한글문서 분류용으로 이용할 복합어로 구성된 분야연상어의 추출법

(An Extraction Algorithm of Compound Field-associated Terms for Korean Document Classifications)

이 상 곤 †

(Samuel Sangkon Lee)

요 약 분야연상어는 어휘자체가 분야정보를 가지므로 인간이 분야를 인지할 때와 유사하게 문서의 분야를 판단한다. 한국어의 경우 180분야로 분류된 약 15,000개의 문서뱅크를 수집하여 구축·실험한 결과 88,782개의 단일 분야연상어가 8,405개로 전체의 약 9%로 압축되며, 재현율 0.77 이상(평균 0.85), 정확률 0.90 이상(평균 0.94)의 높은 추출 정밀도를 얻었다. 구축한 분야연상어를 문서분류의 초기결정에 적용하여 인간에 의한 분야결정과 비교한 결과 약 90%이상의 정답률을 얻었다. 연구결과를 문서분류의 초기단계에 관한 기초연구로 이용하고, 다언어(multilingual) 간의 문서검색에 적용하여 다국어 정보검색에 대한 기초연구로 이용할 수 있다.

키워드 : 복합 분야연상어, 안정성랭크, 계승랭크, 단락검색, 정보추출, 문서분류, 정보검색

Abstract Field-associated Terms itself have field information. So, they determine field of document just like when human being perceives field. In case of Korean, we organized and experimented them by collecting approximately 15,999 document banks that are classified into 180 fields. We obtained high precision of extraction that 88,782 single field-associated terms are contracted into 8,405 ones thus recording compression rate as approximately 9% and recall as above 0.77 (average 0.85), precision as above 0.90 (average 0.94). By applying established field-associated terms to initial determination for document classification and comparing it with filed determination by human being, we got correct answers above approximately 90%. We can use results of research as fundamental research for initial stage and apply it document retrieval between multilingual environment thus utilizing it as fundamental research for multilingual information retrieval.

Key words : Compound Field-associated Term(FT), Stability Rank, Inheritance Rank, Passage Retrieval, Information Extraction, Document Classification, Information Retrieval

1. 서 론

인간은 문서전체를 읽지 아니하여도, 문서에서 대표적인 단어를 보는 것만으로 정치나 스포츠 등의 문서분야를 정확히 인지할 수 있다. 따라서, 문서단편 내의 소수의 단어정보를 이용하여 분야를 정확하게 결정하기 위한 분야연상어의 구축은 중요한 연구과제[14]이다.

인간은 자신의 상식지식으로 특정분야를 인지할 수 없는 경우에도 문서에서 처음으로 출현하는 몇 개의 단어들을 이용하여 연상되는 연상정보를 감각적으로 인식하고, 문서의 내용을 읽어감에 따라 문서에 해당하는 분야를 연상하거나 추측할 수 있다. 또한 문서의 이전내용에서 애매성이 발생하여도 문서의 뒤에서 출현하는 단어에 의해 이전의 문서내용에서 이해하지 못했던 애매성을 해소해 나갈 수 있다. 이와 같이 문서의 단락 내에 몇 개의 단어정보를 이용하여 문서가 포함되는 분야를 정확하게 결정할 수 있는 단어를 "분야연상어[14,15]"라 정의하고, 상식적인 분야연상어의 구축, 유사문서(문장) 검색, 문서요약 등의 기초연구를 수행한다.

제2장에서는 복합어로 구성된 분야연상어의 분석을 논의한다. 복합어를 구성하는 각 구성요소들(단일어들)

· 본 연구는 한국과학재단의 목적기초연구(과제번호 : R05-2003-000-10690-0) 지원으로 수행되었습니다. 재단의 연구지원에 깊은 감사를 드립니다.

† 종신회원 : 전주대학교 정보기술공학부 교수
samuel@jj.ac.kr

논문접수 : 2003년 12월 4일
심사완료 : 2005년 5월 13일

이 가지고 있는 의미계승에 의해 복합 분야연상어를 분석하는 방법을 제안한다. 분야 계승랭크[15]를 정의하고, 이전의 연구에서 제안한 안정성랭크와 조합하여 복합 분야연상어의 효율적 결정에 사용한다.

제3장에서는 인간이 미리 180분야로 분류한 약 15,000개의 파일을 이용한 실험에 의해 제안방법의 유효성을 평가한다. 실험을 위해 판정표에 의해 기준단계로 균등하게 분할하고, 이 기준에 의해 25단계의 균등분할보다 본 논문의 제안방법이 유효함을 입증한다. 추출된 분야연상어의 정밀도를 평가하기 위해 재현율과 정확률을 이용한다. 정밀도 평가결과를 분석하여, 본 연구의 목표인 높은 재현율이 달성됨을 증명한다.

제4장에서는 문서인식의 초기단계에서 분야결정의 실험을 수행하여 그 유효성을 평가한다. 판정문서의 정답데이타를 작성하여 분야연상어에 대한 분야결정이 어느 정도 정확하게 이루어지는지를 평가한다. 특히, 종단분야 이외에 상위의 유사한 중간분야까지를 모두 정답으로 간주하여 평가하면 100%에 가까운 정답결과를 얻을 수 있다.

본 연구결과를 이용하여 한국어, 일본어, 영어 등 언어별 분야연상어 컬렉션을 구축하면, 다국어로 작성된 문서를 각 분야로 분류할 수 있고 문서요약의 기초자료로 이용할 수 있을 것으로 기대한다. 각 나라의 문화(文化)에 맞는 고유한 정보에서 보편적 분야정보를 추출하면 다른 나라의 분야정보로도 쉽게 번역이 가능할 것이다.

2. 복합 분야연상어의 결정

본 장에서는 단일 분야연상어의 분야계승[15]을 기초로 복합 분야연상어의 분석을 논의한다. 기존의 연구들을 살펴보면, 복합어를 분석할 때 각 구성어의 통사적 구성에 대해서는 많은 논의가 있으나, 구성어의 의미적 계승에 대한 연구는 별로 활발하지 못하다. 일반적으로 복합어를 구성하는 단어 중 오른쪽 단어가 복합어의 문법적 주요어가 되며, 이 주요어가 단어 전체의 품사를 결정[1]한다고 알려져 있다.

따라서 이를 토대로 복합 분야연상어의 구성에 관계하는 요소를 분석하여 정리한다. 이 장에서는 단일 및 복합 분야연상어의 계승랭크를 정의하고 안정성랭크와 조합한 복합 분야연상어를 자동으로 결정하는 알고리즘을 제안한다. 더불어, 향후 연구과제인 분야규칙에 대해서도 본 장의 관련사항으로 논의한다.

2.1 분야계승에 의한 복합 분야연상어의 분석

2.1.1 의미에 의한 분야계승

복합명사의 통사적 구성에 대해서는 일반적으로 오른쪽 단어의 품사가 복합어의 문법적 주요어가 되며, 복합어 전체의 품사를 결정한다. 이 때, 왼쪽의 단어가 오른쪽 단어를 수식하는 경우가 대부분이기 때문에 오른쪽 단어의 의미가 복합어 전체에 계승되어 분야정보의 의미계승에 관계한다. 예를 들면, 그림 1은 '냄비'에 관한 의미계승을 나타낸다. "냄비"는 "요리도구"를 연상하며, 분야는 <취미-오락/요리-먹는 것>에 관한 분야연상어가 된다. 이 단어와 다른 단어가 결합하여 복합어 "압력+냄비"과 "칭기즈칸+냄비"인 경우 왼쪽의 단어 "압력"과 "칭기즈칸"은 단지 오른쪽의 단어 "냄비"의 의미를 한정하는 수식어이며, "냄비"의 의미를 계승하고 단일어 "냄비"와 동일한 분야 <요리-먹는 것>을 연상한다.

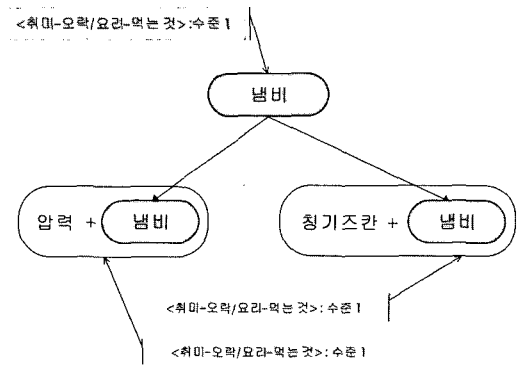

그림 1 "냄비"의 분야정보 계승 예

2.1.2 은유적 전의에 의한 분야계승

복합 분야연상어의 오른쪽 단어가 은유적 전의(轉意)에 의하여 왼쪽의 단어가 분류학 명사[1]가 되며, 연상되는 분야가 변화하는 경우가 있다. 그림 2와 같이 단일어 "전쟁"은 분야 <국제/지역분쟁>을 연상하고, 복합어 "걸프전쟁", "이라크전쟁", "6.25전쟁", "한국전쟁" 등은 전쟁 본래의 의미를 계승하지만, 다른 단어와의 복합어 "입시전쟁"에서 전쟁은 비유적으로 사용되고 있으며, 단

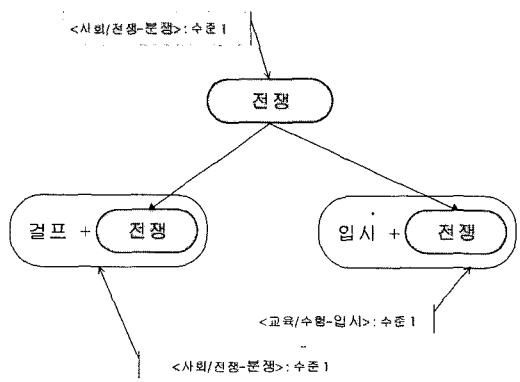

그림 2 "전쟁"의 분야정보 계승 예

어 자체의 의미인 "전쟁"을 의미하는 것은 아니다. 따라서 복합어에서 오른쪽의 "전쟁"은 통사적 주요어이긴 하지만, 어휘적 주요어는 아니다. 이런 경우는 왼쪽의 "입시"가 어휘적 주요부이며, 연상하는 분야는 <교육/수험-입시>가 된다.

이 경우 단일 분야연상어 "전쟁"은 두 가지의 분야를 연상할 수 있으며, 다른 단어와 결합하면 전혀 다른 분야를 연상할 수 있다. 이와 같이 단어의 은유적 전의(혹은 의미적 전의)는 일상생활에서 끊임없이 생성된다. 따라서 각 구성요소(각각의 단일어들)의 의미정보만으로 복합 분야연상어를 결정하는 것은 바람직하지 못하다.

2.1.3 복합어의 왼쪽단어에 의한 분야계승

다음의 그림 3은 오른쪽의 단어가 은유적 전의가 발생하지 않고 왼쪽의 단어가 어휘적 주요부가 되는 경우의 예이다. 예를 들면, 일본어의 경우, 복합어 "장고냄비"는 "냄비"의 의미계승에 의해 <요리/먹는 것>의 분야를 연상하지만, 인간의 두뇌는 분야 <쓰모>를 연상[1] 한다. 어휘적 주요부가 되는 왼쪽의 구성어 "장고"의 분야는 <요리-먹는 것>과 <쓰모>를 모두 계승하고 있다. 이 단어가 "냄비"라는 단어와 복합 분야연상어로 구성되어도 동일한 분야를 동일한 수준으로 계승한다. 따라서 "장고냄비"와 같은 복합어는 쓸모 없이 길게 형성된 분야 과잉 연상어[2]로 간주하고, 본 논문에서는 논의하지 않는다.

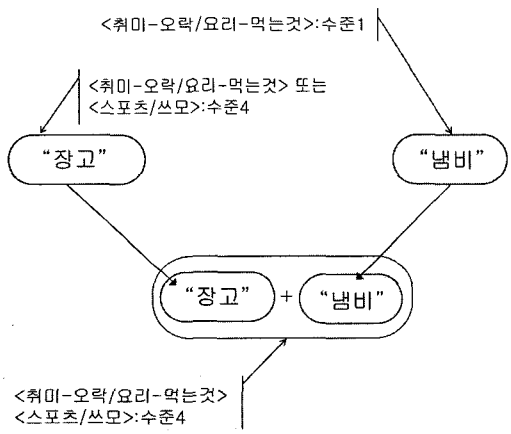

그림 3 "장고냄비"의 분야정보 계승 예

2.1.4 계승랭크의 필요성

복합 분야연상어의 각 구성어들(왼쪽 혹은 오른쪽 구별 없이)은 각각 독립된 분야를 계승하며, 때로는 유사한 분야를 연상하는 성질을 갖는다. 따라서 다음과 같이 계승랭크를 정의한다.

정의. 계승랭크(Inheritance Rank)

복합 분야연상어(w)의 연상분야를 <F>라 하고, 그 복합어의 구성어(x)가 연상하는 분야를 <F'>이라 하자(그림 4 참조). 두 분야 <F>와 <F'>이 다음의 세 가지 경우 중 하나에 해당하면 '유사분야'라 정의한다.

1) <F>와 <F'>이 일치하는 경우,

2) <F'>이 <F>의 상위분야인 경우, 혹은

3) <F>와 <F'>이 종단분야[15]에서 동일한 부모분야를 갖는 경우.

반대로, <F'>과 <F>가 전혀 다른 분야를 연상하면 '다른 분야'라 한다. 복합 분야연상어의 후보 w(w=xy라 하자)와 구성어 x가 한정하는 분야가 유사분야를 갖는 경우, 계승의 정도가 가장 높은 랭크열 'A'로 정의한다. 만약, x가 어떠한 유사분야도 갖지 않고 다른 분야를 갖는 경우에 계승랭크는 'C(가장 낮은 랭크)'로 정의한다. 구성어 x가 수준 5[3]의 비연상어(참고문헌 [15] 참고)인 경우는 중간정도를 나타내는 랭크 'B'로 정의한다.

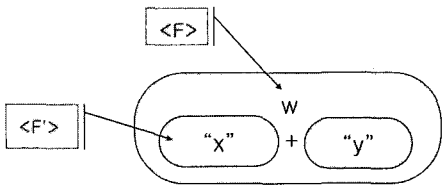

그림 4 유사분야와 다른분야

예를 들면, <야구>를 연상하는 복합어 w="김응용감독"에 대하여 x="김응용"과 y="감독"은 연상분야가 모두 <야구>로서 유사분야를 갖기 때문에, w에 대한 x의 계승랭크 값은 A가 된다. 다른 형태의 예로서 분야연상어 후보 w="성균관대감독"을 생각해 보면, x="성균관대"는 <교육>의 하위분야에 해당하는 분야연상어이므로 계승랭크는 C가 된다.

1) "장고냄비"는 쓰모선수들이 즐겨 먹는 음식이름이다.

2) 기존의 연구는 "분야 중복 연상어"라 정의하였으나, 본 연구에서는 이를 "분야 과잉 연상어"라 한다. 영문용어는 Field-Associated Redundancy Terms라 할 수 있다. 여기서 리던던시는 중복된 생각을 반복하는 단어의 부분요소라 설명할 수 있다.

3) 분야연상어 수준의 정의
(수준 1) 완전 분야연상어 : w는 유일한 종단분야만을 연상한다.
(수준 2) 준완전 분야연상어 : w는 같은 부모분야를 갖는 종단분야 중에서 한정된 복수 개의 종단분야만을 연상한다.
(수준 3) 중간 분야연상어 : w는 완전 분야연상어, 준완전 분야연상어가 아니고, 하나의 중간분야를 연상한다.
(수준 4) 다분야연상어 : w는 완전 분야연상어, 준완전 분야연상어, 중간분야연상어가 아니고, 다수의 중간분야와 다수의 종단분야를 연상한다.
(수준 5) 비연상어 : w는 위의 수준 1~4 이외이고, 어떠한 특정분야도 연상하지 않는다.
자세한 사항은 참고문헌을 참조하기 바란다.

표 1 수준 1에 해당하는 <야구>에 대한 복합 분야연상어 예

복합어 후보	빈도	랭크열	판정기준	판정결과	판정수준
고 교 + 야 구	149	-	-	×	-
김응용 + 감독	127	AcAa	8	○	1
김재박 + 감독	85	AcAa	8	○	1
일반인 + 야구	66	-	-	×	-
야 구 + 연 맹	55	-	-	×	-
법정대 + 진학	53	CbAa	103	●	5
해태+타이거스	23	AbAc	11	○	1
전 력 + 진 단	21	AaCa	27	●	5
친 선 + 게 임	20	BaAa	15	○ 과잉	1
롯 데 + 타 선	19	-	-	× 삭제	-
봉 황 기 + 전	16	CaAa	27	●	5
퍼펙트 + 게임	4	BaAa	15	● 탈락	5
현 역 + 은 퇴	4	AaAa	3	○ 과잉	1
투 수 + 교 체	4	-	-	×	-
김용수 + 투수	3	-	-	×	-

2.2 우선순위의 결정

계승랭크와 안정성랭크[4]를 이용하여 복합 분야연상어가 지시하는 정확한 분야를 결정한다. 이 두 가지 랭크열은 복합 분야연상어를 구성하는 각 구성요소들의 분야정보를 결정하는 '우선순위'로 사용한다.

(1) 랭크열

복합 분야연상어 후보 w의 구성어가 x와 y라 하면, 계승랭크와 안정성랭크를 연결시켜 '랭크열'을 생성한다. 예를 들어, 위의 표 1에서 표시한 바와 같이 <야구>에 대한 복합 분야연상어(김응용감독)에서 x="김응용"의 계승랭크는 'A'이고, 안정성랭크는 'c'이다. 다른 구성어 y="감독"의 계승랭크는 'A'이고, 안정성랭크는 'a'이다.

이를 결합하면, 김응용감독(w)의 랭크열은 "김응용"의 계승과 안정성랭크는 'Ac', "감독"은 'Aa'이다. 따라서 이를 조합하여 얻은 랭크열은 'AcAa'가 된다(표 1의 랭크열 참조).

(2) 25단계의 판정기준

랭크열을 이용하여 분야연상어 후보(w)가 어느 정도 분야를 연상하는가를 나타내는 판정기준을 마련한다(표 2의 판정표 참조). 먼저 계승랭크(영문 대문자 사용)의 조합에 의하여 AA에서 CC까지 다음과 같이 다섯 단계 ① AA, ② AB (혹은 BA), ③ AC (혹은 BB 또는 CA), ④ BC (혹은 CB), ⑤ CC 등으로 우선순위를 정의한다. 단, ③의 AC는 A의 우성과 C의 열성이 서로 맞아 상쇄되기 때문에 우열이 없는 BB와 동일한 단계로 취급한다. 각 계승랭크의 다섯 단계에 대하여 다시 안정성랭크(영문 소문자 사용)를 aa에서 cc까지 다섯 단계(aa, ab(혹은 ba), ac(혹은 bb, 또는 ca), bc(또는 cb), cc 등)로 세분화하여 모두 25단계(5 가지의 계승랭크×5 가지의 안정성랭크)의 '판정표(Decision(L, <F>))'를 결정한다.

표 2 판정표(Decision(L, <F>))의 예

단계	랭 크 열		기준 빈도
	계승	안정성	
1	AA	aa	3
2		ab	6
3		ac (혹은 bb, ca)	8
4		bc	11
5		cc	13
6	AB	aa	15
7		ab	18
8		ac (혹은 bb, ca)	20
9		bc	23
10		cc	25
11	AC (혹은 BB, CA)	aa	27
12		ab	30
13		ac (혹은 bb, ca)	32
14		bc	40
15		cc	48
16	BC	aa	56
17		ab	64
18		ac (혹은 bb, ca)	72
19		bc	80
20		cc	88
21	CC	aa	95
22		ab	103
23		ac (혹은 bb, ca)	111
24		bc	119
25		cc	127

4) [정의] 안정성랭크 : 안정성랭크는 랭크의 순위가 높은 순으로 보통명사를 a로, 인명이외의 고유명사를 b로, 인명에 해당하는 고유명사를 c로 할당한다.

(3) 기준빈도

다음의 표 2에서와 같이 수준 L에 대하여 분야 <F>를 연상하는 분야연상어 후보 w의 집합을 W_SET(L, <F>)라 정의하고, 그 집합 중에서 후보어의 최대빈도, 평균빈도, 최소빈도를 구한다. 최소빈도에서 평균빈도까지, 평균빈도에서 최대빈도까지를 각각 12개로 등분하여 평균빈도를 더한 25단계의 기준빈도를 대응시킨 판정표 Decision(L, <F>)를 표 2와 같이 정의한다. 이 판정표는 우선순위가 높은 후보일수록 제거되는 빈도를 낮게 하여 추출에서 제외될 가능성을 방지하도록 하는데 이용한다. 우선순위가 낮은 후보는 제거되는 빈도를 높게 설정하여 과잉추출 되는 것을 방지한다.

2.3 복합 분야연상어의 결정 알고리즘

이전에 서술한 분야연상어 결정 알고리즘[15]에 의해 복합 분야연상어는 분야연상어 후보로 결정되고, 아래의 결정 알고리즘에 의해 최종적으로 복합 분야연상어로 선택된다. 무조건 계승랭크가 높은 후보를 선택하면 역으로 분야 과잉 연상어를 선택하는 다음의 두 가지 모순이 발생한다.

- 복합 분야연상어 후보(w)의 연상분야 <F>와 구성어 x의 연상분야 <F'>이 다른 분야인 경우,
- 복합어 w의 모든 연상분야 <F>와 모든 구성어 x의 연상분야 <F'>이 유사한 분야이고, w의 수준이 x의 수준보다 높은 경우 등이다.

따라서 다음의 복합 분야연상어 결정 알고리즘을 이용하여 위의 두 가지 모순점을 해결한다. w가 분야연상어 후보이면, 각 구성어에 의해 연상분야의 추적이 가능하지만, 분야 과잉 연상어가 될 가능성이 있다. 따라서 복합 분야연상어 결정 알고리즘 (순서 B1)과 같이 계승랭크에서 발생하는 복합 분야 과잉 연상어 후보를 먼저 제거한다. (순서 B2)에서는 계승랭크를 이용하여 판정표에 의한 복합 분야연상어를 우선적으로 결정한다. 분야연상어 w의 집합 W_SET(L, <F>)의 역 표현인 F_SET(w)는 분야연상어 w의 (L, <F>)를 요소로 하는 집합이다.

● 복합 분야연상어 결정 : 알고리즘[5]

입력 : 복합 분야연상어 후보 w와 W_SET(L, <F>),

출력 : 복합 분야연상어 w의 연상분야와 수준

(순서 B1) : 분야 과잉 연상어의 제거

분야연상어 w=xy에 대하여 다음을 실행한다.

(순서 B1-1) : w의 수준(L)이 1인 경우

F_SET(w)와 F_SET(x)가 같은 요소 (L, <F>)를 갖고 동시에 x의 안정성랭크가 a일 때[6], 혹은 F_SET(y)가 <F>와 다른 분야 <F'>이 되는 요소 (1, <F'>)를 갖지 않으면, w를 W_SET(L, <F>)와 F_SET(w)에서 제거한다(y의 경우도 동일). 이것은 F_SET(w)에서 요소(L, <F>)가 제거되는 것을 의미한다. 수준 1의 분야연상어는 유일한 종단분야를 연상하는 중요한 분야연상어이므로 안정성랭크 a인 조건으로 한정하지만, 다음의 순서 (B1-2)에서는 분야연상어의 수준이 2~4이므로 조건 a를 붙이지 않는다.

(순서 B1-2) : w의 수준이 2~4의 경우

F_SET(w)의 모든 분야 <F>가 F_SET(x)와 F_SET(y)의 모든 분야 <F'>과 유사한 분야이고, w의 수준이 x와 y의 수준을 넘지 않으면 w를 W_SET(L, <F>)에서 제거한다.

(순서 B2) : 판정표에 의한 압축

이상의 처리로 얻어진 W_SET(L, <F>)에 대한 판정표 Decision(L, <F>)를 결정(표 2 참고)하고, w의 계승랭크열과 안정성랭크열에 대응되는 기준빈도 w의 정규화 빈도 Normalization(w, <F>)(앞의 주석 8 참고)보다 적으면, w를 W_SET(L, <F>)에서 제거한다.

(순서 B3) : 수준의 최종결정

이상과 같이 연상분야가 제거된 후보어 w는 모두 수준 5로 변경하여 비연상어로 처리하고, 연상하는 분야수가 감소한 후보어 w는 참고문헌 [15]에서 정의한 수준 1~5에 따라 각 수준을 변경한다.

표 1에 수준 1의 복합 분야연상어 후보를 나타내고, 표 3에는 수준 2~4의 후보어를 나타내었다. 이 표의 각 셀에 표시된 분야 수를 세 개 이내로 제한하고, 수준 3에서는 하위의 종단분야를 열거하였다. 알고리즘의 (순서 B1)에서 제거된 후보어는 표 3의 판정 난에 ●는 탈락을, ○는 과잉을, ×는 삭제를 의미한다. 다음은 위에서 제시한 알고리즘을 쉽게 이해하고, 실제 이용되는 것을 보이기 위해 실제 텍스트를 대상으로 실행한 예제를 가지고 설명한다.

먼저, 수준 1의 (순서 B1)에 대해 설명하면, (순서 B1)의 (1)의 예로서 표 1에서 수준 1의 후보 w="김용수+투수"를 생각해 보자. F_SET("투수")는 요소(1, <야구>)를 포함하고, 안정성랭크는 a이며, "김용수"는 다

5) 참고문헌 [15]에서 제시한 알고리즘 A와 구분하기 위해 다음의 알고리즘 순서를 'B'라 한다.

6) 수준 1의 완전 분야연상어는 유일의 종단분야를 지시하는 중요한 분야연상어이므로 안정성랭크가 'a'로 부착하였으나, 다음의 (순서 B2)의 대상이 되는 수준 2~4는 이 조건을 붙이지 않는다.

표 3 수준 2~4에 해당하는 <야구>에 대한 복합 분야연상어의 예

수준 2의 후보	분야 1	랭크열 1	판정수준 1	분야 2	랭크열 2	판정수준 2	분야 3	랭크열 3	판정수준 3
리그 + 기록	<야구>	AaAa	○	<농 구>	AaAa	○	<축 구>	AaAa	○
선발 + 박찬호	<야구>	AaAc	×	<축 구>	AaAc	×			
TV + 방영권	<야구>	CaCa	●	<럭 비>	CaCa	●			
해태+타이거즈	<야구>	BbBb	●	<럭 비>	BbBb	●			
연계 + 플레이	<야구>	BaAa	○	<농 구>	BaAa	○	<축 구>	BaAa	○
수준 3의 후보	**분야 1**	**랭크열 1**	**판정수준 1**	**분야 2**	**랭크열 2**	**판정수준 2**	**분야 3**	**랭크열 3**	**판정수준 3**
일본 + 선수	<축구>	BbAa	○	<농 구>	BbAa	○	<럭 비>	BbBb	○
우승 + 전선	<야구>	AaCb	○	<축 구>	AaCb	○	<쓰 모>	AaCb	○
한국 + 선수	<육상>	BbAa	×	<테 니 스>	BbAa	×	<골 프>	BbAa	×
역전 + 우승	<야구>	AaAa	○	<쓰 모>	AaAa	○	<골 프>	AaAa	○
선수권 + 선발	<럭비>	AaAa	×	<농 구>	AaAa	×	<축 구>	AaAa	×
수준 4의 후보	**분야 1**	**랭크열 1**	**판정수준 1**	**분야 2**	**랭크열 2**	**판정수준 2**	**분야 3**	**랭크열 3**	**판정수준 3**
유효 + 투표수	<야구>	BaBa	●	<정치/선거>	BaAa	○	<정치/정당>	BaAa	○
선제 + 공격	<야구>	AaAa	○	<테 니 스>	AaAa	○	<국제/중동>	BaAa	○
패자 + 부활전	<야구>	AaBa	○	<국제/중동>	AaBa	○	<정치/평화>	CaBa	○
타이틀 + 방위	<야구>	AaCa	●	<복 싱>	AaAa	○	<레저/취미>	AaAa	○
김응룡 + 감독	<야구>	AcAa	●	<배 구>	BcAa	●	<레저/영화>	CaBa	○

른 분야의 수준 1의 분야연상어가 아니므로 후보어 w는 제거된다. 후보어 w="야구+입문"은 F_SET(w)=F_SET("야구")={(1, <야구>)}이고, F_SET("입문")이 다른 분야 <교육/수험-입시>의 수준 1의 분야연상어로 가능하므로 제거되지 않고, 최종결정이 다음 순서로 미루어진다. 이 현상은 단일 분야연상어와 복합분야연상어를 동일한 학습데이타를 사용하였기 때문이다. 단일어 후보의 선정단계에서 "입문"은 <야구>와 <교육/수험-입시>를 연상하는 수준 4이지만, 사람의 판단으로 분야 <야구>가 제거된다. 이와 같이 다른 분야를 포함하는 후보어가 (순서 B2)에 의해 채택되는 경우는 극히 드물다.

(순서 B1-2)의 수준 2~4에 해당하는 분야연상어에서 표 3과 같이 수준 2의 w="선발+박찬호"를 생각해 보면, F_SET(w)=F_SET("선발")=F_SET("박찬호")이므로 w는 제거된다. 동일하게 수준 3의 "봉황기+처녀출전"도 소거된다.

w="한국+선수"는 F_SET(w)=F_SET("선수")이고, "한국"이 수준 5이기 때문에 w는 재거된다. 표 3에는 없지만, F_SET(w)={(2, <야구>), (2, <축구>)}가 되는 분야연상어 후보 w="선동렬+투수"를 판정할 때 각 구성어 "선동렬"과 "투수"의 모든 연상분야와 w의 분야는 유사분야이다. 그러나 복합어 w의 수준 2는 구성어의 수준 1과 2를 초과하지 않으므로 제거된다.

(순서 B2)는 표 3에서 ×표가 붙지 않은 후보어 22개 (최소빈도는 3, 평균빈도는 32, 최대빈도는 127)가 (순서 B2)의 W_SET(1, <야구>)의 요소가 되며, 표 2의 판정표 Decision(1, <야구>)의 기준빈도가 얻어진다. 예를 들면, "법정대+진학"의 빈도는 53으로 비교적 높지만 구성어는 <교육>의 분야연상어이고, 얻어진 랭크열은 Ccba의 기준빈도 103이 되므로 제거한다. 동일하게 표 3에서 W_SET(L, <F>)에 대응하는 판정표를 구성하여 제거분야를 결정한다. (순서 B2)에서 제거된 표 2의 후보어는 표 3의 판정난에 ●로 표시하였다.

앞의 알고리즘 (순서 B3)을 설명하면, 이상의 제거로 분야연상어가 삭제된 후보어의 수준을 모두 5로 변경한다. 수준 2~4의 후보어에서 일부 분야가 제거된 경우도 마찬가지로 수준변경을 한다. 사람이 수정한 단일 분야연상어의 정보를 이용하여 (순서 B2)의 분야갱신은 인간의 판단과 가까운 분야정보를 복합 분야연상어에 부여함을 의미하기 때문에 (순서 B3)의 수준상승은 인간의 판단에 기반 한 연상작용이지만, 실제의 분야결정과 비교하는 면에서 그 의의가 크다고 말할 수 있다.

표 2와 표 3의 판정수준 난에는 (순서 B3)에 의해 최종 결정된 수준을 표시하였다. 표 1에서는 최종적으로 12개의 복합 분야연상어로 압축되지만, "김재박+감독", "퍼펙트+게임"은 탈락되었다. "친선+게임"과 "현역+은퇴"는 <스포츠>에 대한 분야연상어이지만, 지나치게 많이 추출된 것이다. 이 분야연상어는 <야구>와 유사한 분야이기 때문에 다른 분야를 연상하는 잘못된 분야연상어는 아니다. 표 3에서는 "한국+대표"(<스포츠>의 하위분야인 복수의 종단분야를 연상한다고 가정)는 수준 3에서 수준 2로 변경되고, 동일하게 수준 4의 "유효+투표수"도 <정치>에 관한 복수개의 하위분야를 연상하는 수준 2로 변경된다.

2.4 분야연상어의 규칙

단어 x에 대한 개념을 *Concept(x)*=¥로 표시한다. 예를 들면, *Concept*("전주")=[지명]이 된다. 또한 $w=xy$가 되는 y에 대한 접두사 x의 집합을 *Prefix(y)*, x에 대한 접미사의 집합을 *Suffix(x)*로 한다. $w=xy$ 일 때,

$$\text{prefix}(y)=\{x_1, x_2, \ldots x_n\},$$
$$\text{suffix}(x)=\{y_1, y_2, \ldots y_n\}$$ 이다.

이 때, 분야규칙을 결정하는 알고리즘을 다음에 표시하였다.

● 분야연상어 규칙 결정 : 알고리즘
입력 : 분야연상어 w=xy,
출력 : 분야연상어의 규칙

(순서 C1) : 공통개념의 추출
동일분야를 지칭하는 복합 분야연상어 w=xy에 대한 집합 Prefix(y)를 결정하여, 그 요소 x에 대한 Concept(x)=¥를 추출한다.

(순서 C2) : 규칙후보의 결정
[¥] + y인 규칙후보를 결정한다.

(순서 C3) : 바른 규칙의 결정
Concept(z)=¥가 되는 요소(≠x)에 대한 zy를 자동 생성하여 w=xy와 동일한 수준의 분야연상어가 되는가를 확인하여 올바른 규칙을 결정한다.

(순서 C4) : 접미사의 처리
(순서 C1)에서 Suffix(x)에 대하여 (순서 C2), (순서 C3), (순서 C4)를 동일하게 실행한다.

예를 들어, 동일분야 <과학·기술/생물학·바이오>에의 수준 1의 분야연상어 w="신경조직", "뇌조직"과 y="조직"에 대하여 (순서 C1)과 (순서 C2)에서는 *Prefix*(y)={"신경", "뇌"}가 되는 요소 x에서 Concept=¥=[늑막·근육·신경·내장]인 개념을 검출하여, 규칙후보 ¥ + "조직"을 얻었다. (순서 C3)에서 같은 개념인 단어 z로부터 합성되는 복합어 "조직", "상피조직"은 <생물학·바이오>에의 수준 1의 분야연상어가 되기 때문에 [늑막·근육·신경·내장] + "조직"은 <생물학·바이오>에의 분야연상어 규칙이다. 본 논문에서는 분야연상어의 규칙에 대한 충분한 실험은 수행하지 않으며 상세한 평가도 생략한다. 이것은 미래에 연구를 계속할 과제이므로 본 논문에서 더 이상 논하지 않는다.

3. 분야연상어 컬렉션, 실험 및 평가

본 장에서는 사람의 손으로 수정구축한 단일 분야연상어를 이용하여 알고리즘에 의한 복합 분야연상어의 정밀도를 평가한다. 제안된 방법은 단일 분야연상어에 대해 사람이 직접 수정하였기 때문에 판정자에 따라 주관적인 생각의 개입이 있을 수 있지만, 분야연상어의 수정과 올바른 분야연상어의 판정을 동일인이 수행하도록 하여 구축된 데이타의 혼란을 최소화하였다. 평가실험은 판정표의 기준단계를 균등하게 나누어 고찰하였다. 추출 정밀도의 평가는 재현율과 정확률을 이용하여 고찰한다. 이 장에서 설명하는 구체적인 수치는 샘플링(sampling)에 의한 비율의 계산에서 얻어진 결과임을 먼저 밝힌다.

3.1 실험 데이타

부록 A의 분야체계는 사람이 참고문헌 [15]를 참고하여 인터넷 상에서 약 1년 간 수집하였다. 또한 실험에 사용된 문서는 주로 조선일보 신문기사에서 수집하였다. 또한 수집된 데이타가 매우 적은 분야는 종단분야의 최저 데이타 양을 수십 킬로바이트로 제한하였다. 주석 7에서 설명한 것과 참고문헌 [15]에서 사용한 알고리즘의 순서 (A2)에서의 식 (1)을 간단히 하기 위해

$$\frac{Normalization(w, \langle F\rangle)}{m}$$

인 평균치를 이용하였으나, 이번 실험에서는

$$\frac{Normalization(w, \langle F\rangle)}{(\beta m)}$$

와 같이 기준치 β를 도입하였다. 여기서, β > 1이면 누적 가산되는 자식분야수가 많아지며, 수준 2의 후보가 증가하면 수준 3의 후보는 감소한다. 분야연상어 후보의 결정 알고리즘에서 사용된 기준치 α와 β를 결정하기 위해 여러 차례 실험한 결과 α와 β의 기준치를 각각 0.92와 0.91로 결정하였다. 정규화 되지 않은 빈도 *Frequency*(w, <F>)[7]가 1인 단어 w는 분야연상어 추출의 실험대상에 포함시키지 않았다.

분야전체에서 추출된 단일 분야연상어 83,894개에 대하여 본 논문에서 제시한 분야연상어 후보의 결정 알고리즘에 의한 분야연상어 후보 수는 49,798개(수준 1, 2, 3, 4의 순으로 각각 28,367개; 4,400개; 320개; 16,711개)이며, 사람이 수정한 후에 결정된 단일 분야연상어의 수는 15,328개(수준의 순서로 각각 7,354개; 1,846개; 191개; 5,937개)가 되었다. 이 작업은 표 1과 표 3에서 제시한 방법에 의해 효율적이고 빠르게 진행되어 한 사람이 3주 동안에 완성하였다. 이상의 실험결과는 <스포츠>를 중심으로 한 자식분야의 실험으로 비례계산에 의해 결정된 결과이다. 현재에도 분야연상어의 구축과 확장작업이 계속되고 있으나, 거의 이 값으로 결정될 것으로 예상된다. 복합 분야연상어의 선정과 안정성랭크의 판단

7) 인간이 수집한 분야연상어가 각각의 종단분야에 균일하게 수집되었다고 보기 어렵기 때문에 종단분야 〈T〉에 출현하는 모든 단어의 합계빈도 Total_Frequency(w)를 계산하고, 종단분야 〈T〉에 출현하는 단어 w의 빈도를 Frequency(w, 〈T〉)라 하면, 다음의 식과 같이 정규화된 빈도 Normalization(w, 〈T〉)를 정의한다.

$$Normalization(w, \langle T\rangle) = \left\{\frac{Frequency(w, \langle T\rangle)}{Total\_Frequency(w)}\right\}$$

을 사람이 직접 판단하였기 때문에 분야연상어의 수는 변동될 가능성도 있으나, 시뮬레이션 결과는 크게 다르지 않을 것으로 예상된다.

3.2 복합 분야연상어의 평가

분야연상어 후보의 결정 알고리즘에 의해 복합 분야연상어 후보를 결정한다. 학습데이타에서 얻어진 약 18만 개의 복합어는 분야연상어 후보의 결정 알고리즘에 의해 후보어 총 수는 87,782개(수준 1, 2, 3, 4의 순으로 각각 74,257개; 5,815개; 181개; 8,529개)이며, 복합 분야연상어 결정 알고리즘에 의해 복합 분야연상어는 8,405개(수준 1, 2, 3, 4의 순으로 각각 6,188개; 915개; 98개; 1,204개)로 줄일 수 있었다.

본 논문에서 제안하는 방법을 평가하기 위해 여섯 개의 중간분야 <스포츠>, <취미·오락>, <건강·의료>, <정치>, <국제·지역>, <과학기술·학문>만을 선정하여 각각 네 가지 수준의 분야연상어 314개, 216개, 34개, 186개를 사람이 직접 결정하였다. 단, 수준 4의 분야연상어는 복수의 분야에 속하는 분야연상어이기 때문에 <스포츠>내의 분야연상어 후보에서 결정하였다. 분야연상어의 수를 E, 추출결과에 포함된 분야연상어의 수를 F, 추출된 분야연상어의 수를 G라고 하면, 재현율(recall)을 $R = \frac{F}{E}$, 정확률(precision)을 $P = \frac{F}{G}$로 표시한다.

재현율과 정확률을 그림 5에 표시하였다. 수준 1과 3의 여섯가지 종류의 결과를 각각 기호 □과 ●을 이용하여 원호 내에 표시하였다. 수준 2와 4는 복수분야의 평균치로 계산되어 각각 ○과 ■으로 표시하였다. 또한 비교실험을 하기 위해 표 4와 같이 일곱 가지의 실험 종류를 선정하고 이들 방법들을 서로 조합하여 그 결과를 분석하였다.

그림 5에 표시한 바와 같이, 실험 (A)와 (A, B1)를 제외한 모든 실험방법은 판정기준을 사용하였기 때문에 고정된 값의 분포(실험 (A)와 (A, B1)은 추출 단어 수의 변경결과를 표시)를 나타낸다. 그림 5에 의해 다음과 같이 다섯 가지의 평가결과를 알 수 있다.

(평가 ①) 방법 (A) 만으로는 정확률을 향상하지 못하지만, (A, B1)에서는 정확률이 향상하고, 복합 분야연상어 후보 결정 알고리즘의 (순서 B1)에서 분야 과잉 연상어 제거의 유용성을 알 수 있다.

(평가 ②) 안정성과 계승랭크를 단독으로 사용한 (A, Stability-5)와 (A, Inheritance-5)에서는 정확률의 개선은 적으나, (A, B1, Stability-5)와 (A, B-1, Inheritance-5)에서는 개선이 현저하다. 그 이유는 B1에 의한 분야 과잉 연상어의 제거 수가 Stability-5와 Inheritance-5에 의해 제거된 수에 비해 많기 때문이다.

(평가 ③) 두 개의 랭크를 조합한 제안방법 (A, B1, 25)에서 다시 정확률이 크게 향상함을 알 수 있다.

(평가 ④) 계승과 안정성랭크의 우선순위를 변경한 (A, B1, Reverse-25)는 제안방법보다 정확률이 저하하는데 그 원인은 안정성랭크를 우선함으로서 계승랭크가 높은 분야연상어가 추출되기 때문이다.

(평가 ⑤) 판정표에서 균등하게 분할 설정한 (A, B1, Uniform-25)의 제안방법은 정확률이 저하한다. 그

표 4 비교실험의 종류

순번	실험의 종류	설명
1	A	분야연상어 후보의 결정 알고리즘(집중률이 높은 순서) 만으로 분야연상어를 결정한 경우
2	B1	복합 분야연상어 후보의 결정 알고리즘(순서 B1)에 의해 분야 과잉 연상어를 제거한 경우
3	Stability-5	안정성랭크와 다섯 단계의 판정기준을 사용한 경우이다. 다섯 단계의 결정은 최대빈도와 평균빈도 사이를 두 단계로, 최소빈도와 평균빈도 사이를 두 단계로 균등하게 분할하여 설정한 경우
4	Inheritance-5	계승랭크 만으로 다섯 단계의 판정기준을 사용한 경우(다섯 단계의 분할설정은 (3)과 동일)
5	Reverse-25	25 단계의 판정기준에 의한 안정성랭크와 계승랭크를 우선 사용한 경우(다섯 단계의 분할 설정은 (3)과 동일)
6	Uniform-25	25 단계의 판정기준에 대하여 최대와 최저빈도 사이를 25단계로 균등 분할하여 설정한 경우
7	25	25 단계로 설정한 경우

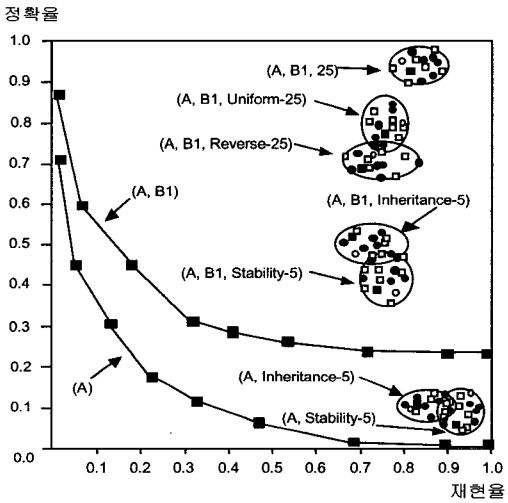

그림 5 정확률과 재현율에 의한 실험결과의 비교

이유는 균등분할에서 우선순위가 높은 분야연상어의 기준빈도가 제안한 25단계보다 크고, 낮은빈도에 존재하는 올바른 분야연상어가 제거되었기 때문이다.

이상의 결과로부터 본 논문에서 제안하는 방법은 재현율 0.77 이상(평균 0.85)을 유지하며, 정확률 0.90 이상(평균 0.94)을 실현하여 복합 분야연상어에 대한 유용한 추출법이라고 평가할 수 있다. 과잉 추출된 분야연상어 중에서 대상분야와 유사한 분야정보를 갖는 분야연상어도 포함시키면 정확률은 0.98 이상이 되어 비교적 노이즈가 적은 분야연상어를 추출할 수 있다. 본 논문에서 제시하는 복합 분야연상어 후보의 결정 알고리즘(순서 B3)에서 분야수의 감소에 의한 분야연상어의 수준상승을 고려하였으나, 이 수준변경을 수행하지 않는 경우는 제안방법 (A, B1, 25)의 재현율과 정확률이 각각 약 3%와 5% 저하되었기 때문에 (순서 B3)의 수준변경은 유용한 선택이었다고 생각된다.

요약하면, 사람이 수정하여 결정한 단일 분야연상어의 정보를 토대로 복합 분야연상어 후보를 자동으로 결정하는 알고리즘을 제시하였다. 제안된 방법은 단일 분야연상어에 대하여 사람이 직접 수정하였기 때문에 판정자에 의한 주관적인 개입이 생각되지만, 평가실험에서와 같이 25단계의 균등분할보다 제안방법에 의한 분할이 더 유효하다. 추출 정밀도에 대한 평가는 재현율과 정확률을 이용하여 수준 1, 2, 3, 4에 대한 고찰을 하였으며, 본 연구의 목표인 높은 재현율이 달성되었다.

3.3 문서단편에서 분야결정의 평가

학습데이타 이외의 문서[8]에 대하여 문서 내의 단락을 20문자 단위로 단계적으로 확장할 수 있는 윈도우를 하나 준비한다. 인간이 유일하게 하나의 종단분야를 인식한 단계에서 단락의 범위, 확정된 분야, 분야를 결정한 분야연상어 집합 X를 구축하였다. 문서의 시작부터 카운트하여 200 문자를 넘어도 분야가 결정되지 않는 단락[9]은 실험데이타로 채용하지 않는다. 이 단락 내의 단어에 대하여 단일어와 복합어의 분야연상어(23,733개)와 일치하는 분야연상어 집합 Y를 결정하여 다음의 방법에 따라 분야를 결정하였다.

수준 1, 2, 3, 4의 분야연상어가 해당하는 분야에 각각 12점, 8점, 4점, 2점씩 득점을 주었다. 단, 중간분야의 분야연상어에 대해서는 그 하위의 모든 종단분야에도 동일한 득점을 부여하였다. 그리하여 점수의 집계에서 최고득점의 종단분야를 정답인 분야(정답분야)로 결정하였다. 그림 6은 판단문서의 문자수에 대한 결과를 표시한다.

- 정답률 : 사람이 판정한 정답분야와 시스템이 결정한 분야를 비교한 비율
- 공유하는 분야연상어수의 비율 : 사람의 판단에 의한 분야연상어 집합 X와 Y에 공통으로 존재하는 분야연상어의 비율
- 문서수의 비율 : 문자수(단락의 길이)에 대한 문서수의 비율

그림 6과 같이 인간의 분야판정에 대하여 90% 이상의 높은 정답률을 얻을 수 있었다. 특히, 결정된 분야가 정답분야와 유사분야인 경우를 포함하면, 정답률은 약 97%가 된다. 문서수의 비율에서 알 수 있는 바와 같이, 단편문서는 약 80문장 이내가 약 90% 정도를 점유하며, 대단히 빠른 시간에 분야가 결정되어 단락에 대한 분야결정에 유용함을 알 수 있었다. 공유하는 분야연상어의 비율은 문서가 길어질수록 낮아지는데, 그 이유는 본 논문의 방법은 출현하는 모든 분야연상어를 이용하지만, 인간은 문서길이에 관계없이 분야판정에 사용하는 분야연상어가 거의 변화하지 않고, 대단히 적은 수(본 실험은 평균 2.3개)로 분야를 결정하기 때문이다. 분야결정에 유용한 분야연상어를 동적으로 축소하는 메커니즘은 차후에 계속 진행하여야 할 연구과제이다.

문서전체의 단어정보를 이용하는 방법은 특정분야의 단락을 검색하기 어려운 문제[14]가 있다. 예를 들면, 서두에 「일본 고베 대지진의 부흥」에 관한 화제가 출현하고, 그 뒤에 바로 이어서 「고교야구」에 관한 화제가 아주 길게 계속되는 문서는 전자의 "지진"에 대한 화제가 은폐된다. 따라서 분야연상어를 이용한 분야결정 방법은

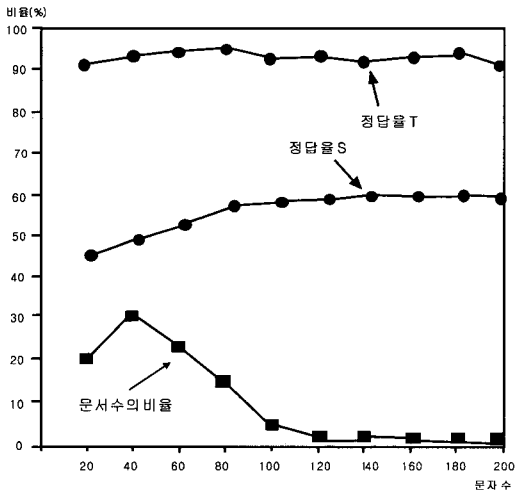

그림 6 분야결정의 실험결과

8) 문서의 추출원은 학습데이타와 같고, 출현된 문장의 선두단어가 수준 1의 분야연상어가 되는 문장은 제거하였다.

9) 300개의 대상문서 중 36문서는 문서 내에서 화제가 변화하여 인간의 판단으로도 종단분야를 결정할 수 없었다.

단락검색[5,14,16]이나 혹은 비교적 정밀도가 높은 Pin Point 검색에 유용하게 사용될 것으로 생각된다. 특히, 어휘적 연쇄(Lexical Chain)에 의한 단락검색법[11]에서는 유용한 분야연상어를 묶어서 연쇄로 이용하면 검색 효율의 개선을 기대할 수 있을 것이다. 본 논문의 목적은 양질의 분야연상어 구축이기 때문에 본 절에서는 득점 가산에 의한 간단한 분야결정 방법을 이용하였다. 차후에 단락검색을 위한 분야 결정법에 관해서도 연구할 계획이다.

4. 결정분야의 평가

이상으로 구축된 분야연상어는 어휘자체가 분야정보를 갖기 때문에 분야를 결정할 때 인간의 뇌와 동일한 인지작용에 의한 판단이 가능하다. 특히, 문서에 대해 인간은 대단히 적은 수의 분야연상어에서 분야를 결정할 수 있기 때문에 본 장에서는 문서의 분야결정에 관한 실험을 수행하여 분야연상어의 유효성을 평가한다. 분야연상어의 각 수준에 따라 가중치로 사용되는 득점을 가산하여 점수가 가장 높은 오직 하나의 분야를 최종적으로 채택한다.

실험결과에 의해 본 논문의 방법은 약 90% 이상의 정답률을 얻어 유효성이 검증되었다. 본 연구로 수집된 분야연상어를 이용하여 문서에 포함된 텍스트별 분류나 문서의 요약기술, 분야체계에 대한 확장정보 등을 연구하여 다른 검색기법에 관한 연구에 적용하고자 한다.

4.1 초기 인식의 분야결정

실험을 위해 264개의 문서를 선택하여 인간이 문서단편에서 한 문장 단위로 정답분야와 분야연상어를 추출하였다. 단, 분야를 판정할 수 없는 문장은 뒤에 따르는 문장과 연계하여 판단하였다. 이들의 정답데이타로 단편문서, 확정분야, 또는 분야결정의 요인이 되는 분야연상어 집합 X를 구축하였다.

예를 들어, 아래의 샘플데이타에서 슬래시(/)로 분류한 데이타는 문장번호, 실제문장, 연상분야와 그것을 결정하는 분야연상어를 나타낸다. 네 번째 문장까지 분야가 확정되지 않고, 다섯 번째 문장에서 정확한 분야와 분야연상어가 결정된다.

/1/올해는 팬의 수가 줄어드는 것을 염려한 야구계가 한 숨진 1년이었다./<스포츠/야구>/팬/야구계/ ····· **(예문 1)**

/2/ ...

/3/ ...

/4/주위의 열광에 의해 확실하게 떠오르는 새로운 스타상을 팬들은 느꼈다./<>//

/5/최종성적은 210안타 3할 8푼 5리/<스포츠/야구>/안타/ ····· **(예문 2)**

이 조각문서(단락에 해당)에서 출현하는 단어와 단일 혹은 복합 분야연상어(23,733개)와 일치하는 분야연상어 집합 Y를 추출하고, 다음의 방법으로 분야를 결정했다.

● 실제 문장에서의 분야결정 예

각 수준의 득점은 수준 1의 분야연상어가 출현하는 문장의 분야는 12점, 수준 2, 3, 4의 분야연상어는 각 수준별로 8, 4, 2점의 득점을 부여한다. 단, 안정성랭크가 'b' 혹은 'c'인 분야연상어는 득점의 ½만 가산한다. 중간분야의 분야연상어는 그 하위의 모든 종단분야에 동일한 득점을 부여한다.

위의 (예문 1)에 대해 분야결정의 예를 아래의 (예문 3)과 같이 표시하였다. "팬"이 수준 4의 분야연상어이고, 중간분야 <스포츠>와 <취미·오락>을 연상하고, 안정성랭크는 a이기 때문에 득점은 중간분야 아래에 있는 모든 종단분야에 2점씩 주어진다. 다음 분야연상어 "야구계"는 분야 <야구>의 분야연상어로서 수준 1, 안정성랭크는 'a'이기 때문에 종단분야 <야구>에 대하여 12점이 주어진다. 이 단계에서 최다득점 14점인 종단분야 <야구>가 결정된다.

/1/올해는 팬이 줄어드는 것을 염려한 야구계가 한 숨진 1년이었다./

(분야연상어, 수준, 안정성랭크, 확정분야, 득점)

("팬", 4, a, <스포츠>, 2)

("팬", 4, a, <취미-오락>, 2)

("야구계", 1, a, <스포츠/야구>, 12)

·····

<취미-오락> 2득점

<스포츠> 2득점

<스포츠/야구> 12득점 ····· **(예문 3)**

아래의 예도 동일하게 득점이 가산되지만, "박찬호"는 수준 1의 완전 분야연상어이고, 안정성랭크는 'a'이고, "LA Dodgers"는 수준 1의 분야연상어이지만 팀 명이기 때문에 안정성랭크는 'b'가 되어 수준 1의 득점 12의 ½인 6득점이 가산된다. 따라서 이 문장에서는 수준 1의 분야연상어가 두 개 존재하지만, 총득점은 16점이 된다.

/1/최대의 영웅은 박찬호(LA Dodgers)이었다.

("박찬호", 4, a, <스포츠/야구>, 10)

("LA Dodgers", 1, b, <스포츠/야구>, 6)

<야구> 16득점 ····· **(예문 4)**

4.2 결정분야의 실험 및 평가

이상과 같이 판정된 분야결정에 대하여 그림 6과 그림 7에 평가결과를 표시하였다.

(1) 정답률 : 사람이 평가한 정답분야와 비교하여 정답률 T를 구한다. 비교실험을 위해 단일 분야연상어만을 사용하여 정답률 S를 표시한다.

(2) 공유하는 분야연상어수의 비율 : 사람이 결정한 분야연상어 집합 X와 Y에 공통으로 존재하는 분야연상어의 비율, 위와 동일하지만, 단일 분야연상어만으로 구성된 집합을 S로 표시한다.

(3) 문서수의 비율 : 문자수(문자길이)마다 단편문서수의 비율

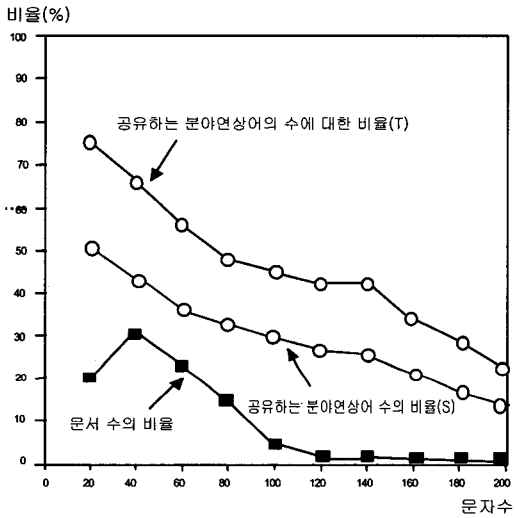

그림 7 공유하는 분야연상어 수의 비율

그림 6은 정답률과 문서수의 비율을 표시하고, 그림 7은 공유하는 분야연상어 수의 비율과 문서수의 비율을 표시한다. 그림에서 보는 바와 같이 인간의 분야판정과 비교해 보면 90% 이상의 높은 정답률을 얻고 있다. 특히, 최종 결정된 분야가 정답분야와 유사한 분야도 포함하면 정답률은 97%가 된다. 문서 수의 비율에서 알 수 있는 바와 같이 단편문서 중 약 80문자 이내의 문서비율이 약 90%를 점유하고 있으며, 대단히 빠른 단계에서 분야가 결정되고 있어 그 유용성을 알 수 있다.

분야연상어의 총수에 대응하는 단일 분야연상어의 비율은 약 65%이기 때문에 비율 T에 대한 비율 S의 저하율도 65%에 가까운 수치가 된다. 그러나 정답률 T에 대한 정답률 S의 저하율은 문자수가 적은 문서에 대해서는 이보다 낮은 50%가 되고 있다. 이것은 분야결정에 큰 영향을 미치는 수준 1의 분야연상어의 총수에 대응하는 수준 1의 단일 분야연상어 수가 약 54%정도이고, 역으로 분야 결정력이 약한 수준 4의 분야연상어의 총수에 대한 수준 4의 단일 분야연상어 수가 약 83%인 것이 이유인 것으로 보인다. 요약하면, 분야연상어 중에서 분야결정력이 강한 수준 1의 분야연상어가 점유하는 비율이 단일 분야연상어의 수준 1의 비율보다 많은 복합 분야연상어는 분야결정에 대단히 유용하다고 말할 수 있다.

공유하는 분야연상어의 비율은 문서의 길이가 길어질수록 낮아진다. 그 이유는 컴퓨터는 출현하는 모든 분야연상어를 조사하지만, 인간은 문서의 길이에 관계없이 분야판정에 이용하는 분야연상어의 수가 거의 변하지 않고, 대단히 적은 수(실험에서 평균 2.3개)를 이용하기 때문이다. 이와 같이 분야결정 중에 유용한 분야연상어를 동적으로 압축하는 메커니즘은 계속 연구하여야 할 과제이다.

문서전체의 단어정보를 사용하는 방법은 특정분야의 단락을 검색하는 것은 어려운 문제이다. 전체문서가 <야구>에 대한 문서이지만, 부분적으로 "동남아 지진해일"의 화제가 포함되어 있는 문서가 있다면, 문서전체의 정보를 이용하면 "지진해일"의 화제는 은폐되어 버린다. 따라서 분야연상어를 이용한 분야결정은 단락검색 혹은 정밀도가 높은 Pin Point 검색을 실현하기 위한 방법으로 생각된다. 특히, 어휘적 연쇄에 의한 단락검색법[5]에 대해서는 유용한 분야연상어의 연쇄를 작성하면 검색효율의 개선을 기대할 수 있을 것이다. 본 논문의 목적은 분야연상어 컬렉션의 구축이므로 본 절에서는 득점가산에 의해 비교적 간단한 분야결정법을 이용하였으나, 단락검색(Passage Retrieval)을 위한 분야결정 방법은 향후에 연구를 진행할 계획이다.

4.3 다국어간 분야번역과 문서요약에 응용

상식적 분야체계에 대한 다국어(multilingual)의 학습 데이타에서 분야연상어를 각 언어별로 구축하면 분야체계에 대한 다국어정보를 구축할 수 있다. 다국어로 구성된 문서로 추출된 분야정보에 보편성을 갖도록 고려하면 다국어 문서의 요약과 번역도 가능할 것으로 기대한다.

본 논문에서 제시하는 분야정보는 문서가 내포하는 상세한 의미나 정보를 인간에게 제공할 수는 없으나, 인간과 컴퓨터의 초기적인/기초적인 통신수단으로 이용하면 중요한 초점정보를 사용자에게 제공할 수 있다. 그러나 표층적인 단어를 번역하는 간단한 처리만으로 적당한 분야를 결정할 수 없는 경우도 있다. 예를 들면 <야구>에 대한 분야연상어 "김응용감독"이나 "나가시마감독"을 영어로 번역하더라도 반대로, 야구분야의 분야연상어 "Babe Ruth"를 한국어나 일본어로 번역하더라도 분야연상어로 적당하다고 볼 수 없기 때문에 단지 <야구>만이 분야정보로서 유용한 정보가 된다. 이와 같이

001: 애리조나 김병현(24)이 3연패 끝에 감격의 시즌 첫 승을 따냈다.
(분야연상어, 수준, 안정성랭크, 분야, 득점)
(none, x, x, (/), 0)
<> 0득점

002: 김병현은 20일(이하 한국시간) 부시스타디움에서 벌어진 세인트루이스 카디널스와의 원정경기에서 7이닝 5안타 3실점(볼넷 3개, 탈삼진 1개)으로 호투하고, 팀이 4대3으로 승리해 데뷔 첫 선발승을 거뒀다.
(분야연상어, 수준, 안정성랭크, 분야, 득점)
(원정경기, 4, a, (/스포츠), 2)
(이닝, 1, a, (/스포츠/야구), 12)
(안타, 1, a, (/스포츠/야구), 12)
(볼넷, 1, a, (/스포츠/야구), 12)
(탈삼진, 1, a, (/스포츠/야구), 12)
(선발승, 1, a, (/스포츠/야구), 12)
<야구> 62득점

003: 고비가 많았던데 반해 실점은 최소화하는 위기 관리 능력이 돋보였다.
(분야연상어, 수준, 안정성랭크, 분야, 득점)
(none, x, x, (/), 0)
<> 0득점

004: 2회 선두 4번 짐 에드몬즈에게 볼넷을 내준 뒤 5번 에드가 렌테리아에게 가운데 펜스를 원바운드로 넘어가는 2루타를 허용해 무사 2, 3루의 절대 위기.
(분야연상어, 수준, 안정성랭크, 분야, 득점)
(펜스, 4, a, (/스포츠/야구), 2)
(루타, 1, a, (/스포츠/야구), 12)
<야구> 14득점

005: 그러나 7번 티노 마르티네스를 알은 중견수 플라이로 처리하고, 7번 엘리 마레로의 땅볼을 직접 잡아 홈에 송구해 3루 주자 에드몬즈를 태그아웃시켜 실점을 막았다.
(분야연상어, 수준, 안정성랭크, 분야, 득점)
(중견수, 1, a, (/스포츠/야구), 12)
(땅볼, 1, a, (/스포츠/야구), 12)
(태그아웃, 1, a, (/스포츠/야구), 12)
<야구> 36득점

그림 8 <야구>의 정답데이타, 분야결정, 득점 예

각 나라의 문화(文化)와 그 나라 고유의 분야정보는 각 나라의 문화에 적합하도록 독립적으로 구축되어야 한다.

예를 들어, 그림 8은 야구문서의 예이다. 각 국의 독자는 쉽게 분야를 결정할 수 있지만, 보편적인 분야연상어(대역사전에서 대응되는 안정성랭크가 a인 분야연상어)가 적고, 단순히 번역하는 것만으로 적당한 분야를 판단하는 일은 곤란하다.

4.4 분야체계의 확장

부록 A의 분야체계는 문서 수집의 용이성을 우선으로 고려하여 수집하였기 때문에 상식적인 분야체계로 사용하기에는 부적절한 분야도 존재한다. 따라서, 향후에 분야체계를 수정·보완하여야 한다. 저자의 실험실에서 이미 한국어 분야체계를 위한 문서데이타의 수집을 시작하고 있다. 이에 의해 본 연구의 실험평가도 보다 정확하게 고찰될 것으로 생각된다. 또한 확장시킬 분야체계에 대해 계속 연구하여야 한다. 그 체계의 일부는 참고문헌 [15]의 부록 B를 이용하였다.

5. 결 론

본 논문에서는 단일어에 대한 분야연상어 정보[15]를 이용하여 일상생활에서 끊임없이 생성되는 복합 분야연상어를 효율적으로 결정하는 방법을 제안하여 180분야의 학습데이타에 대한 실험결과를 기반으로 제안방법의 유효성을 평가하였다. 구축된 분야연상어가 문서 내의 단락의 분야결정에 유효한 것인가에 대해 논의하였다.

분야연상어를 단일과 복합 분야연상어로 분류하여 단일 분야연상어를 형태소사전에 등록된 표제어와 일치하도록 한정하였다. 이것은 단일 분야연상어의 분야정보를 형태소사전에 그대로 등록하기 위한 실용성을 고려한 것이다. 또한, 본 연구에서는 분야체계를 미리 정의한다고 하였으나, 분야연상어 구축은 어떠한 분야체계에도 손쉽게 적용될 수 있으므로 보편성은 충분하다고 생각된다. 다음은 향후에 수행할 과제에 대하여 논의한다.

- 분야연상어가 시간경과에 의해 변화하는데 주목하여 안정성랭크를 이용하였으나, 이 랭크에 대해 보다 상세한 분석에 의해 정확한 랭크의 정의가 존재할 가능성이 있기 때문에 향후 검토할 계획이다.
- 인명에 해당하는 고유명사에 대해서는 "김대중대통령"의 실체를 이해하는 것과 이를 "박찬호투수"와 구별하는 것이 애매하므로 문서에서 독립된 단어의 실체를 이해하는 방법이 필요하다.

본 논문에서는 학습데이타에서 분야연상어의 후보와 그 수준을 자동적으로 결정하는 알고리즘을 제안하였다. 학습데이타의 불균형성에 대해서는 상대빈도를 이용하여 빈도를 정규화하고, 분야연상어가 특정분야에 집중하는 기준율 α를 정의하였다. 이 연구에 대해 다음과 같은 과제가 생각된다.

- 적은 수의 문서가 수집된 분야에서 정규화 한 빈도가 대단히 크게되는 점을 보완할 필요가 있다.

제안한 알고리즘은 분야트리의 루트가 되는 <전체분야>에서 기준율 α이상에 집중하는 특정분야를 탐색하고 그 조작을 하위의 분야들로 진행하여 종단분야에 도달하면 수준 1의 분야연상어와 특정분야를 결정하였다. 조작 중 언제나 동일한 기준치 α를 이용하였으나 다음 문제들을 검사하여야 한다.

- 기준율 α를 만족하지 않으나, α에 대단히 가까운 분야연상어의 취급은 어떻게 할 것인가? 이 점은 실제 수집된 분야연상어 데이타를 충분히 분석하여야 한다.

먼저, 단일 분야연상어의 분야계승에 기초한 복합 분야연상어의 분석을 논의하였고, 의미계승과 분야계승에 관련한 본 연구의 방법은 대단히 유용하며, 다음의 과제가 필요하다.

- 안정성랭크가 'c'인 것(인명에 대한 고유명사)이 분야

계승에 얼마만큼 관계하는가? 다시 말하면, 연상분야의 지속성과 분야계승의 정도(세기)에 대해 연구하여야 한다.

- 분야규칙에 대해서도 미래의 연구과제로 본 논문과의 관련성을 논의할 필요가 있다.

180분야로 분류된 약 15,000 파일의 실험결과에 의해 제안방법의 유효성을 증명하였다. 인간의 지식으로 구축한 단일 분야연상어를 이용하여 알고리즘에 의해 압축된 복합 분야연상어의 정밀도를 평가하였으나, 다음의 후속연구가 필요하다.

- 구축한 문서데이타가 신문데이타에 의존하고 있으므로 반드시 일반문서에 대한 후속실험을 수행할 필요가 있다.

초기단계에서 분야결정의 실험을 수행해 분야연상어의 유효성을 평가하였다. 실험에서는 사람이 단락의 정답데이타를 작성해 분야결정의 정밀도를 평가하였다. 여기서 수준별로 득점을 부여하는 방법을 사용하였기 때문에 다음과 같은 보충연구가 필요하다.

- <프로야구>, <고교야구>, <소프트볼> 등의 분야로 구성된 문서에서 "투수", "홈런" 등은 수준 1의 분야연상어 컬렉션에서 제거되기 때문에 분야체계와 수준의 변화에 동적으로 반응하는 방법이 필요하다.

지금까지는 일반적인 분야체계를 대상으로 논의를 진행하였으나, 단일 분야연상어를 사람이 판단할 수 있기 때문에 각 기관의 이용 목적에 맞도록 분야체계를 결정하여야 한다. 본 논문은 명사연속의 복합어를 대상으로 하였으나, 용언이나 조사를 포함한 명사구, 명사와 용언의 조합 등과 같이 출현하는 공기정보와 분야연상어의 관계도 검사할 필요가 있다. 이상에서 논의한 바와 같이 분야연상어에 관한 연구는 끊임없이 계속되어야 할 연구과제이다. 결론적으로 본 논문은 자연언어처리 과정을 인간의 인지과정과 유사한 복합 정보의 비교 및 분석방법의 형태로 흉내낼 수 있는 가능성을 보였다는 점에서도 그 의의가 있다고 말할 수 있다.

참 고 문 헌

[1] Edwin Williams, On the Notions "Lexically Related and Head of a Word," Linguistic Inquiry, Vol. 12, No. 2, pp. 245-274, 1981.

[2] Fumiyo Fukumoto et al., "Automatic Clustering of Articles Using Dictionary Definition," Transactions of Information Processing Society of Japan, Vol. 37, No. 10, pp. 1789-1799, 1996 (in Japanese).

[3] M. J. Blosseville et al., "Automatic Document Classification: Natural Language Processing, Statistical Analysis, and Expert System Techniques Used Together," Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '92), pp. 51-58, 1992.

[4] Masami Hara et al., "Keyword Extraction Using a Text Format and Word Importance in a Specific Field," Transactions of Information Processing Society of Japan, Vol. 38, No. 2, pp. 299-309, 1997 (in Japanese).

[5] Mochizuki, H., Makoto, I., and Okumura, M. "Passage-Level Document Retrieval Using Lexical Chains," Journal of Natural Language Processing," Vol. 6, No. 3, pp. 101-126, 1999 (in Japanese).

[6] Naoyuki Nomura, "ConceptBase-A NL-based IT Solution Core," Proceedings of the 1999, the 18th International Conference on Computer Processing of Oriental Language (ICCPOL '99), p. 235, 1999.

[7] Norbert Fuhr, "Models for Retrieval with Probabilistic Indexing," Information Processing & Management, Vol. 25, No. 1, pp. 55-72, 1989.

[8] Salton, G. and McGill, M. J., "Introduction of Modern Information Retrieval," McGraw-Hill Book Company, 1983.

[9] Salton, G., "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer," Addison-Wesley Publishing Company, 1989.

[10] Tokunaga, T. and Iwayama, M., "Text Categorization based on Weighted Inverse Document Frequency," Natural Language Processing, Vol. 100, No. 5, 1994. (in Japanese)

[11] Tsuji, T., Nigazawa, H., Okada, M., & Aoe, J., "Early Field Recognition by Using Field Association Words," Paper Presented at the Proceedings of the 18th International Conference on Computer Processing of Oriental Language (ICCPOL '99), 1999.

[12] Yoshitaka Hayashi et al., "Efficient Method for Extracting Keywords of Compound Words Using Pattern Matching Machines," Transactions of Information Processing Society of Japan, Vol. 38, No. 4, pp. 815-825, 1997 (in Japanese).

[13] 남영신, 우리말 분류 사전, 성안당, 2001.

[14] 이상곤, "분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할 방법", 정보처리학회논문지(B), 제 10권, 제 1호, pp. 57-66, 2003.

[15] 이상곤, 이완권, "분야연상어의 수집과 추출 알고리즘", 정보처리학회논문지(B), 제 10권, 제 3호, pp. 347-358, 2003.

[16] 이상곤, "분야연상어를 이용한 화제분야의 계산방법과 단락검색", 정보처리학회논문지(B), 제 12권, 제 1호, pp. 57-68, 2005.

부록 A. 분야체계와 수집데이타의 양

분야체계의 루트에 바로 아래에 있는 자식분야를 다음에 기술한다. 분야명은 < >내에 기술하고, 괄호는 파일 수와 KByte를 표시하였다. 이 분야의 하위분야는 괄

호로 표시한다.

<분야전체(15,435 ; 42,092)>
<스포츠(1,856 ; 5,527)>
골프, 야구, 배드민턴, 농구, 유도, 야구, 레슬링, 배구, 테니스, 태권도, 쓰모, 복싱, 축구, 럭비, 수영, 검도, 동계스포츠(스키, 스케이트, 스키점프, 봅슬레이), 육상(포환던지기, 해머던지기, 원반던지기, 100m, 마라톤, 삼단뛰기), 모터스포츠(F1, 모터크로스, 보드)
<취미-오락(1,680 ; 4,891)>
애니메이션, 희극, 장기, 컴퓨터게임, 여행, 영화, 경마, 요리-식음료, 예술, 독서, 음악, TV
<과학기술-학문(735 ; 7,074)>
우주개발, 해양개발, 군사기술, 건축, 원자력, 전기전자, 재료, 화학, 수학, 물리학, 고고학, 언어학, 생물학-바이오, 컴퓨터(S/W, H/W)
<자연(102 ; 517)>
지구과학, 지진-화산, 천문학, 기상학
<건강-의료(514 ; 3,708)>
진단, 병명(O-157, 아토피성피부염, 에이즈, 암, 당뇨병, 간 질환), 건강(다이어트, 스트레스, 콜레스테롤, 혈압)
<환경(1,618 ; 2,782)>
인구증가, 공해, UN정책, 자연재해, 오존파괴, 쓰레기문제, 에너지, 온난화, 도로-교통
<교육(1,622 ; 4,102)>
교육기관, 교사-교수, 수험-입시, 외국어교육, 교육장소, 학교행사, 자격, 교육교재, 학벌, 교육문제(왕따, 장기결석)
<사회(1,104 ; 1,824)>
광고, 유행, 문화활동, 전쟁-분쟁, 저널리즘, 사건-사고, 재해(화재, 지진, 수해, 태풍)
<생활(998 ; 1,742)>
주택, 식생활, 여성, 보험, 연금, 가족-가정, 복지, 세금
<국제관계(2,179 ; 3,991)>
아시아, 중국, 일본, 한국, 북한, 오세아니아, 유럽, 미국, 캐나다, 중동, 남미, 아프리카, 중동, 소련, 북극-남극
<정치(2,026 ; 4,910)>
사법, 국회, 압력단체, 지방자치, 외교, 헌법, 정당, 정치이론, 선거, 한국정치, 국제정치, 세제, 행정-내각, 방위
<경제(1,011 ; 4,024)>
세계경제, 노동, 외화, 금융, 미국경제, 일본경제, 한국경제, 경영, 재무회계, 금융일반, 재정, 무역, 농업, 어업, 경기-물가, 주식-채권, 마케팅

이 상 곤

1994년 전주대학교 영어영문학과(이학사) 1996년 전북대학교 컴퓨터과학과(이학사) 1998년 전북대학교 전산통계학과(이학석사). 2001년 日本 도쿠시마대학교 지능정보공학과(공학박사). 2001년~2002년 원광대학교 음성정보 기술산업 지원센터 연구원. 2002년~현재 전주대학교 정보기술공학부 컴퓨터공학전공 조교수. 관심분야는 한국어 정보처리, 한글공학, 정보검색, 문서분류 및 요약, 키워드추출, 컴파일러, 인공지능