전문용어 및 정보추출에 기반한 문서분류시스템

이 경 순　　최 기 선

한국과학기술원 전산학과, 전문용어언어공학연구센터
{kslee, kschoi}@world.kaist.ac.kr

Text Categorization Based on Terminology and Information Extraction

Kyung-Soon Lee　　Key-Sun Choi
Department of Computer Science, KORTERM, KAIST

요　약

본 연구에서는 문서분류시스템에서 자질의 표현으로 전문분야사전을 이용한 분야정보와 개체정보추출을 통한 개체정보를 이용한다. 또한 지식정보를 보완하기 위해 통계적인 방법으로 범주 전문용어를 인식하여 자질로 표현하는 방법을 제안한다. 문서에 나타난 용어들이 어떤 특정 전문분야에 속하는 용어들이 많이 나타나는 경우 그 문서는 용어들이 속한 분야의 문서일 가능성이 높다. 또한, 정보추출을 통해 용어가 어떠한 개체를 나타내는지를 인식하여 문서를 표현함으로써 문서가 내포하는 의미를 보다 잘 반영할 수 있게 된다. 분야정보나 개체정보를 알 수 없는 용어에 대해서는 학습문서로부터 전문분야를 자동 인식함으로써 문서표현의 지식정보를 보완한다. 전문분야, 개체정보 및 범주전문용어에 기반해서 표현된 문서의 자질에 대해서 지지벡터기계 학습에 기반한 문서분류기를 이용하여 각 범주에 대해 이진분류를 하였다. 제안된 문서자질표현은 용어기반의 자질표현에 비해 좋은 성능을 보이고 있다.

1. 서 론

문서분류시스템은 문서의 내용을 파악하여 미리 정의된 두 개 이상의 범주로 분류하는 것으로, 문서를 어떻게 표현할 것인가를 다루는 자질추출(feature extraction) 부분과 추출된 자질을 기반으로 해서 어느 범주로 할당할 것인가를 결정하는 문서분류(document classification) 부분으로 구성된다.

문서의 표현을 위한 자질추출에 관한 연구에는 단일어를 자질로 이용하는 방법, 문법적인 어구를 자질로 이용하는 방법, 그리고 시소러스를 이용하여 용어의 의미를 자질로 이용하는 연구가 있다. 단일어를 자질로 이용하는 방법은 자질 추출이 간단하다는 장점이 있으나, 단어들간의 문법적인 관계나 의미적인 관계를 표현하기가 어렵다[Fagan, 1987]. 문법적인 어구에 기반한 자질 표현은 구문관계에 의해 생성된 단어들인 문법적인 어구를 자질로 표현하는 것으로, 단일어를 사용하는 것보다 언어적인 중의성이 적다는 장점을 가지지만, 너무 적은 출현빈도를 가진다는 문제점이 있다[Lewis, 1992, 장병규, 1997]. 시소러스를 이용한 의미기반 자질표현[강원석, 1999]은 용어가 가지는 의미를 획득해야 하는 문제가 있다.

문서에 범주를 할당하는 분류기법으로는 결정트리 학습에 기반한 분류기, 베이지언 확률에 기반한 분류기, 신경망에 의한 분류기, 최근접 이웃(k-Nearest Neighbor)에 기반한 분류기 등이 많이 이용되고 있다. 지지벡터기계(SVM: support vector machine)[Vapnick, 1995] 학습에 기반한 분류는 학습자료들에서 나타나는 패턴으로부터 지지벡터를 자동 생성하고, 이 지지벡터를 이용해서 분류를 하는 것으로, 우수한 성능을 보이고 있어서 최근 패턴 인식 연구분야에서 많이 이용되고 있다.

본 연구에서는 문서의 자질 표현 방법으로 전문분야사전을 이용한 분야정보, 개체정보추출을 통한 개체정보, 그리고 새로운 전문용어의 출현으로 인한 사전정보를 보완하기 위해 통계적인 방법을 이용하여 자동적으로 추출한 범주전문용어를 자질로 표현하는 방법을 제안한다.

2절에서는 문서의 표현으로 사용된 전문용어, 범주전문용어 인식 및 개체정보 추출에 대해 설명하고, 3절에서는 지지벡터기계 학습에 기반한 분류기를 이용해서 용어기반의 자질표현과의 실험을 비교, 분석하고, 4절에서는 결론을 맺는다.

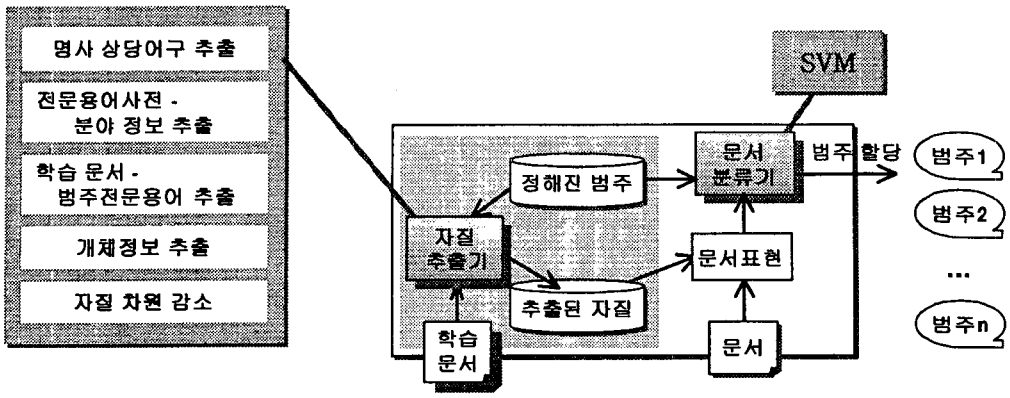

그림 1 : 전문분야정보 및 개체정보에 기반한 문서분류시스템

2. 전문분야정보 및 개체정보에 기반한 자질표현

전문분야정보 및 개체정보에 기반한 문서분류시스템의 전체적인 구조는 그림1과 같이 자질추출기와 문서분류기로 구성되어 있다.

자질추출기에서는 학습문서에서 명사상당어구를 추출하고, 그것이 전문분야에서 사용되는 용어인지를 검사를 한다. 전문분야에서 사용된 용어이면, 그 용어를 전문용어로 포함하고 있는 분야에 대한 정보를 추출한다. 용어가 사람이름, 지역명, 조직 등과 같은 것을 나타내는 것인지를 인식해서 개체정보를 추출한다. 분야정보나 개체정보를 알 수 없는 용어에 대해서는 자질과 범주와의 연관성을 통계적인 방법으로 계산해서 범주와 연관성이 높은 용어들을 범주전문용어로 자동 인식하여 범주정보를 추출한다. 이렇게 추출한 전문분야정보, 범주정보, 개체정보 그리고 용어 그 자체를 문서들을 대표할 수 있는 자질로 선택한다. 자질의 수가 너무 많아지면 문서의 범주를 학습하는데 어렵기 때문에 용어들 중에서 중요한 자질들을 선택하여 문서의 자질로 표현한다.

문서분류기에서는 선택된 자질과 학습문서를 이용해서 각 범주에 대해 지지벡터기계를 학습시켜, 문서가 그 범주에 속하는지 속하지 않는지를 결정하여 범주를 할당한다.

2.1 전문용어사전을 이용한 분야정보 추출

문서에 나타난 용어들이 어떤 특정 전문분야에 속하는 용어들이 많이 나타나는 경우 그 문서는 용어들이 속한 분야의 문서일 가능성이 높다. 그러므로, 전문분야사전에 포함되어 있는 용어들은 문서의 분류 특성을 결정하는데 커다란 영향을 가진다고 할 수 있다. 전문분야사전은 해당 분야에서 사용되는 용어들을 정의하고 있으므로, 전문분야사전을 이용해서 용어의 분야정보를 추출한다.

30여개의 전문분야사전을 이용하여 경제, 사회, 법률, 화학 분야 등 용어의 분야정보를 추출하여, 각 용어에 대해 나타날 수 있는 분야정보를 역파일(inverted file)구조로 색인한다. 문서에서 나타난 각 용어에 대해 색인구조를 참조하여 전문용어이면 분야정보를 문서의 자질로 표현한다. 다음은 용어가 자지는 분야정보를 역파일구조로 나타낸 것으로, '공비'라는 용어는 수학, 화학, 군사 분야에서 전문용어로 정의하고 있다. 문서에 공비라는 용어가 나타나면 3개의 분야정보를 표현하게된다.

공비: <수학><화학><군사>
발효: <경제><화학><금형><미생물학><유전학>
<고분자><해양><건축설비><식품과학>
전쟁: <군사>
체포: <법률><군사>

어떤 용어는 여러 분야에서 사용될 수 있는데, 이러한 용어들은 가중치를 낮게 하고, 분야가 뚜렷한 용어일수록 높은 가중치를 부여함으로써 문서분류에서의 영향을 조정할 필요가 있다. 전문용어의 분야변별정도를 반영하기 위해 역분야빈도수(*iTf* : inverse Terminology frequency)를 정의하였다. Tf는 어떠한 용어가 얼마나 많은 전문분야사전에서 정의되고 있는가를 나타낸다. 여러 개의 전문분야사전에 나타나는 용어는 해당 분야의 특성을 나타내는 정도가 약하다고 할 수 있다. 전체 전문분야사전의 수를 용어의 Tf로 나눈 값을 용어의 중요도로 반영함으로써 분야의 중요도를 조절하였다. 다음은 전문분야사전을 이용해서 계산한 Tf의 예이다.

아군	Tf:1
질서	Tf:1
세계	Tf:1
방법	Tf:5
국민소득	Tf:3

'방법'이나 '세계'와 같은 용어가 하나의 전문분야사전에서 나타나고 있지만, 일상용어에 가까운 용어라고 볼 수 있다. 이것은 전문용어사전의 제작 특성에 따라 일반용어에 가까운 용어를 포함하는 경우가 있고 그렇지 않은 경우도 있다. 따라서 하나의 전문용어사전에 나타나는 용어일지라도 전문성이 없는 용어일 수도 있다. 이럴 경우 앞에서 계산한 Tf값만으로 전문용어의 중요도를 측정하는 것은 정확하지 않을 수 있다. 이를 보완하기 위해 학습문서의 통계정보를 이용한다.

어떠한 용어가 학습문서에서 얼마나 많은 범주에 걸쳐서 나타나고 있는지를 측정하기 위해 역범주빈도수(***icf*** : inverse category frequency)[조광제, 1997]를 이용한다. cf는 용어가 얼마나 많은 범주에 걸쳐서 나타나고 있는지를 나타내는 값으로, cf의 값이 클수록 범주의 변별성이 떨어지는 용어라고 할 수 있다.

다음은 30여개의 전문분야사전과 49개의 범주를 기준으로 했을 때 각 용어가 갖는 Tf값과 cf값을 나타낸다.

아군	Tf:1	cf:2
질서	Tf:1	cf:21
세계	Tf:1	cf:45
방법	Tf:5	cf:46
국민소득	Tf:3	cf:1

'방법'과 '세계'와 같은 용어가 Tf값은 작은 값을 갖고 있어서 그 분야에서 중요하게 쓰이는 용어로 인식될 수 있으나, cf값을 통해서 일상용어에 가까운 용어라는 것을 알 수 있어 용어의 분야 중요도를 낮게 할수 있다. 이와 같이 전문분야사전과 학습문서를 이용해서 용어에 대한 분야정보와 분야에서의 중요도를 측정한다.

2.2 범주전문용어 추출

전문분야 인식에서 모든 전문분야에 대해 사전을 이용하기가 어려우므로 통계적인 방법을 이용해 범주와 연관성이 높은 용어를 범주 전문용어로 자동 인식하여 표현함으로써 지식정보를 보완한다.

하나의 범주에서 어떠한 용어가 하나의 문서에 집중적으로 나타나는 것보다 같은 범주에 속하는 여러 문서에 걸쳐서 자주 나타나는 경우 범주와의 연관성이 높을 것이라는 가정에서 범주내문서빈도수(***dfc***: document frequency in category)를 정의하였다. dfc는 어떠한 용어가 하나의 범주내에서 얼마나 많은 문서에서 나타나고 있는가를 나타낸다. 범주빈도수(cf)가 작고 범주내문서빈도수(dfc)가 클수록 그 용어는 그 범주를 대표할 수 있는 용어라고 할 수 있을 것이다.

본 연구에서는 범주전문용어를 인식하는 방법으로 범주와 관련된 자질을 추출하는 방법의 하나인, 카이제곱(χ^2)값을 이용하였다.

<철학>	<정치외교>	<법>	<군사>	<물리학>	<스포츠>
철학	정당	법률	탄알	물체	경기
맹자	국제연합	법원	전술	전류	선수
춘추	대통령	헌법	전투	전압	스포츠
유교	국회	형사	어뢰	질량	종목
순자	정치	재판	기관총	에너지	축구
성악설	국가	법	전략	빛	동작
성선설	북한	권리	화기	도체	상대편
경서	민주	부칙	핵무기	진공관	배구
해겔	선거	민사	소총	전하	공
...	...	...	...	...	...

표 1 : 추출된 범주전문용어의 예

χ^2는 두 사건의 독립성 여부를 판단하는 통계적인 방식[Devore, 1995]으로, 문서 범주화를 위한 자질의 중요성을 판단할 경우에는, '어떤 문서가 범주에 관련이 있다'라는 사건과 '어떤 자질이 그 문서에 나타난다'라는 두 사건의 독립성 여부를 판단하는데 이용된다. 두 사건이 독립적이라면 그 자질은 그 문서들을 범주화하는데 영향을 미치지 않는다고 판단한다. 그러므로, χ^2값이 큰 자질은 범주와 관련이 높은 용어라고 할 수 있다. 이 값은 잠정적으로 icf와 dfc가 반영되어서 계산된다고 할 수 있으므로, 각 범주에 대해서 χ^2값이 큰 용어들을 그 범주의 전문용어로 인식하여 용어에 대한 범주정보를 자질로 표현하였다.

표1은 실험집합에 정의된 범주의 일부인 <철학>, <정치·외교>, <법>, <군사>, <물리학>, <스포츠> 등의 각 범주에 대한 각 자질의 χ^2값을 계산해서 높은 값을 갖는 순서로 용어들을 나타낸 것이다.

문서에 나타난 용어에 대해서 자동 추출된 범주전문용어정보와 비교해서 범주정보를 문서의 표현으로 나타낸다. 즉, '전류'라는 용어가 문서에 나타났을 때, <물리학>범주정보를 문서의 표현으로 나타내게 된다.

전류: <물리학>
맹자: <철학>
헌법: <법>
강수량: <지구과학> <한국지리> <외국지리>

2.3 개체정보(named entity) 추출

정보추출의 연구[MUC]에서는 용어가 사람, 조직, 지위, 지명, 화폐, 날짜 등 어떠한 개체를 나타내는지를 인식하고, 동일지시어의 인식, 사건에 대한 템플릿을 구축하는 등 문서의 내

용을 이해하기 위해 지식을 자동 추출하려는 노력이 활발하다.

본 연구에서는 사전과 패턴에 기반하여 개체의 이름을 인식하여 자질로 한다. 인식의 대상이 되는 개체는 인명, 지명, 조직, 화폐, 시간, 길이, 무게, 각종 단위 등 20여 개체이다.

기본지식을 요구하는 인명, 지위, 조직 등의 개체는 수동 구축된 고유명사사전[남지순, 1999]을 이용하여 인식하였다. 다음은 고유명사 사전에 나타난 용어들의 예를 나타낸다. '//'는 문서에서 공백이 들어갈 수도 있음을 나타낸다.

<인명>	<지위>	<조직>
경덕/왕	각하	가지산/파
경명/왕	간사	간도//국민회
김//대건	간사장	감로/사
김//대문	간수장	감리//위원회
김//도수	간호원	건국//대학교
김//도연	감독관	건국//동맹
김//도원	감독원	강서시파
...	...	...

사전에 포함된 정보가 충분하다면 문서에 나타나는 모든 개체들을 인식할 수 있으나, 사전에 포함되어 있지 않은 정보는 인식을 할 수가 없다. 그리고, 어떠한 개체에는 개체와 관련되어서 나타나는 어휘정보에 따라서 이름을 부여할 수 있는 것들이 많다. 이러한 어휘정보를 패턴으로 기술하여 사전에서 찾을 수 없는 개체들을 인식한다. 다음은 개체인식을 위해 정의한 패턴의 일부이다. nq, nno등은 품사를 나타내고, n*은 명

```
사람(PERSON)
        nq+{씨/nbn}
        nq+{$POSITION$}
지위(POSITION)
        nq+{회장, 사장, 위원장, .. }
지명(LOCATION)
        n*+ {지역, 장소, 주변, 일대}
날짜(DATE)
        nno+{시,일,월,년,년대}/nbu
        nno+년/nbu nno+월/ncn+nno+일/nbu
화폐(CURRENCY)
        nnc+{원,달러}/nbu
        nnc+억/nnc+{원,달러}/nbu
조직(ORGANIZATION)
        n* + {위원회, 대회, 대학}
길이,높이,거리
        nnc+{미터, 인치, 피트, mm, cm, m, ft }
무게
        nnc+{톤, 그램, 킬로그램, g, kg, t,}
...
```

사류를 모두 포함할 수 있음을 나타낸다.

사람을 나타내는 정보에는 사람이름에 '씨'라는 어휘정보를 덧붙여서 사용되는 경우가 많은데, 이를 인식하기 위한 패턴으로 'n*+씨' 를 기술한다. 지명를 나타내는 정보에는 '지역, 장소, 주변, 일대' 등과 같은 어휘정보가 앞에 나타난 품사를 지명으로 인식할 수 있는 근거를 마련해준다고 할 수 있으므로, 이를 패턴으로 기술한다. 화폐를 나타내는 표현이나 날짜를 나타내는 표현은 품사와 몇몇 어휘정보를 통한 패턴으로 인식하는 것이 아주 효율적이다.

이와 같이 개체의 이름에 대한 정보추출을 통해 문서를 표현함으로써 문서가 내포하는 의미를 더 잘 반영할 수 있을 것이다. 용어에 대해서는 χ^2값에 따라 자질을 순위화해서 가장 큰 값을 갖는 용어들을 우선적으로 문서의 자질로 이용한다.

3. 실험 및 평가

3.1 실험집합

제안된 문서표현인 전문분야, 개체정보 및 범주전문용어에 기반해서 표현된 문서의 자질에 대해서 성능을 평가하기 위해 ETRI-KEMONG 실험 집합[강원석, 1999]을 이용하여 실험하였다. 실험집합에 포함된 문서의 수는 21,618개이고, 범주는 12개의 범주와 76개의 범주로 구성되어 있다.

범주12는 철학, 종교. 사회, 과학, 생물, 산업, 예술, 어학, 문학, 지리, 역사, 스포츠 · 레저의 12개의 범주로 구성되어 있고, 범주76은 범주12를 세부 분류한 것으로, 과학의 세부 분류는 과학 일반, 수학, 물리학, 화학, 천문·우주학 등이 있고, 산업의 세부분류는 산업 일반, 기술 · 가전, 건강과 의학 · 인체, 농업, 수산업, 임업 등이 있어 76개의 범주를 갖는다.

본 연구의 실험에서는 범주76에 대한 실험과 범주76 중에서 사람이 판별하기에 애매하지 않고 학습문서의 수가 어느 정도 이상에 해당하는 49개의 범주를 선택해서 실험을 하였다.

3.2 지지벡터기계(SVM)학습에 기반한 문서분류기

본 연구의 실험에서는 지지벡터기계 학습에 기반한 문서분류기를 이용하여 문서의 범주를 결정하였다.

지지벡터기계[Vapnick, 1995,1998]는 통계적인 학습 방법으로 Vapnik에 의해 고안된 것으로, 구조적 위험 최소화 원리를 채택하고 있어서 일반화 오류의 상한값을 최소화한다. 패턴의 분류에서 최적의 경계를 결정하기 위해 지지벡터(SV)를 이용하는데, 이 지지벡터는 학습자료가 나타내는 패턴으로부터 생성되며, 서로 다른 부류에 속하는 지지벡터들 사이의 거리에서 중간 지점을 최적의 결정의 경계로 한다. 지지벡터 학습의 기본 아이디어는 단순하면서도 실제 응용에서 우수한 성능을 보이고 있어, 분류나 회귀(regression) 문제에 적용되고

문서표현 \ 성능	49 범주			76 범주		
	정확율/재현율	F1-측정	향상률	정확율/재현율	F1-측정	향상률
용어(W)	50.16%/49.25%	49.70%		51.39%/39.49%	44.66%	
분야-용어(TW)	54.26%/49.64%	51.85%	+4.32%	51.81%/39.44%	44.79%	+0.28%
개체-용어(EW)	51.38%/48.82%	50.07%	+0.74%	52.30%/39.65%	45.10%	+0.99%
범주-용어(CW)	54.19%/52.97%	53.57%	+7.79%	53.11%/42.63%	47.30%	+5.90%
분야-개체-용어(TEW)	55.09%/49.41%	55.01%	+10.68%	54.01%/39.57%	45.68%	+2.27%
범주-개체-용어(CEW)	57.11%/53.06%	52.10%	+4.83%	54.92%/42.52%	47.93%	+7.32%
분야-범주-용어(TCW)	56.77%/53.33%	55.00%	+10.65%	54.18%/42.78%	47.81%	+7.05%
분야-범주-개체-용어(TCEW)	57.21%/53.99%	55.55%	+11.78%	55.55%/42.38%	48.08%	+7.65%

표 2: 자질 표현에 따른 문서분류시스템의 성능평가

있다[Hearst 1998].

문서는 하나 이상의 범주에 할당될 수 있으므로, 하나의 범주에 대해서 하나의 지지벡터기계를 학습시켜서 문서가 그 범주에 속하는지 속하지 않는지를 결정하였다. 자질의 값은 문서에서 나타났으면 1의 값을 갖도록 표현하고, 문서에 나타나지 않았으면 표현하지 않는다. 어떠한 문서에 대해 지지벡터기계 분류결과는 그 범주에 할당되는지 여부를 나타낸다.

3.3 실험결과 및 분석

실험은 용어기반, 전문분야정보-용어기반, 개체정보-용어기반, 범주정보-용어 기반, 전문분야정보-범주정보-용어 기반, 전문분야정보-개체정보-용어 기반, 범주정보-개체정보-용어 표현, 전문분야정보-범주정보-개체정보-용어 기반으로 자질을 표현했을 때 문서분류시스템의 성능의 차이를 비교하였다.

평가는 정확율과 재현율을 측정하였고, 정확율과 재현율을 하나의 값으로 표현해줄 수 있는 F_1-측정을 이용하였다. F_1값은 정확률과 재현율의 곡선에서 정확율과 재현율이 같아지는 지점에서의 값인 손익분기점(break-even point)과 유사하다는 특성을 가진다[Moulinier, 1996].

표2는 49개의 범주와 76개의 범주에 대해서 학습문서를 70%, 실험문서를 30%로 하고, 자질의 개수를 각 5000개와 10,00개로 하였을 때 각 문서자질표현에 따른 결과를 나타낸다. 그림3은 49개의 범주에 대한 실험에서 용어기반 자질표현과 제안된 전문분야정보-범주정보-개체정보-용어기반 문서표현에서의 각 범주에 대한 F1값을 나타내고 있다. 대부분의 범주에서 제안된 문서표현에 의한 분류가 높은 성능을 보이고 있다.

제안된 자질 표현이 용어기반의 표현에 비해 좋은 성능을 보이고 있으나, 기대했던 만큼 높지 않은 것은 일반적으로 용어가 가지고 있는 의미의 중의성 문제 때문인 것으로 보인다.

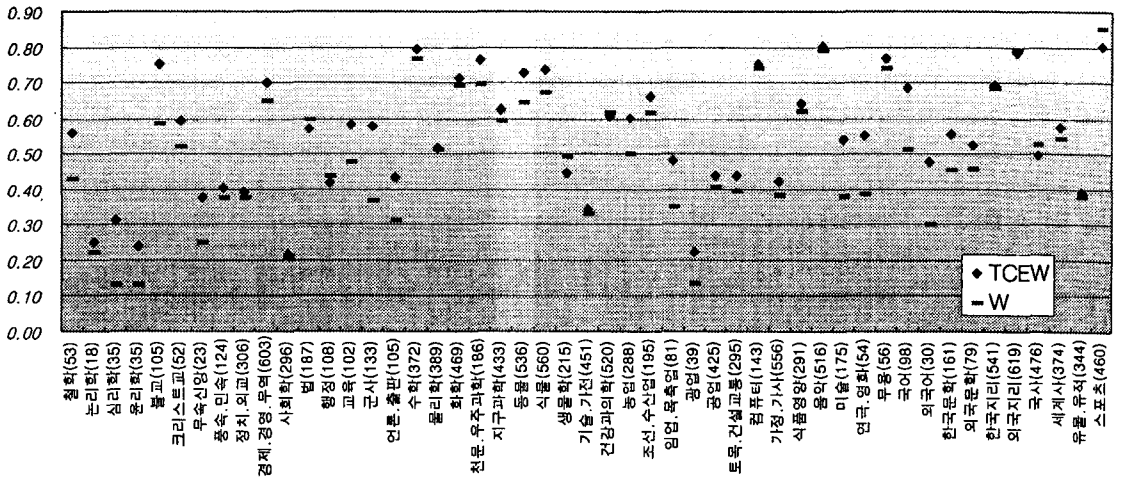

그림 3: 49개의 범주(학습문서수)에 대한 F_1값

용어의 중의성으로 인해 분야정보나 개체정보에 관련 없는 분야나 개체정보를 같이 표현하고 있어서 변별성의 효과를 떨어뜨리고 있다. 예로, '발효'라는 용어는 <경제>분야의 전문용어 사전에서는 'activation'의 의미로, <화학> <금형> <미생물학> <유전학> <고분자> <해양> <건축설비> <식품과학> 분야에서는 'fermentation'의 의미로 정의되어 있는데, 실제 문서에서는 경제분야를 뜻하는 것인데도 경제분야 이외의 다른 분야들에 대한 분야정보를 모두 포함하게 된다. 이러한 문제로 인하여 어떠한 문서에서는 거의 모든 분야에 대한 용어가 쓰인 것으로 나타나는 경우도 있다.

개체가 나타내는 정보에도 중의성을 가지고 있는 경우가 많다. 예로, '장비'라는 용어는 인명을 나타낼 수도 있고, 보통명사로 쓰일 수도 있는데, 사전에 의한 패턴매칭에서는 '장비'를 인명으로 인식하여 표현한다. '지사'라는 용어는 사람의 지위를 나타내는 의미로 쓰일 수도 있고, 회사의 지사를 나타낼 수도 있는데 이에 대한 해결 없이 그대로 사용하고 있기 때문에 문서의 표현에 관련없는 정보를 포함함으로써 변별력이 떨어지고 있다.

분야정보나 개체정보의 중의성을 문맥을 통해서 해결한다면 보다 우수한 성능을 보일 수 있을 것이다.

4. 결 론

본 연구에서는 문서분류시스템에서 자질의 표현으로 전문분야사전을 이용하여 분야정보와 정보추출에서 개체의 인식을 통한 개체정보를 추출하고, 지식정보를 보완하기 위해 학습문서로부터 범주전문용어를 자동으로 인식하여 자질로 표현하는 방법을 제안하였다. 지지벡터기계 학습에 기반한 문서분류기로 49개 범주에 대한 분류실험에서 제안된 자질표현방법이 용어기반의 자질표현 방법에 비해서 11.78%의 향상을 보이고 있다.

본 연구에서는 분야정보에 대한 애매성과 개체정보에 대한 애매성을 단순히 가중치의 계산에서 조정하면서 그대로 문서의 표현으로 하여 변별성이 약했는데, 문맥을 통해서 용어가 가지는 중의성 해결을 한다면 보다 높은 성능향상을 기대할 수 있을 것이다.

참고 문헌

강원석, 강현규. 1999. 시소러스도구를 이용한 실시간 개념기반 문서분류시스템, *한국 정보과학회 논문지*, 제26권 1호, pp. 167-177.

남지순. 1998. 한국어 백과명사 전자사전의 구축(I): 인명관련 백과명사의 연구. *CAIR-TR-98-74.*

조광제, 김준태. 1997. 역카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류, *한국 정보과학회 봄 학술발표논문집*(B), pp. 507-510.

장병규. 1997. *문서범주화에서 연어를 기반으로 한 문서표현.* 석사학위논문, 한국과학기술원 전산학과.

Devore, Jay L. 1995. *Probability and Statistics for Engineering and the Sciences.* Morgan Kaufmann Publishers, Inc., fourth edition.

Fagan, Joel L. 1987. *Experiments in Automatic Phrase Indexing For Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods.* Ph.D. thesis, Department of Computer Science, Cornell University.

Hearst, Marti A. 1998. Support Vector Machines, *IEEE Information Systems*, 13(4):18-28.

Lewis, David D. 1992. *Representation and Learning in Information Retrieval.* Ph.D. thesis, Department of Computer and Information Science, Graduate School of the University of Massachusetts.

Moulinier, Isabelle. 1996. A framework for comparing text categorization approaches. In *AAAI Spring Symposium: Machine Learning in Information Access*, March.

Proceedings of *the Seventh Message Understanding Conference(MUC-7).* http://www.muc.saic.com/

Vapnick, Vladimir N. 1995. *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

Vapnick, Vladimir N. 1998. *Statistical Learning Theory*, John Wiley & Sons. Inc.