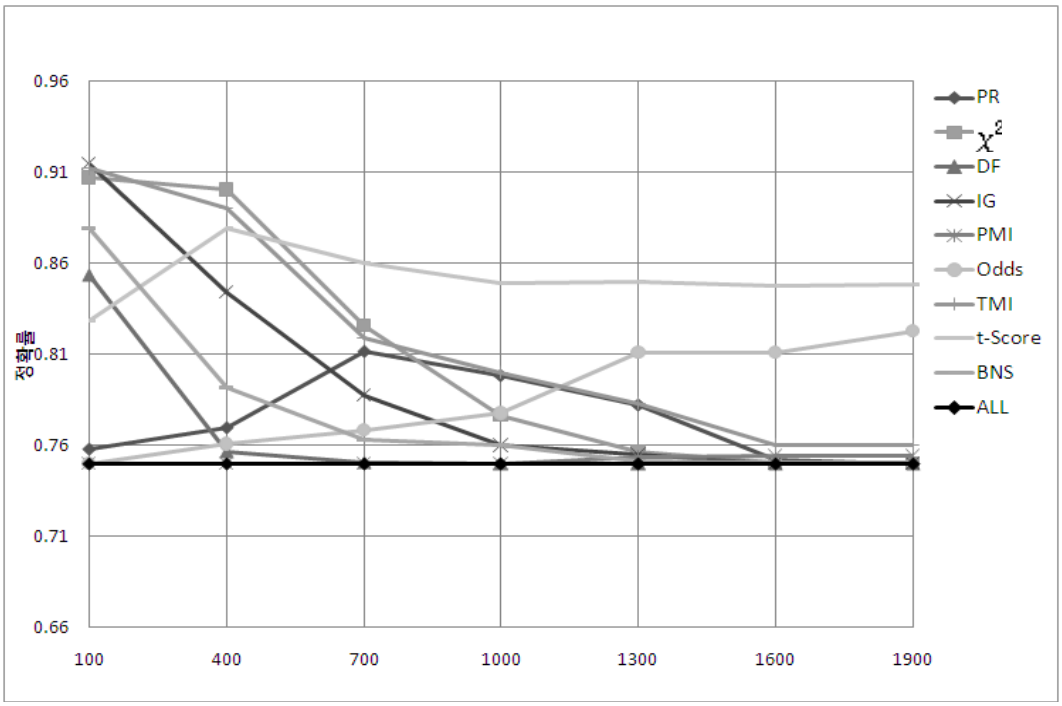
(a) Naive Bayesian



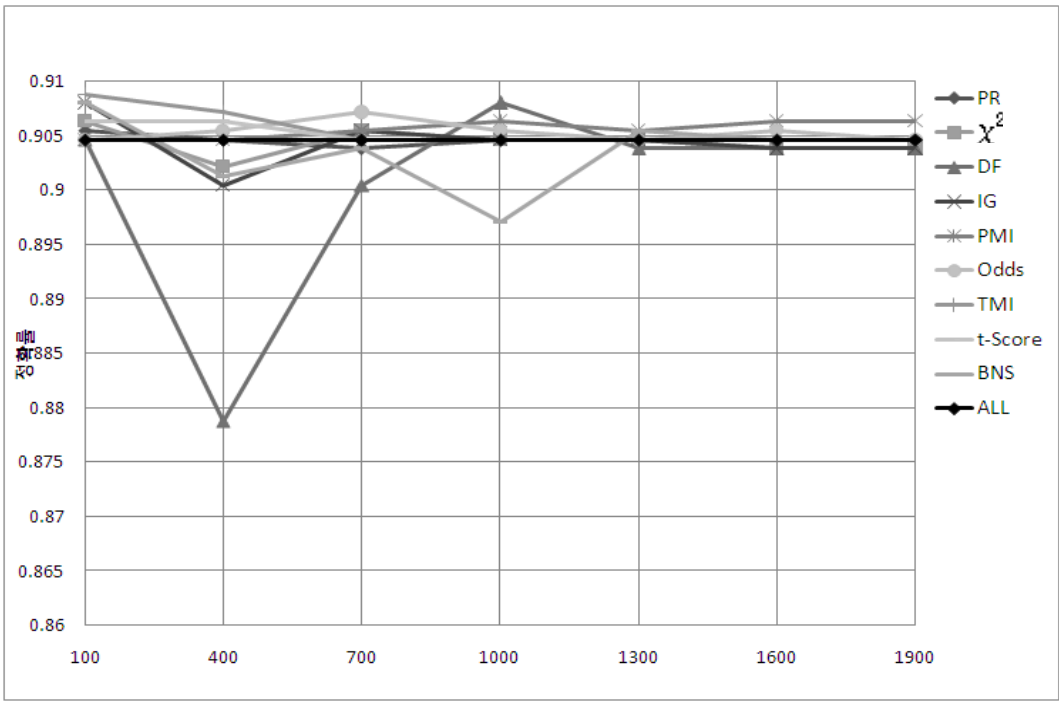
(b) Rocchio

그림 4.1 온톨로지를 적용하지 않은 자질선택기의 분류 성능

Figure 4.1 Performance variation of feature selectors without applying ontology



(c) kNN



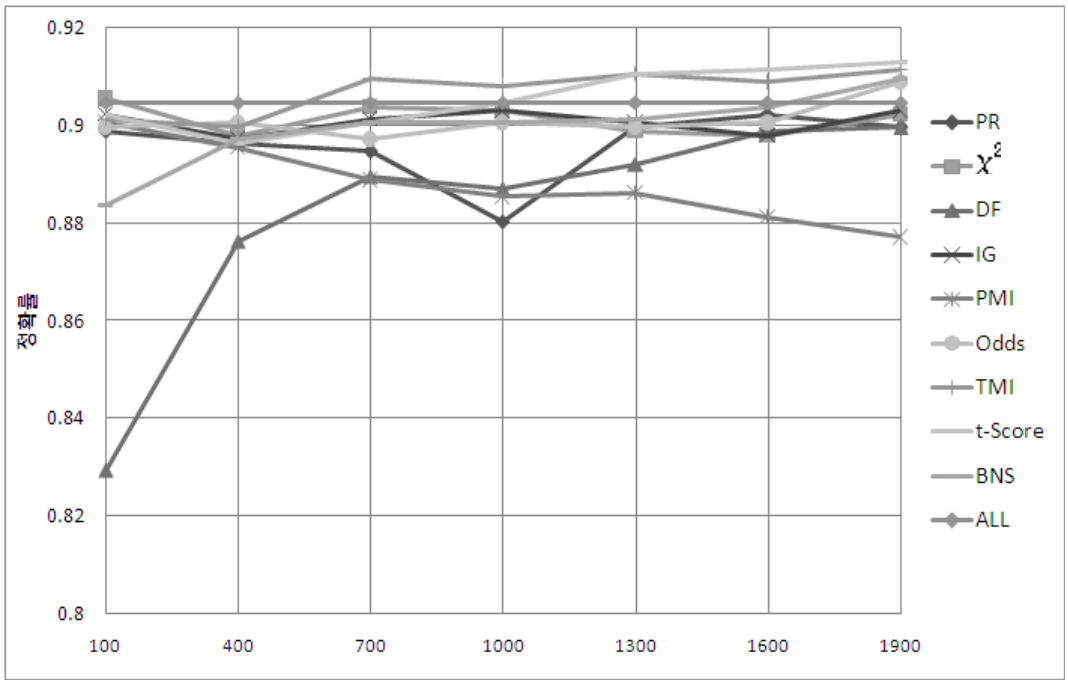
(d) SVM

그림 4.1 온톨로지를 적용하지 않은 자질선택기의 분류 성능(계속)

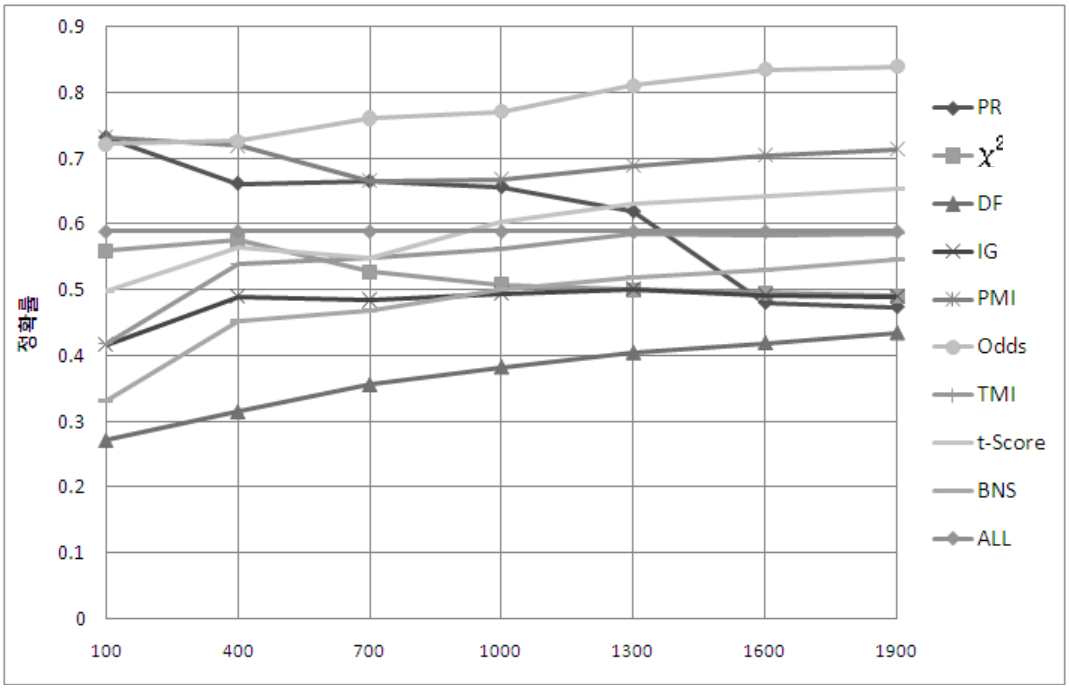
Figure 4.1 Performance variation of feature selectors without applying ontology(cont.)

(2) 온톨로지의 적용

온톨로지를 적용할 경우는 온톨로지의 적용시점에 따라 두 가지 방법이 있다. 즉 자질을 선택하지 전과 후에 온톨로지를 적용하는 방법이 있다. 그림 4.2는 자질을 선택한 후에 온톨로지를 적용하였을 때 각 분류기와 자질선택기 그리고 자질 수에 따른 성능을 보이고 있다. 그림 4.1에서와 같이 그림 4.2에서도 ALL은 기존 시스템의 자질을 선택하지 않은 경우의 성능을 나타낸다. 전체적으로 온톨로지를 적용했을 경우는 온톨로지를 적용하지 않았을 경우와 비슷한 성향의 결과를 보였다. 다만 kNN의 경우 자질선택에 따라서 다소 혼란스러운 결과를 보였다. 이는 실험 문서의 수가 다소 부족하여 보인 결과일 것으로 추측된다. 자질선택에 따른 결과는 Naive Baysian에서 t-Score의 정확률이 높았으며 Rocchio와 SVM, kNN의 경우는 Odds의 정확률이 높았다.



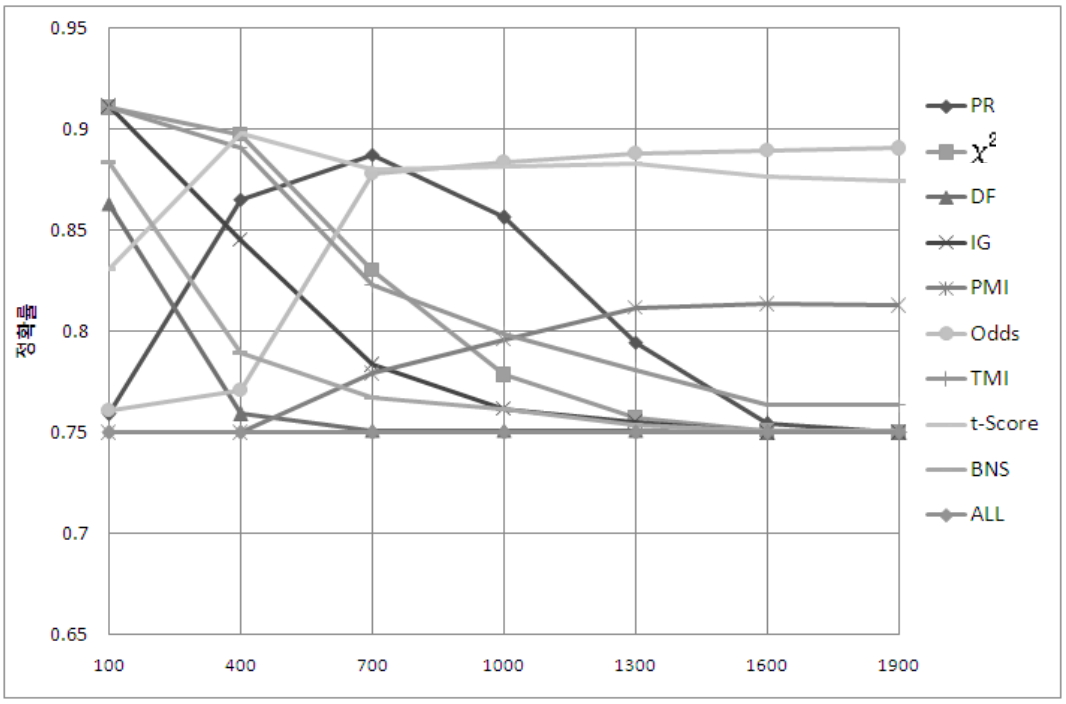
(a) Naive Bayesian



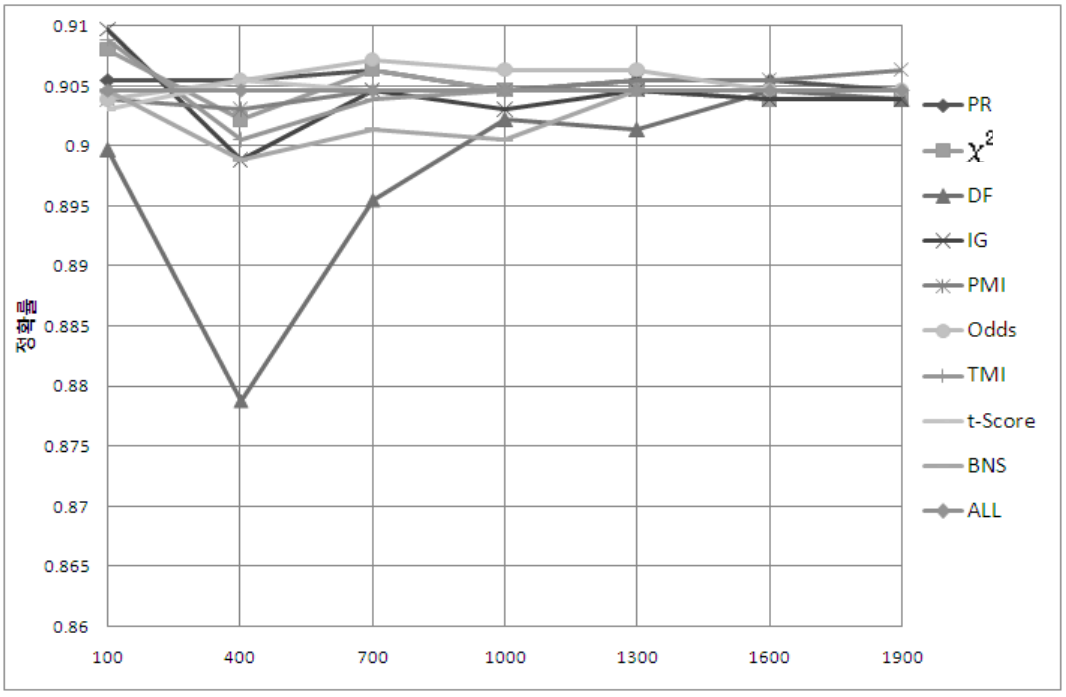
(b) Rocchio

그림 4.2 자질선택 후 온톨로지 적용시 자질선택의 분류 성능

Figure 4.2 Performance variation of feature selectors with applying ontology after feature selection



(c) kNN

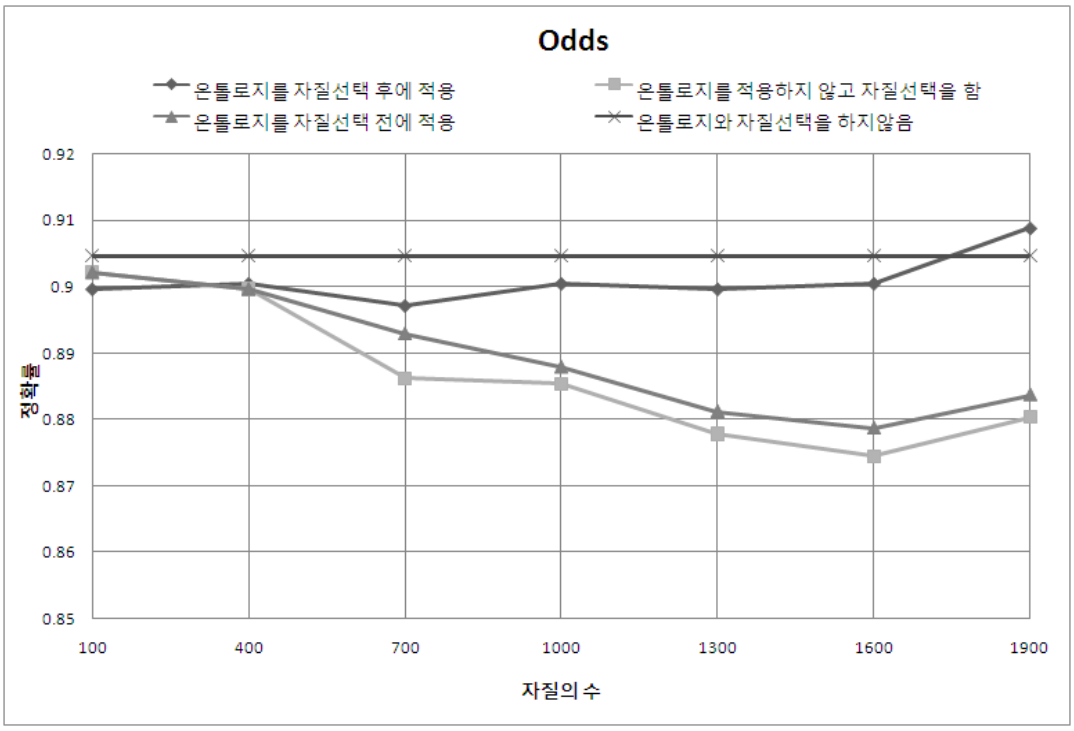


(d) SVM

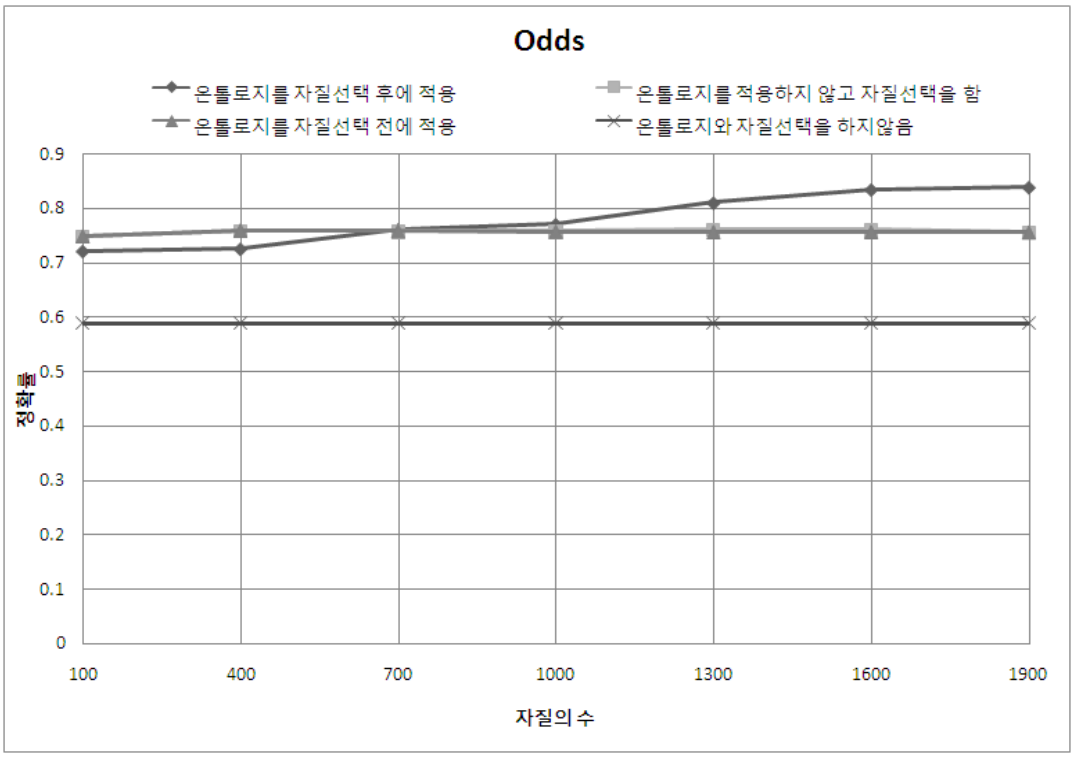
그림 4.2 자질선택 후 온톨로지 적용시 자질선택의 분류 성능(계속)

Figure 4.2 Performance variation of feature selectors with applying ontology after feature selection(cont.)

그림 4.3은 온톨로지 적용에 따른 성능의 변화를 관찰하기 위해서 자질선택 방법을 Odds로 고정시키고 각 분류기에 따른 성능을 분석한 결과이다. 자질선택만 적용할 경우(Feat)보다는 거의 모든 환경에서 온톨로지를 적용할 경우(befor/afterFeatOnto)가 좋은 성능을 보였다. 이와 같은 결과는 온톨로지의 적용이 좋은 자질을 추출하는 데 큰 도움이 됨을 알 수 있다. 또한 온톨로지의 적용은 자질추출 전과 후의 결과를 비교할 경우, SVM 분류기의 경우에 약간의 동요가 있었으나 대부분의 분류기의 경우에 자질추출 후 온톨로지의 적용의 성능이 더 좋았다. 이는 온톨로지를 적용하되 자질을 선택한 후 적용하는 것이 더 효과적임을 알 수 있다.



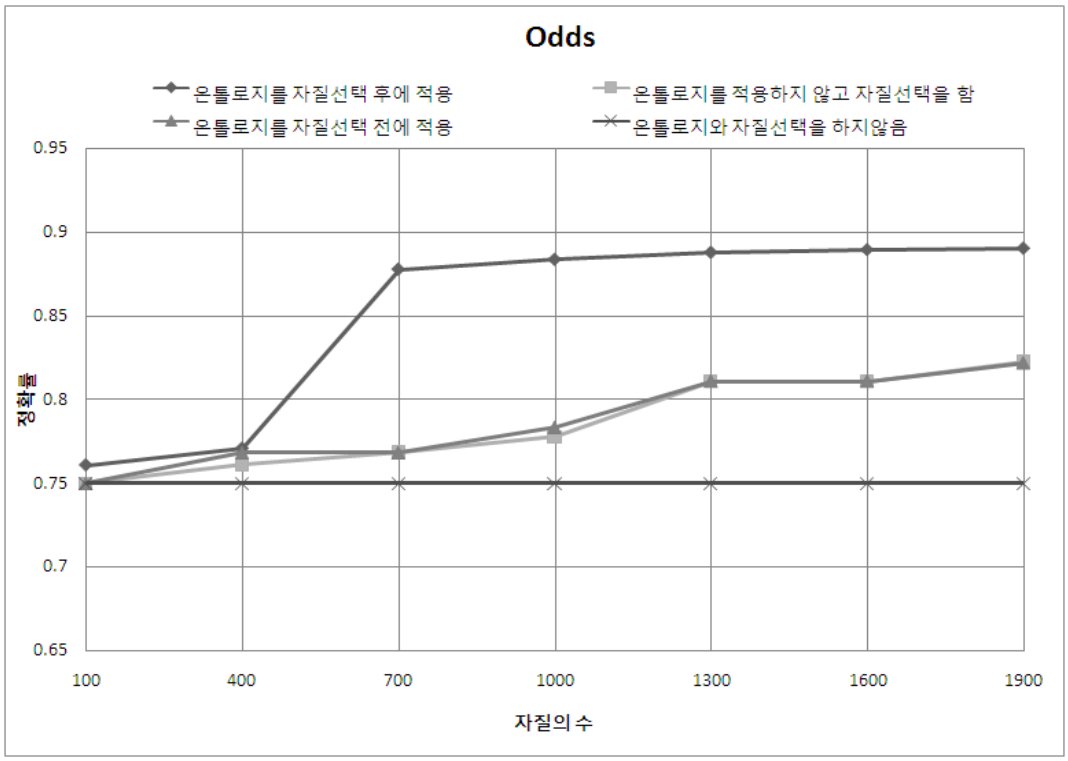
(a) Naive Bayesian



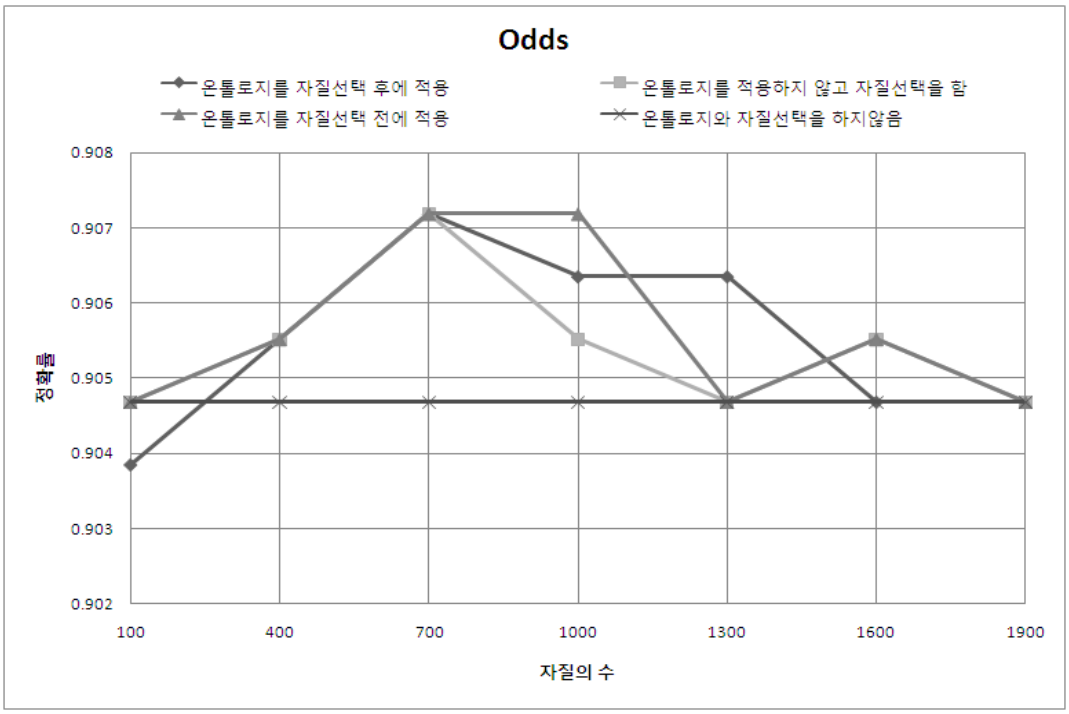
(b) Rocchio

그림 4.3 온톨로지 적용에 따른 분류 성능

Figure 4.3 Performance variation according to applying ontology



(c) kNN

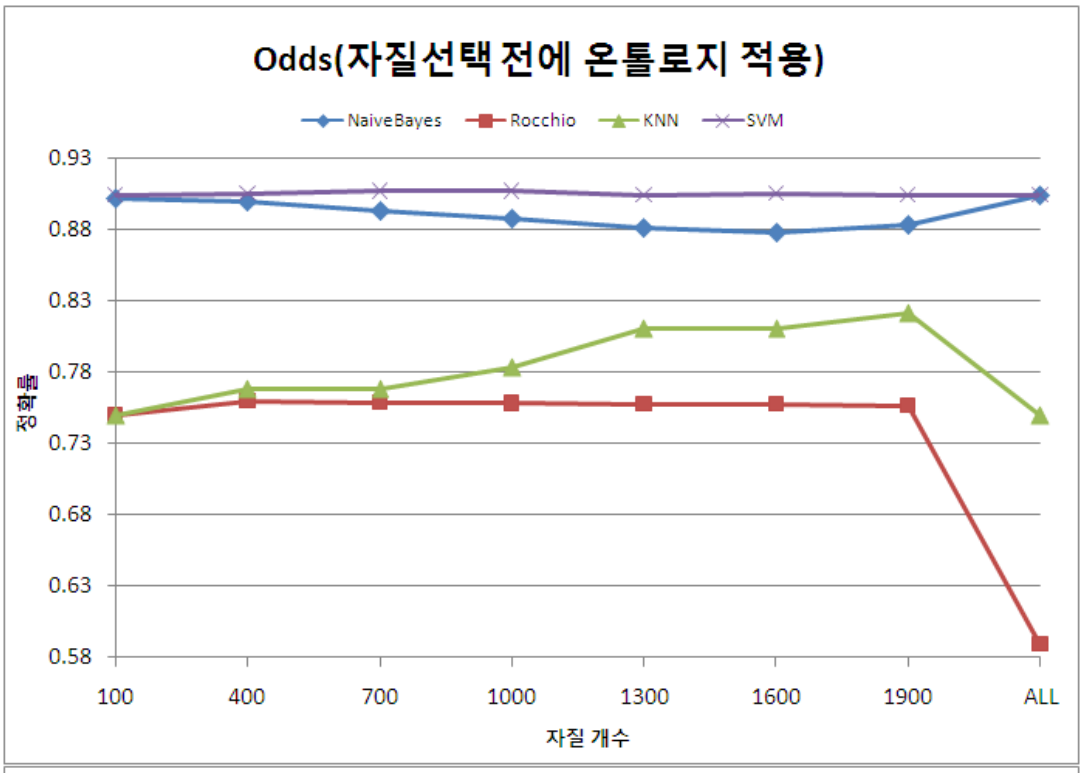


(d) SVM

그림 4.3 온톨로지 적용에 따른 분류 성능(계속)

Figure 4.3 Performance variation according to applying ontology(cont.)

그림 4.4는 온톨로지를 적용할 경우 분류기에 따른 성능 변화를 보이고 있다. 적용 시점에 상관없이 모든 경우에는 SVM이 가장 좋은 성능을 보였으며, 다음으로 Naive Baysian, kNN, Rocchio 순으로 좋은 성능을 보였다. 앞에서도 언급했지만 Naive Baysian의 경우를 제외한 모든 분류기에서 자질을 선택하거나 온톨로지를 적용하는 경우가 좋은 성능을 보였으나 Naive Baysian은 실행속도를 고려할 경우에는 자질선택이나 온톨로지를 적용하지 않는 것이 반드시 효과적이라고는 말할 수 없을 것이다. 자질선택 후에 온톨로지를 적용한 경우는 700개 정도의 자질로도 비교적 높은 성능을 낼 수 있다는 것을 알 수 있었다. 표 4.3은 3.1절에서 기술한 Odds 선택기에서 선택된 자질을 온톨로지를 적용하는 두 번째 방법을 적용한 후의 자질의 수이다. 표 4.3을 보았을 때 자질의 수가 커질수록 감소하는 폭이 커진다.



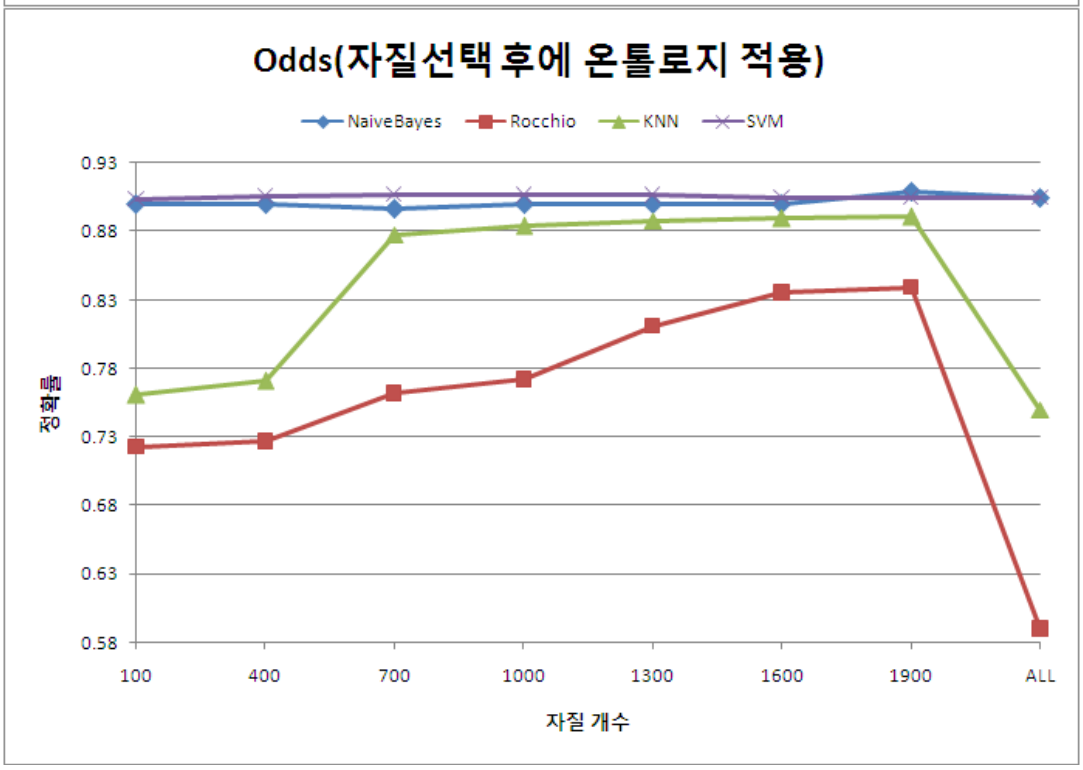


그림 4.4 분류기에 따른 분류 성능(자질선택: Odds ratio)

Figure 4.4 Performance variation according to classifiers based on the feature selector, Odds ratio

표 4.3 자질선택 후 온톨로지 적용에 따른 자질의 수 변화

Table 4.3 The variation of feature size according to applying ontology after feature selection

자질 수	적용후의 자질 수								
	Odds	PR	BNS	χ^2	DF	IG	PMI	TMI	t-Score
100	100	99	97	98	98	99	100	99	100
400	399	396	390	392	393	389	400	392	400
700	698	692	682	683	683	681	697	683	698
1000	995	987	981	977	972	976	997	978	996
1300	1293	1283	1276	1273	1261	1273	1296	1278	1296
1600	1591	1572	1568	1566	1556	1566	1596	1573	1596
1900	1888	1852	1863	1853	1847	1852	1896	1871	1896

(3) 성능 분석

본 논문의 실험결과 온톨로지를 적용하고 자질선택을 하였을 경우 상대적으로 기존의 온톨로지를 적용하지 않은 자질선택보다 높은 성능을 보였다. 온톨로지를 적용함에 있어서는 전처리단계에서 온톨로지를 적용하는 것보다 자질 선택 후에 선택된 자질에 온톨로지를 적용하는 방법이 좀 더 좋은 결과를 얻을 수 있었다. 선택한 자질의 수가 많아질수록 성능도 올라가고 제안한 두 번째 방법의 적용시의 줄어드는 자질의 수의 폭도 커졌다. Naive Baysian의 경우는 문서의 모든 용어를 자질로 하여 문서분류를 하는 것이 가장 성능이 높았고 SVM의 경우 분류기 자체의 성능자체가 좋아서 변화하는 폭이 거의 차이가 나지 않았다.

제 5 장 결론

본 논문에서는 온톨로지를 자질추출에 적용하는 방법을 제안했다. 기존의 다양한 분류 방법들과 자질선택 방법을 조합하여 실험을 하였다. 성능 평가는 문서분류의 정확도로 제안한 자질추출 방법의 성능을 평가하였다. 온톨로지는 실험에 사용된 문서들의 분야에서 추출된 전문용어들에 대한 것을 사용하였다.

실험 결과에 의하면 기존의 단순 자질추출보다 본 논문에서 제안한 유의어, 동의어, 상위어 의미구조의 온톨로지를 사용하여 자질을 줄이는 방법이 성능을 향상시켰다. 온톨로지의 적용을 자질의 선택 전과 후에 하는 것에 따라 성능의 차이를 보였고, 온톨로지를 자질선택 후에 적용할 때 더 좋은 성능을 얻을 수 있었다. 이것은 자질선택 전에 온톨로지가 적용되어 원본 데이터가 상위어로 바뀌더라도 그 바뀐 단어가 반드시 자질이 되리라는 보장이 없기 때문이라고 판단된다.

분류기의 경우, Naive Baysian의 경우는 자질을 모두 사용하는 것이 가장 성능이 좋다. 이것은 확률기반인 Naive Baysian의 특성 때문이다. SVM의 경우는 분류기 자체의 성능이 뛰어나 자질선택에 따른 큰 효과를 보기 어려웠다.

자질 선택기의 경우 모든 분류기에서 평균적으로 Odds ratio 방법이 가장 좋은 성능을 나타내었다. 그 다음이 Probability Ratio(PR) 방법이었다. 이들 둘의 결과를 평가하였을 때 자질선택 후의 온톨로지의 적용은 자질의 수가 커질수록 성능의 향상을 보였다. 선택하는 자질의 수가 클수록 감소되는 자질의 비율도 증가하고 성능 또한 향상되는 결과를 보였지만 자질의 수에 정비례 한 성능의 증가는 보이지 않았다. 700~1300개 사이의 자질을 선택하였을 때 가장 큰 폭으로 성능 향상을 보였다.

본 논문에서 사용한 온톤로지는 U-WIN의 전기 전자 분야에 한정된 부분만을 적용하여 실험을 하였지만 기존의 방법보다 향상된 성능을 보였다. 향상된 성능의 폭이 작은 이유는 온톨로지의 적용범위가 좁아서 온톨로지와 문서와의 교차확률이 작기 때문이다. 이후 더욱 잘 구축된 큰 온톨로지를 이용할 경우 좀 더 높은 성능을 얻을 수 있을 것이다. 또한 다양한 분야의 문서와 다양한 분야를 포함하는 온톨로지를 사용하여 일반적인 문서의 자질추출에서의 영향을 실험해야 할 것이다.

참고문헌

[1] 고영중, 서정연, 「문서관리를 위한 자동문서범주화에 대한 이론 및 기법」, 정보관리연구, 33권, 2호, pp. 16-32, (2002).

[2] 정재헌, 이상구, 「정보검색을 위한 효율적인 시소러스 구조에 관한 연구」, 한국정보과학회 봄 학술발표논문집 , 22권, 1호, pp. 949-952 , (1995).

[3] 최기선, 「온톨로지의 구축과 학습: 상하의 관계」, 정보과학회지, 24권, pp. 24-30, (2006).

[4] 최호섭, 임지희, 배영준, 최수일, 옥철영, 「온톨로지 구축 방법과 사례」, 정보과학회, 4권, pp. 31-44, (2006).

[5] Abasolo, J. M., Gómez, M., 「MELISA. AnOntology-based agent for information retrieval in medicine」, ECDL 2000 Workshop on the Semantic Web(SemWeb2000), pp. 73-82, (2000).

[6] Buckley, C., Salton, G., Allan, J., 「The Effect of Adding Relevance Information in a Relevance Feedback Environment」, Proceedings of th International Association for Computing Machinery SIGIR Conference, pp. 292-300, (1994).

[7] Berger, H. and Merkl, D., 「A comparison of text-categorization methods applied to n-gram frequency statistics」, Proceedings of the 17th Australian Conference on Artificial Intelligence, pp. 998-1003, (2004).

[8] Daelemans, W. and van den Bosch, A., ≪Memory-Based Language Processing≫, Cambridge University Press, (2005).

[9] Connolly D., van Harmelen F., Horrocks, I, McGuinness, D. L., Patel-Schneider P. F., and Stein L. A., ≪DAML+OIL (March 2001) Reference Description≫, W3C, (2001).

[10] Eyheramendy, S. and Madigan, D. A. N., 「Feature Selection Score for Text Categorization」. Proceedings of the International Workshop on Feature Selection for Data Mining, pp. 1-8, (2005).

[11] Finin, T., Labrou, Y., Mayfield, J., 「KQML as an agent communication language」, Software agents, MIT Press, Cambridge, MA, pp. 292-361, (1997).

[12] Forman, G., 「An extensive empirical study of feature selection metrics for text classification」, Journal of Machine Learning Research, vol. 3, pp. 1289-1305, (2003).

[13] Frakes, W. B., and Yates, R. B., ≪Information Retrieval Data Structures and Algorithms≫, Prentive-Hall, (1997).

[14] Genesereth, M. R., Fikes, R. E., ≪Knowledge Interchange Format:Version 3 Reference Manual≫, Techical report, Stanford University Logic Group, (1992).

[15] Gruber, T. R., 「A Translation Approach to Portable Ontology Specifications」, Knowledge Acquisition, vol. 5, no. 2, pp. 199-220, (1993).

[16] Guyon, I., Gunn, S., Nikravesh, M. and Zadeh L. A., ≪Feature Extraction-Foundations and Applications≫, Springer, (2006).

[17] Joachims. T., 「A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization」, Proceedings of the 14th International Conference on Machine Leaning, pp. 143-151, (1997).

[18] Joachims, T., 「Text categorization with support vector machines: learning with many relevant features」, Proceedings of the European Conference on Machine Learning, pp. 137-142, (1998).

[19] Joachims, T., ≪Learning to Classify Text Using Support Vector Machines≫, Kluwer, (2002).

[20] Lee, G., Le, J.-H., Rho, H., Park, Y.-T., Choi, J. and Seo, J., 「Interactive NLI agent for multiagent web search model」, Proceedings of International Workshop on Intelligent Agents on the Internet and Web, in 4th World Congress on Expert Systems, pp. 67-74., (1998).

[21] Manning, C. D., Raghavan, P. and Schütze, H., ≪Introduction to

Information Retrieval≫, Cambridge University Press (http://informationretrieval.org/), (2007).

[22] McCallum, A. and Nigam K., 「A Comparison of Event Models for Naive Bayes Text Classification」, Proceedings of the AAAI/ICML-98 Workshop on Learning for Text Categorization, Technical Report WS-98-05, pp. 41-48, (1998).

[23] Pedersen, T., 「Fishing for exactness」, Proceedings of the South-Central SAS Users Group Conference, pp. 188-200, (1996).

[24] Pedersen, T., Kayaalp, M. and Bruce, R., 「Significant lexical relationships」, Proceedings of the 13th National Conference on Artificial Intelligence, pp. 455-460, (1996).

[25] Salton, G. and Buckley, C., 「Term-weighting approaches in automatic text retrieval」, Information Processing Management, vol. 24, no. 5, pp. 513-523, (1988)

[26] Salton, G. and McGill, M. J., ≪Introduction to modern information retrieval≫. McGraw-Hill, (1983).

[27] Sebastiani, F., 「Machine learning in automated text categorization」, Association for Computing Machinery Computing Survey, vol. 34, no. 1, pp. 1-47, (2002).

[28] Voorhees, E. M., 「Overview of TREC 2005」, Proceedings of the 15th Text Retrieval Conference Proceedings, (2006).

[29] Willbur, J. W. and Sirotkin, K., 「The automatic identification of stop words」, Journal of Information Science archive, vol. 18, pp. 45-55, (1992).

[30] Yang, Y. and Pederson., J. O., 「A comparative study on feature selection in text categorization」, Proceedings of the 14th International Conference on Machine Leaning, pp. 412-420, (1997).