한국어의 분야 연상 지식의 추출 방법에 관한 연구

이 상 곤○

전주대학교 공과대학 컴퓨터공학과
samuel@jj.ac.kr

Automatic Dictionary Construction of Indonesian Field-Associated Terms by Using Korean Associated Knowledge

요 약

인간은 문서전체를 읽지 않고 대표적인 단어를 보는 것만으로 정치나 스포츠 등의 분야를 정확히 인지할 수 있다. 문서 전체는 물론 부분 텍스트(단락)에 출현하는 소수의 단어 정보에서 문서의 분야를 정확히 결정하기 위한 분야연상어의 구축은 중요한 연구과제이다. 미리 분야체계를 정의하고, 각 분야에 해당하는 문서를 인터넷이나 서적을 통해 수집한다. 본 논문은 수집 문서의 분야를 정확히 지시하는 분야연상어를 수집하는 방법을 제안한다. 문서의 분야결정 시점을 고려하여 분야연상어의 수준을 정하였다. 인도네시아어의 분야연상어 사전을 자동으로 구축하기 위해 먼저 한국어로 구축한 분야 연상 지식을 추출하는 방법을 제안한다.

주제어: 분야 연상어, 분야 트리, 종단분야, 연상 지식, 지식 추출, 분야연상어 수준

1. 서론

근래에는 컴퓨터 하드웨어의 소형화, 대용량화, 저가격화, LAN의 고속화 등에 의해 컴퓨터는 금융기관, SNS, 회사의 시스템 등을 비롯하여 사회의 여러 곳에 침투하여 폭넓은 용도로 사용되고 있다. 컴퓨터는 대규모의 정보를 보존하고 빠른 검색 및 불필요한 데이터의 삭제 등 여러 장점이 있다. 이에 따라 컴퓨터를 이용한 디지털화된 문서 데이터의 양도 크게 늘어나고 있다. 컴퓨터를 이용하여 문서를 자동으로 처리하는 소프트웨어의 필요성이 높아지고 있으며, 방대한 양의 문서를 주제 분야에 대응하여 자동으로 분류하는 기술이 절실하다.

문서를 분류한다는 것은 문서를 내용에 따라 주제별로 자동 분류하는 방법이다. 문서 분류 방법은 기계학습과 같이 통계적 방법을 이용하는 것이 주류이다. 그러나 통계적인 방법을 사용하기 위해서는 학습 데이터로 대량의 문서를 준비할 필요가 있고 문서 분류의 최종 결과도 매우 정확해야 한다[1]. 통계적 방법을 사용하지 않는 문서 분류 방법 중의 하나로 분야연상어의 방법이 있다[3]. 분야연상어란 문서의 특정 분야를 연상할 수 있는 직관적인 단어를 말한다. 인간이 직관적으로 문서 내에서 분야연상어를 추출함으로써 문서의 주제를 파악할 수 있는 단어이다.

방대한 분야연상어의 데이터로 구성된 사전을 사용하여 주제 분야를 추출하는 시스템은 전자 문서를 분류하는 기술 중의 하나이다. 주제 분야를 추출하는 시스템은 어떤 임의의 문서를 읽어 들여 그 문서와 관련 있는 가장 가까운 분야를 추출하는 시스템이다.

본 연구에서는 오랫동안 연구해 온 한국어의 분야연상어를 기반으로 인도네시아어의 문서를 자동으로 분류할 목적으로 사용하기 위한 분야연상어 사전의 자동 구축을 목표로 한다. 2장에서는 분야연상어의 정의와 문서 분류 방법에 대하여 설명하고, 문제점에 관하여 논의한다. 본 논문에서 제안하는 분야연상어 사전의 구축 방법에 관하여 서술한다. 마지막으로 3장에서는 결론과 향후의 연구 과제에 대하여 서술한다.

2. 분야연상어와 분야 체계의 정의

2.1 분야연상어

본 절에서는 분야연상어의 정의와 개요에 대하여 자세히 설명한다. 분야연상어란 인간이 특정 분야를 연상하는 것이 가능한 단어를 말한다. 예를 들면 <농구>의 분야연상어로 “기브 앤 고(give and go)[1]”, “트래블링(travelling)[2]” 등의 보통 명사나 혹은 “마이클조단”, “로스앤젤레스 레이커스(미국 프로농구 구단)” 등의 인명이나 조직명이 있다.

1) 둘이서 주고받으면서 나아가는 플레이. 농구에서 패스한 다음 방어자를 따돌리기 위해 움직이고, 리턴 패스를 받아 슛하려 하는 플레이
2) 선수가 볼을 가지고 3보 이상 움직이는 반칙

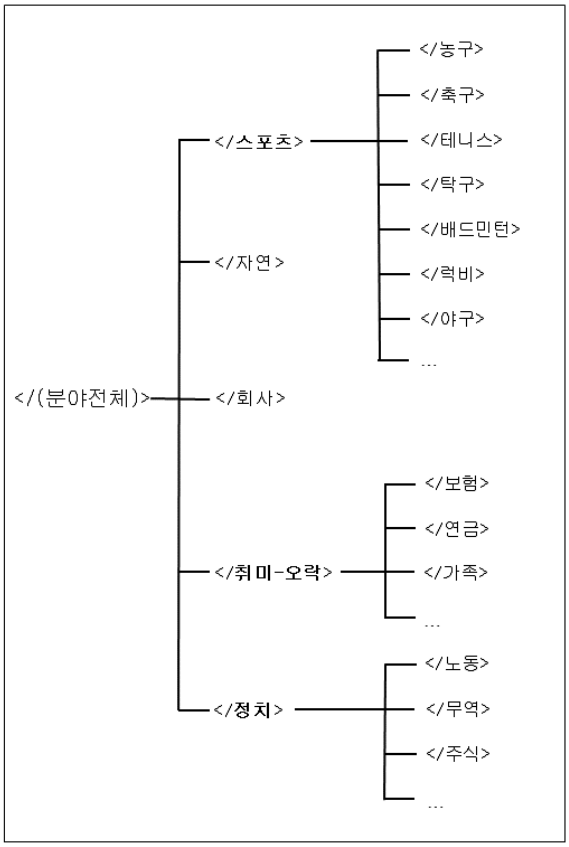
<표 1> 연상되는 분야와 분야연상어

분야	분야연상어(득점)
<농구>	기브 앤 고(80), 사이드 핸드 패스(80), 한국농구협회(60), FIBA(40), 트래블링(30) 등
<패션>	베스트 드래스 상(80), 패션쇼(60), 크리스찬 디오르(60), 한국청바지협의회(50) 등
<경제>	한국경제(60), 한국은행(40), 엔고(30), 엔화환전(30), 물가(30), 금융시장(30), 금리정책(30) 등
<정치>	외교정책(80), 국회의원(40), 하원(40), 상원(40), 일원제(30), 상임위원(30), 재정(30), 국가(30) 등

단, “경우”, “사용” 과 같은 단어는 특정 분야를 연상할 수 없기에 분야연상어로 적절하지 않다. 분야연상어는 미리 정의되어 있는 분야 체계를 기반으로 구축되어 분야연상어 사전에 등록되어 있다. 등록되어 있는 분야연상어에는 연상할 수 있는 분야에 대한 득점이 부여되어 있으며, 그 값은 분야를 연상할 수 있는 정도에 따라 설정된다. 위의 <표 1>에 분야연상어와 득점의 예를 나타내었다. 이 표에서 “한국경제”와 “금융시장”은 <금융>이라는 다른 분야의 연상이 가능한 단어이지만 “한국경제”는 <경제>만을 연상할 수 있는 단어이다. 따라서 “금융시장”의 득점은 30점으로 낮은 득점이 설정되어 있다. 또한 분야연상어에는 단일 단어(단일어) 혹은 복합어로 구성된 분야연상어[1]가 있다. 여기서 단일어란 형태소 사전에 등록되어 있는 단어를 의미하며 복합어는 두 개 이상의 단일어로 구성된 단어를 의미한다. 분야연상어 중에서 단일어인 것을 단일 분야연상어, 복합어로 구성되어 있는 단어를 복합 분야연상어라고 부르기로 한다[1, 3]. 예를 들면 앞에서 언급한 “트래블링”과 같은 분야연상어는 농구 분야의 단일 분야연상어이며, “한국농구협회”는 각각 “한국” 과 “농구”와 “협회”를 조합한 복합어는 복합 분야연상어이다. 본 논문에서 분야의 표기는 다음과 같다. 농구와 같이 분야명은 < > 안에 표기(<농구>)하고, 트래블링처럼 분야연상어는 “ ” 안에 표기(“트래블링”)하기로 한다.

2.2 분야 체계

분야 체계란 주제 분야의 집합을 의미한다. 분야 체계는 주제 분야의 상-하위 관계를 트리 구조로 표현하고 있다. 이 트리 구조를 분야 트리라 부른다. 분야 트리의 리프(leaf)에 해당하는 주제 분야를 종단 분야(terminal field)라 부르고, 그 이외의 노드는 모두 중간 분야(intermediate field)라 부른다. 또한 직접적인 상-하위 관계를 갖는 주제 분야의 상위 분야를 부모 분야(parent



(그림 1) 분야 체계의 예

field)라 하며 하위 분야를 자식 분야(child field)라 부른다. 위의 (그림 1)에 분야 트리의 예를 표시하였다. 예를 들면, <스포츠>나 <생활>, <경제> 등은 중간 분야라 부르고 <야구>, <보험>, <주식> 등은 종단 분야라 부른다. 분야 지정은 분야명의 경로(path)를 기호 <S>라 기술하고 루트(root)에 해당하는 <전체 분야>는 생략한다. 또한 특별한 모순이 생기지 않는 경우에는 경로 지정을 생략하고 종단 분야만으로 문서의 해당 분야를 기술한다. 예를 들면 분야 경로 <S> = </스포츠/럭비>는 <스포츠>의 하위 종단 분야 <럭비>만으로 표시한다.

2.3 분야연상어 수준의 결정

분야연상어는 연상하는 분야의 범위에 차이가 있다. 즉 단지 하나의 종단 분야나 중간 분야를 연상하는 단어가 있는 반면에 복수의 종단 분야 또는 중간 분야를 연상하는 단어도 존재한다. 분야연상어는 그 연상하는 범위에 따라 다음과 같이 다섯 가지의 수준으로 분류할 수 있다.

[정의] 분야연상어의 수준(Level)

(수준 1) 완전 분야연상어 : 단 하나의 종단 분야만을 연상하는 단어

(수준 2) 준완전 분야연상어 : 같은 부모 분야를 갖는 종단 분야만을 연상하는 단어

(수준 3) 중간 분야연상어 : 단 하나의 중간 분야만을 연상하는 단어

(수준 4) 다분야연상어 : 복수의 중간 분야 혹은 종단 분야를 연상하는 단어

(수준 5) 비분야연상어 : 주제 분야를 하나로 연상하지 않는 단어

<표 2> 분야연상어와 수준의 할당 예

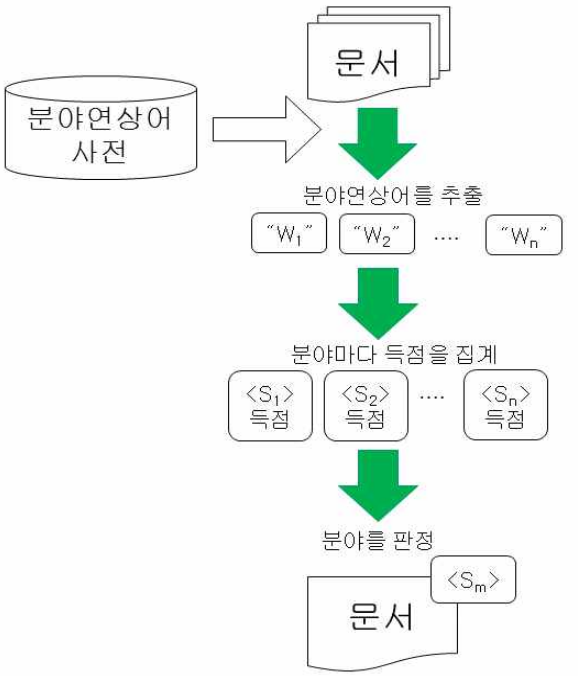
분야연상어	연상 분야	수준
레이커스	</스포츠/농구>	1
아베신조	</정치/정치>	1
프리킥	</스포츠/미식축구>	2
	</스포츠/축구>	
이율 곡선	</돈/주식>	2
	</돈/금융>	
시합	</스포츠>	3
금융 시장	</돈>	3
드래곤 볼	</만화>	4
	</영상/애니메이션>	
Nike	</패션/의류>	4
	<스포츠/스포츠웨어>	
물	없음	5

수준 1의 완전 분야연상어는 “국기원”과 같이 종단 분야 <태권도>를 단번에 연상하는 단어를 말한다. 수준 2의 준완전 분야연상어는 “단식”, “복식”처럼 복수의 종단 분야 <테니스>, <탁구>, <배드민턴>을 연상하는 단어를 말한다. 수준 3의 중간 분야연상어는 “시합”과 같이 종단 분야를 연상할 수는 없지만 하나의 중간 분야를 연상하는 단어를 말한다. 수준 4의 다분야연상어는 “승패”처럼 복수의 종단 분야 <취미/장기> 또는 중간 분야 <스포츠>를 연상하는 단어를 말한다. 수준 5의 비분야연상어는 “경우”나 “사용”처럼 특정한 주제 분야를 한정하지 않는 단어이다. <표 2.2>에서 이러한 분야연상어의 수준을 나타내었다.

2.4 분류 기술에 적용

분야연상어를 이용한 문서 분류 방법에 대하여 설명하면 다음과 같다. 문서 분류 방법의 흐름도를 아래의 (그림 2)에 나타내었다. 먼저 처리 순서를 대략적으로 구분하면 다음과 같이 세 가지로 나눌 수 있다.

- **[단계 1] 분야연상어의 추출**
- **[단계 2] 분야마다 득점을 집계**
- **[단계 3] 분야를 판정**



(그림 2) 문서 분류 방법의 흐름도

아래에 위의 세 가지의 문서 분류 처리 단계를 자세하게 서술한다.

[문서 A]
주요한 산업별 노동조합인 한국기간산업노동조합연합회(기간노연)에 가입하는 철강이나 조선해운, 비철금속노조가 12일 임금 인상이나 노동 조건 개선 등의 요구서를 일제히 회사 측에 제출하고, 2016년 봄에 투쟁 교섭을 본격적으로 시작하였다.

철강 최대 기업의 포스코 본사에서는 오전 10시 서울경인금속가구종업협동조합의 김성곤 회장이 「생산성 향상과 국제 경쟁력 강화를 위해 그룹 또는 관련기업에서 일하는 모든 사람들의 노동 조건 개선에 지원을 요청한다」고 말하고 2년에 총 8,000 만원의 임금 개선 등을 담은 요구서를 김○○상무에게 보냈다.

철강업계는 중국의 경기 위축에 의해 실적이 빠른 속도로 악화되고 있다. 포스코의 경영진은 「둘러싼 상황이 더욱 더 힘들어지는 가운데 고정적인 임금 증가에 이어지는 시책은 도저히 받아들일 수 없는 상황이다」그리고 낮은 임금 상승에 대해서는 중요한 생각을 표명하였다.

그 후, 자동차나 전기 업계의 노조도 요구서를 제출하며 3월 중순의 회의에 결론을 내기 위해 교섭을 계속 중이다.

(밑줄 친 단어 : 분야연상어)

(그림 3) 문서 A에서 인식된 분야연상어

- **[단계 1] 분야연상어의 추출**

분야연상어 사전을 이용해 문서 내에 포함된 분야연상어를 전부 추출한다. 예를 들면 (그림 3)에 있는 문서에서 나타난 분야연상어를 밑줄 친 단어로 표시하였다. 그리고 추출한 분야연상어의 분야와 득점을 아래의 <표 3>에 나타내었다.

<표 3> (그림 3)에서 추출된 분야연상어의 분야와 득점

분야연상어	분야(득점)
“노조”	</비즈니스> (40)
“노동조합”	</비즈니스> (40)
“관련기업”	</비즈니스> (40)
“산업노동조합”	</비즈니스> (40)
“업적”	</비즈니스> (40)
“임금”	</비즈니스> (30)
“경영”	</비즈니스> (30)
“투쟁 교섭”	</비즈니스> (30)
“노동조건”	</비즈니스> (30)
“생산성향상”	</비즈니스> (30)
“경기”	</경제> (30)
“국제경쟁력”	</경제> (30)
“조선”	</선박> (30)
“철강”	</철강원자재> (30)
“자동차”	</자동차> (30)

- **[단계 2] 분야마다 득점을 집계**

[단계 1]의 (그림 3)과 <표 3>에서 보는 바와 같이 문서에서 추출한 분야연상어가 전부 같은 한 분야를 연상하지 않을 수 있다. 어떤 분야가 가장 적절한가를 판정하기 위해 분야연상어에 부여되어 있는 득점을 분야별로 집계한다. 어떤 분야 $<S>$를 연상할 수 있는 분야연상어가 $n(n \geq 1)$ 종류로 추출되었을 때 분야 $<S>$의 득점은 분야연상어를 $w_i(i = 1, 2, ..., n)$로 하여 아래의 식 (1)로부터 계산한다. 여기서 분야연상어의 출현 횟수는 실제 시스템에 반영하기 전에는 고려하지 않기로 한다.

$$<S>\text{의 득점} = \sum_{i=1}^{n}(w_i\text{에 부여되어 있는 득점}) \quad \cdots\cdots\cdots\cdots (1)$$

문서 A에 대한 집계 결과를 <표 4>에 제시하였다.

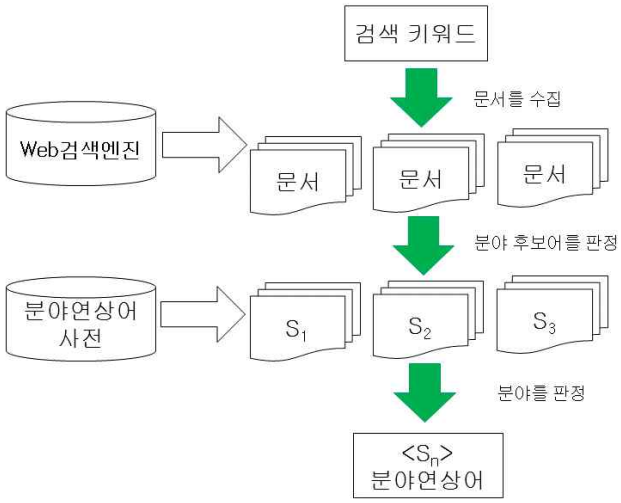
- **[단계 3] 분야 판정**

집계 득점이 가장 높은 분야를 문서의 최종 분야로 판정한다. 문서 A의 분야는 아래의 <표 4>를 참고하여 </비즈니스>가 된다. 만약 득점이 가장 높은 분야가 복수 개인 경우는 복수 분야를 문장의 분야로 판정한다. 복수의 분야가 추출되었을 때 대처하는 방법은 단락 분할 방법[2]을 이용할 수 있다. 여기서 화제 분야의 출현과 화제 분야의 계속 혹은 전환 여부는 참고문헌 [4]를 참고하여 설계한 시스템을 참고 문헌 [6, 7]을 이용하였다. 또한 문서로부터 분야연상어가 전혀 추출되지 않았을 경우는 <분야 미정(field Neutral)>이라고 판정한다[2, 6].

2.5 분야연상어의 웹 추론

<표 4> 득점 집계 결과

분야	득점
<비즈니스>	350
<경제>	60
<선박>	30
<철강원자재>	30
<자동차>	30



(그림 4) 분야연상어의 웹 추론 방법

웹(web) 추론이란 검색 엔진과 문서 분류를 이용해 단어 하나하나에 대해 분야를 자동으로 추정하는 방법이다. 본 연구자가 2006년도에 검색과 분류가 동시에 가능한 엔진을 설계하고 구현[7]하였으나 이후 웹 추론이란 새로운 전문용어가 만들어져 본 논문에서는 웹 추론이란 단어를 이용한다. 이 방법의 흐름도를 위의 (그림 4)에 제시하였다. 처리 순서는 다음과 세 단계로 구성하였다.

- **[단계 1] 웹 검색 엔진으로부터 문서를 취득**
- **[단계 2] 분야 후보를 판정**
- **[단계 3] 분야를 판정**

- **[단계 1] 웹 검색 엔진으로부터 문서를 취득**

분야연상어 후보를 검색 키워드로 하여 웹 검색 엔진으로부터 문서를 20건 취득한다. 이 문서는 요약문이다. 요약문이란 웹 검색 엔진의 검색 결과 페이지에서 제목 바로 밑에 표시되는 문서라 한다.

본 논문에서 HTML 문서가 아닌 요약을 이용하는 이유는 두 가지이다. 첫 번째는 HTML 문서와 비교했을 때 요약된 글은 짧은 시간에 문서를 수집할 수 있다는 장점이 있다. 이것은 HTML 문서의 경우는 1건씩 접근할 필요가 있지만 요약의 경우는 한 번의 접근으로 수십 건의 요약을 취득할 수 있기 때문이다. 두 번째는

HTML 문서보다 검색 키워드와 연관성이 높은 문서를 취득할 수 있지만 광고 문구도 함께 수집할 위험이 있다. 그러나 요약에 광고 문구가 포함되는 경우가 매우 드물기 때문에 일단 적절한 것으로 본다. 따라서 이 요약문은 HTML 문서 보다 자동 분류 측면에서 유용하다고 할 수 있다. 이상과 같은 두 가지 점에서 HTML 문서를 사용하지 않고 요약문을 본 논문의 시스템의 입력으로 사용한다.

- **[단계 2] 분야 후보어를 판정**

위의 [단계 1]에서 수집한 문서의 1건에 대하여 문서 분류 모듈을 실행하여 분야 후보어를 판정한다. 또한 분야 후보 각각의 판정 건수도 계산한다. 그 결과를 아래의 <표 5>에 제시하였다.

<표 5> 문서 1건 마다의 분야 판정 결과

분야	판정 건수
</정치>	220
</법률>	130
</해외-국제>	121
</사전>	120
</선거>	120
</포플리즘>	30
</대통령 후보>	30
</분야 미정>	1

- **[단계 3] 분야를 판정**

[단계 2]에서 판정된 분야 후보어 중에서 판정 건수가 가장 많은 것을 분야연상어 후보어의 “후보 분야(candidate field)”라 판정한다. 위의 <표 5>를 살펴보면 </정치> 분야의 판정 건수가 가장 높기 때문에 후보 분야를 </정치>라고 일단 추정하였다.

2.6 분야연상어 사전의 구축

2.6.1 자동 구축 방법

분야연상어 구축에는 분야 체계의 종단 분야마다 수집된 대규모의 문서를 학습 데이터로 이용한다. 학습 데이터에 대하여 형태소 해석을 하여 명사의 출현 빈도를 각각의 분야마다 계산한다. 그 후에 다음의 2.6.2절에서 설명한 바와 같이 각 명사가 어느 분야에 집중적으로 나타나는지를 계산하여 단일 분야연상어 및 그 수준이 결정된다.

2.6.2 단일 분야연상어 수준의 결정 알고리즘

단일 분야연상어 결정 알고리즘에서는 학습 데이터로부터 구해진 단어의 빈도 정보를 이용한다. 그러나 각 종단 분야에 학습 데이터를 균일하게 수집하는 것은 매우 어려운 작업이다. 따라서 종단 분야 <S>에 해당하는 단어의 빈도는 종단 분야 <S>에 출현한 모든 단어의 빈도를 T(<S>)로 정의하여 단어의 분야 <S>의 빈도를 F(w, <S>)로 표기한다. 다음과 같은 정규화 된 값인 R(w, <S>)를 식 (2)와 같이 사용한다.

$$R(w, \ <S>) = \frac{F(w, \ <s>)}{T(<s>)} \times \gamma \quad \cdots\cdots \ (2)$$

여기서 $\frac{F(w, \ <s>)}{T(<s>)}$ 는 굉장히 작은 값이므로 적절한 상수에 의해 $R(w, <S>)$를 정수로 조정한다. 또한 중간 분야 $<S'>$에 대한 단어 w의 빈도 $R(w, <S'>)$는 $<S'>$의 하위에 존재하는 모든 종단 분야 <S>의 빈도 $R(w, <S>)$를 합산한 것으로 한다. 분야 $<S'>$ 을 분야 <S>의 부모 분야라 할 때 분야 <S>에 관한 단어의 집중률은 다음의 식(3)과 같이 정의한다.

$$P(w, \ <S>) = \frac{R(w, \ <s>)}{R(<s'>)} \quad \cdots\cdots \ (3)$$

[단일 분야연상어의 결정 알고리즘]

- **입력 : 단어 w, 각 분야 <S>에 대하여 R(w, <S>), 분야 트리**
- **출력 : w가 분야연상어가 된다면, 연상하는 분야와 수준**

- **[단계 1] 완전 분야연상어(수준 1)의 결정**

임계값 α를 선정하여 분야 체계의 루트 $<S> = <$전체 분야$>$의 자식 분야 <S/C>에 대하여 단어 w가 특정한 분야에 집중하는지 혹은 집중하지 않는지를 다음의 식 (4)로 판별하고 이 조건을 만족하면

$$P(w, \ <S/C>) \geq \alpha \quad \cdots\cdots \ (4)$$

<S/C>를 <S>로 바꿔서 다시 하위의 자식 분야에서 같은 방법으로 판정을 반복한다. 이 처리를 반복함으로써 <S/C>가 종단 분야가 되면 w를 분야 <S/C>의 완전 분야연상어로 한다. 이 처리로 조건식을 만족하는 <S>의 자식 분야 <S/C>가 존재하지 않는 경우는 다음의 [단계 2]로 진행한다.

- **[단계 2] 준완전 분야연상어(수준 2)와 중간 분야연상어(수준 3)의 결정**

분야 <S>의 $m \geq 2$ 개의 자식 분야 <S/C>로부터

$$P(w, \ <S/C>) \geq o\frac{R(w, \ <s>)}{m} \quad \cdot \cdot \cdot \cdot \ (4)$$

이 되는 <S/C>를 추출하여 $P(w, \ <S>)$의 큰 순서로 누적 가산하여 $k(1 < k < m)$ 개를 가산하여 첫 번째 합계 값이 α를 넘는 경우 k개의 자식 분야 <S/C>가 모두 종단 분야이면 w를 분야 <S/C>의 준완전 분야연상어로 결정한다. 모든 것이 종단 분야가 아니면 다음으로 진행한다. 단 누적 가산한 값이 α 를 넘지 않으면 w를 분야 <S>의 중간 분야연상어로 결정한다.

- **[단계 3] 다분야연상어(수준 4)의 결정**

k개의 자식 분야 <S/C>로부터 종단 분야 <S/C>를 추출하여 w를 분야 <S/C>의 다분야연상어로 한다. 종단 분야 이외의 자식 분야 <S/C>를 부분 분야 트리의 뿌리 <S>로 바꾸고 [단계 1]과 [단계 2]에서 결정된 완전, 준, 중간 분야연상어의 분야에 대하여 w를 다분야연상어로 한다.

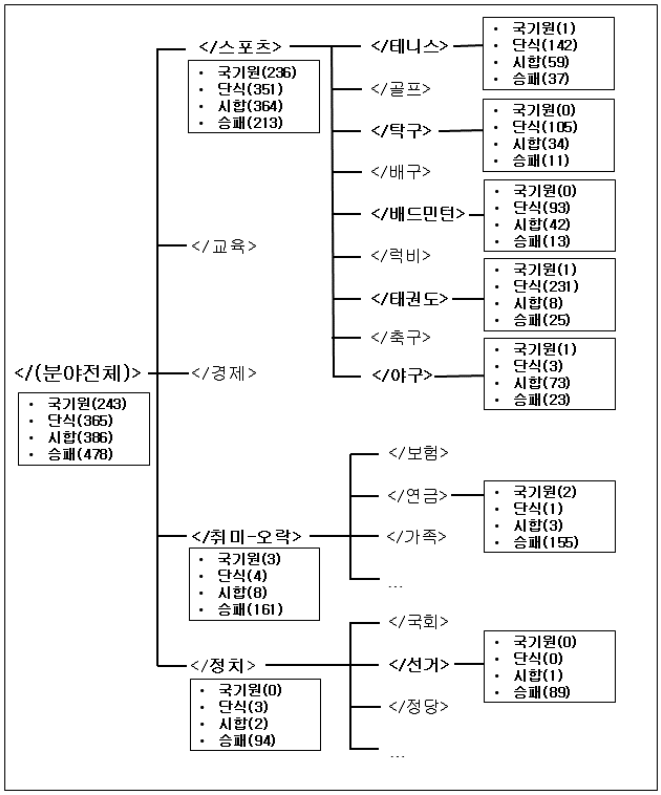
(알고리즘 끝)

단일 분야연상어 결정 알고리즘을 사용하여 분야연상어가 결정되는 예를 나타낸다. (그림 7)은 “국기원”, “단식”, “시합”, “승패”의 각 분야에서의 빈도를 괄호 안에 나타내었다. 또한 <분야 전체>의 자식 분야 수는 12, <스포츠>, <취미-오락>, <정치>의 자식 분야 수는 각각 19, 13, 14이며, 기준치 $\alpha = 0.9$로 하였다.

3. 결론

본 논문에서는 분야연상어를 정의하고, 단일어에 대한 분야연상어 정보를 이용하여 일상생활에서 끊임없이 생성되는 복합 분야연상어를 효율적으로 결정하는 방법으로 확장한다[1].

분야연상어를 단일과 복합 분야연상어로 분류하여 단일 분야연상어를 형태소사전에 등록된 표제어와 일치하도록 한정하였다. 이것은 단일 분야연상어의 분야정보를 형태소 사전에 그대로 등록하기 위한 실용성을 고려한 것이다. 또한, 본 연구에서는 분야체계를 미리 정의한다고 하였으나 분야연상어 구축은 어떠한 분야체계에도 손쉽게 적용될 수 있으므로 보편성은 충분하다. 덧붙여, 본 연구의 결과는 문서 분류를 위한 지식베이스로 이용할 수 있는데, 몇몇 단어를 분야연상어로 잘못 인식하거나 혹은 분야연상어를 일반 단어로 오인식하는 문제가 발생하여도 전체 시스템에는 영향을 주지 않으므로 결함 허용(fault tolerance) 능력이 있다.



(그림 5) 분야연상어의 수준 결정 예

참고문헌

[1] 이상곤, "한글 문서분류용으로 이용할 복합어로 구성된 분야연상어의 추출법", 정보과학회논문지: 소프트웨어 및 응용, 제32권, 제7호, pp. 636-649, 2005.

[2] 이상곤, "분야연상어를 이용한 화제분야의 계산방법과 단락검색", 정보처리학회논문지(B), 제12권, 제1호, pp. 57-68, 2005.

[3] 이상곤, 이완권, "분야연상어의 수집과 추출 알고리즘", 정보처리학회 논문지(B), 제10권, 제3호, pp. 347-358, 2003.

[4] 이상곤, "분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할방법", 정보처리학회 논문지(B), 제10권, 제1호, pp. 57-66, 2003.

[5] 장정효, 손주성, 이상곤, 안동언, "연상 지식을 이용한 문서 분류 엔진의 구현", 제25회 정보처리학회 춘계 학술발표대회 논문집, 제13권, 제1호, pp. 625-628, 2006.

[6] 이원휘, 김도연, 이상곤, "그래픽컬한 분야인식기의 설계 및 구현", 정보과학회 가을 학술발표 논문집, 제31권, 제2호, pp. 769-771, 2004.

[7] 이원휘, 최현, 이상곤, "분야연상어 추출 방법의 설계와 구현", 정보처리학회 2004년도 춘계 학술발표 논문집, 제11권, 제1호, pp. 651-654, 2004.