분야연상어 추출 방법의 설계 및 구현

이원휘[1], 최현[2], 이상곤[1]
전주대학교 정보기술공학부[1]
전주대학교 교육대학원 컴퓨터교육전공[2]
e-mail : {wony, kaiana, samuel}@jj.ac.kr

Design and Implementation for Extraction of Field-Associationed Terms

Won-Hee Lee[1], Hyun Choi[2] and Samuel Sangkon Lee[1]

요 약

우리는 특정 문서를 읽을 때 문서 전체를 읽지 않더라도 대표적인 몇 개의 단어를 보는 것만으로 정치나 경제, 스포츠 등의 분야를 정확히 인지할 수 있다. 문서 전체를 대상으로 하지 않고 부분텍스트에서 출현하는 소수의 단어정보에서 문서의 분야를 정확히 결정하기 위해 분야연상어의 구축은 중요한 연구과제이다. 인간이 미리 분야체계를 정의하고, 각 분야에 해당하는 문서를 인터넷이나 서적을 통해 수집한다. 본 논문은 수집문서의 분야를 정확히 지시하는 분야연상어를 자동으로 수집하는 시스템을 설계하고 구현하는데 목적이 있다. 문서의 분야결정 시점을 고려하여 분야연상어의 수준, 안정성 랭크, 집중률, 빈도정보를 이용하여 단일 분야연상어를 수집하는 방법을 제안하고 구현한다.

1. 서론

최근 컴퓨터의 이용증가에 따른 문서의 전자화에 맞추어 문서의 자동 분류에 관한 연구개발이 대단히 활발하다. 인간은 문서 전체를 읽지 아니하여도, 문서에서 대표적인 단어를 보는 것만으로 <정치>나 <스포츠> 등의 문서분야를 정확히 인지할 수 있다[4]. 한편, 인간은 자신의 상식지식으로 특정분야를 인지할 수 없는 경우에도 문서에서 처음으로 출현하는 몇 개의 단어를 이용하여 연상되는 정보를 인지적으로 인식하고, 문서의 내용을 읽어감에 따라 문서에 해당하는 분야를 연상하거나 추측할 수 있다. 또한 문서의 이전내용에서 애매성이 발생하여도 문서의 뒤에서 출현하는 단어에 의해 이전 문서내용의 애매성을 해소해 나갈 수 있다.

이처럼 몇 개의 단어 정보를 이용하여 문서가 포함되는 분야를 정확히 결정할 수 있는 단어를 참고문헌 [4]에서 정의한 “분야연상어”를 이용하여 분야연상어의 자동적인 추출 시스템을 설계하고 구현한다.

2. 분야연상어

2.1 단일과 복합 분야연상어

단어를 더 이상 분할이 불가능한 의미를 가진 최소단위라 할 때, 형태소 사전에 등록되어 있는 단어를 “단일어”라 하며, 두 단어 이상의 단일어로 구성된 단어를 “복합어”라 부른다. 이들 단일어와 복합어로 구성된 분야연상어를 각각 단일 분야연상어와 복합 분야연상어[4]라 한다.

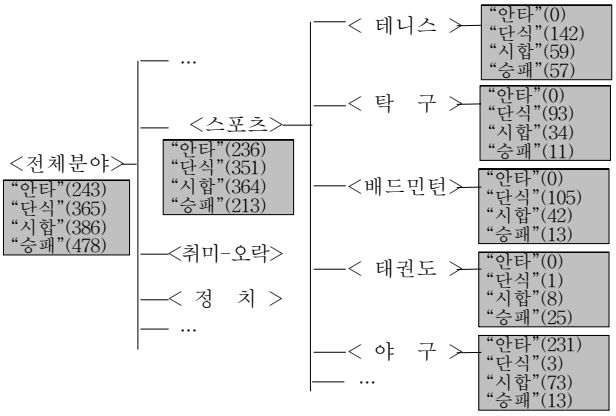
한 개의 접사와 명사로 구성되는 일반적인 복합어 “도시화”, “연습실”, “부가세” 등은 세분화함으로써 분야정보를 잃어버리기 쉽기 때문에 단일어로 취급한다. 또는 학교명이나 회사명 등의 고유명사는 단일어로 취급되어야 한다.

2.2 분야트리

본 논문에서 사용되는 분야트리는 참고문헌[1, 4]를 참조하여 분야트리의 단말노드를 “종단분야”, 그 이외의 모든 노드를 “중간분야”라 지칭한다. 또한 “부모분야”와 “자식분야”로 상위분야와 하위분야를 지칭한다. 분야의 지정은 <path>로 기술하나 모순이 생기지 않는 한 <전체분야>는 생략하고, 종단분야만을 기술한다. 예를 들어 <path>=<스포츠/축구>는 <스포츠>의 하위 종단분야 <축구>를 표시한다.

2.3 분야연상어의 수준별 랭크

분야연상어는 분야를 결정할 때 여러 개 존재할 수 있고, 연상되는 범위는 단일, 복수 혹은 상위, 하위 분야를 한정하는 등 여러 경우가 존재할 수 있다. 따라서 분야체계 내에서 연상되는 분야의 범위에 제약을 가하여 분야연상어의 수준을 다음과 같이 정의한다.



[그림 1] 분야트리와 분야연상어의 예

[정의] 분야연상어의 수준

수준 1 : 완전 분야연상어 : 분야연상어후보 w는 유일한 종단분야만을 연상한다.

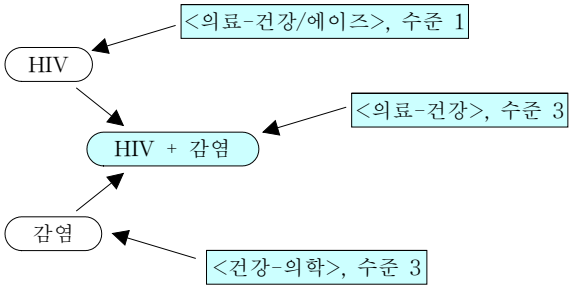
수준 2 : 준완전 분야연상어 : 분야연상어후보 w는 같은 부모분야를 갖는 종단분야 중에서 한정된 복수 개의 종단분야만을 연상한다.

수준 3 : 중간 분야연상어 : 분야연상어후보 w는 완전 분야연상어, 준완전 분야연상어가 아니고, 하나의 중간분야를 연상한다.

수준 4 : 다분야연상어 : 분야연상어후보 w는 완전 분야연상어, 준완전 분야연상어, 중간 분야연상어가 아니고, 다수의 중간분야와 다수의 종단분야를 연상한다.

수준 5 : 비연상어 : 분야연상어후보 w는 위의 수준 1~4 이외이고, 어떠한 특정분야도 연상하지 않는다.

예를 들어, "HIV감염"은 <건강-의료/병명/에이즈>의 수준 1의 완전 분야연상어이지만, 이 복합어의 구성 중 "감염"는 <건강-의학> 분야의 수준 3의 중간 분야연상어이고, 복합어 "HIV감염"은 분야 <에이즈>의 수준 1의 완전 분야연상어 "HIV"를 구성어로 포함하고 있다. 이와 같이 복합어의 각 구성어가 동일한 분야를 중복하여 포함하고 있는 단어를 "분야 중복 연상어"라 부른다.



[그림 2] HIV감염의 예

3. 분야연상어의 결정

본 장에서는 학습데이터에서 분야연상어의 후보와 그 수준을 자동으로 결정하는 방법과 알고리즘에 대하여 언급한다. 학습데이터는 각 종단분야에 대하여 균일하게 수집되었다고 볼 수 없기 때문에 전체 분야에 대하여 특정한 분야연상어의 합계빈도를 계산하여 각 종단분야에 출현하는 단어빈도로 정규화 한 값을 이용한다. 이 정규화된 빈도정보에 의해 특정 단어가 특정분야에 집중하는 집중률을 정의한다.

본 논문의 방법은 한 개의 분야연상어 후보에 대하여 분야트리의 상위에서 하위로 한번 탐색하여 그 분야연상어가 지시하는 분야를 빠르고 정확하게 결정할 수 있다. 분야트리의 루트에서 기준율 이상으로 집중하는 특정분야를 집중적으로 탐색하여 그 조작을 하위분야로 진행한다. 조작이 종단분야에 도달하면 수준 1의 완전 분야연상어와 정확한 연상분야를 결정하고, 기준율 이하의 분야에서 출현하는 분야연상어이면 수준 2, 수준 3 혹은 수준 4의 분야연상어 탐색으로 알고리즘의 제어를 변경한다.

3.1 수준의 결정 알고리즘

본 논문에서 사용되는 알고리즘은 참고문헌[4]에서 제안한 알고리즘을 이용한다. 따라서 본 논문에서는 자세한 설명은 제외하고 알고리즘에서 이용되는 수식만을 언급한다.

- 정규화된 빈도

$$Normalization(w,\langle T\rangle) = \left\{ \frac{Frequency(w,\langle T\rangle)}{Total\ \ Frequency(w)} \right\} \times r \quad \cdots\cdots \quad (1)$$

- 집중률

$$Concentration(w,\langle C\rangle) = \frac{Normalization(w,\langle C\rangle)}{Normalization(w,\langle P\rangle)} \quad \cdots\cdots\cdots (2)$$

- 완전분야연상어(수준 1)

$$Concentration(w,\langle P\rangle) \geq a \quad \cdots\cdots\cdots\cdots \quad (3)$$

- 준완전(수준 2) 혹은 중간 분야연상어(수준 3)

$$Concentration(w,\langle P\rangle) \geq \frac{Normalization(w,\langle C\rangle)}{m} \quad \cdots\cdots \quad (4)$$

- 다분야연상어(수준 4)

k개의 자식분야에서 종단분야 <P/C>를 추출하고, w를 분야 <P/C>의 다분야연상어로 결정한다. 종단분야를 제외한 자식분야 <P/C>를 부분트리의 루트 <P>로 수정하여 수식 (3)과 수식 (4)를 다시 실행하면 복수 개의 중간분야와 종단분야가 얻어진다. w를 분야 <P>의 "다분야연상어"로 결정한다. 참고문헌[5]의 알고리즘을 이용하여 실제 추출 예를 보자

3.2 수준 결정의 예제

이 절에서는 분야연상어 "안타", "단식", "시합", "승패" 등의 수준별 결정 예를 설명한다. 분야트리에서 출현하는 각각의 분야연상어 빈도수를 [그림 1]의 괄호 안에 표시하였다. <분야전체>의 자식분야 수는 13, 이 중 <스포츠> 분야의 자식분야 수를 19로 제한하며, 임계값 α는 수 십 차례의 실험을 통해 얻은 값 0.92를 사용한다.

[순서 1] 완전분야연상어 추출 예

"안타", <P>=<분야전체>에 대하여, 단어 "안타"가 가장 많이 출현하는 분야 <스포츠>를 선택하여 <C>=<스포츠>에 대한 집중율은 다음과 같이 계산한다.

$$Concentration(\text{“안타”}, \langle \text{스포츠} \rangle) = \frac{Normalization(\text{“안타”}, \langle \text{스포츠} \rangle)}{Normalization(\text{“안타”}, \langle \text{분야전체} \rangle)} = \frac{236}{243} \fallingdotseq 0.97(\geq a)$$

위의 식에 의해 “안타”는 분야 <P/C>=<분야전체/스포츠>에 집중한다. 다음의 <스포츠> 분야에서 “안타”의 빈도가 가장 높은 분야 <야구>를 선정하여, <P>를 <스포츠/야구>로 고쳐서 하위의 분야 <C>=<야구>에 대하여 판정하면,

$$Concentration(\text{“안타”}, \langle \text{야구} \rangle) = \frac{Normalization(\text{“안타”}, \langle \text{야구} \rangle)}{Normalization(\text{“안타”}, \langle \text{스포츠} \rangle)} = \frac{231}{236} \fallingdotseq 0.98(\geq a)$$

가 되고, 현재의 분야 <야구>가 종단분야이므로 “안타”는 <야구> 분야에 대한 완전분야연상어(수준 1)로 결정된다.

[순서 2] 준완전 분야연상어 추출 예

“단식”은 <P/C>=<전체분야/스포츠>에서 다음의 집중률이 얻어진다.

$$Concentration(\text{“단식”}, \langle \text{스포츠} \rangle) = \frac{Normalization(\text{“단식”}, \langle \text{스포츠} \rangle)}{Normalization(\text{“단식”}, \langle \text{분야전체} \rangle)} = \frac{351}{365} \fallingdotseq 0.96(\geq a)$$

그러나 분야 <스포츠>의 하위분야 중에서 임계값(α) 이상의 값으로 집중하는 분야가 다음의 계산에서 보는 바와 같이 존재하지 않는다.

$$Concentration(\text{“단식”}, \langle \text{테니스|배드민턴|탁구|야구|태권도} \rangle) = \frac{Normalization(\text{“단식”}, \langle \text{테니스|배드민턴|탁구|야구|태권도} \rangle)}{Normalization(\text{“단식”}, \langle \text{스포츠} \rangle)} = \frac{142|105|93|2|1}{351} \fallingdotseq 0.405|0.299|0.265|0.009|0.003$$

따라서 앞의 알고리즘의 준완전 혹은 중간분야연상어 수식을 이용하여

$$\frac{Normalization(\text{단식}, \langle \text{스포츠} \rangle)}{m} = \frac{351}{19} \fallingdotseq 18.74$$

를 구할 수 있다.

<스포츠>의 하위분야 중에서 이 수치보다 빈도가 높은 자식분야는 빈도 142, 105, 93을 갖는 각각의 분야 <테니스>, <탁구>, <배드민턴>이며, 집중률 Concentration (w, <P/C>)을 각각 계산하면,

$$Concentration(\text{“단식”}, \langle \text{테니스} \rangle) = \frac{142}{351} \fallingdotseq 0.41$$

$$Concentration(\text{“단식”}, \langle \text{배드민턴} \rangle) = \frac{105}{351} \fallingdotseq 0.30$$

$$Concentration(\text{“단식”}, \langle \text{탁구} \rangle) = \frac{93}{351} \fallingdotseq 0.27$$

가 되고, 이들 후보 중 집중률의 합이 임계값(0.92)을 초과할 때까지 집중률이 큰 자식분야 순으로 빈도율을 더하면 “단식”은 분야 <스포츠>의 세 개의 자식분야 <C> =<테니스>, <배드민턴>, <탁구>에 집중하여 출현하고, 이들은 모두 종단분야이므로 “단식”은 준완전 분야연상어(수준 2)가 된다.

[순서 3] 중간 분야연상어 추출 예

순서 2와 비슷하게 “시합”의 경우, <P>=<전체분야>의 유일한 하위분야 <C>=<스포츠>에 집중하지만, 그 하위의 분야에 집중하는 분야는 존재하지 않는다. 따라서 “단식”의 경우처럼

$$\frac{Normalization(\text{“시합”}, \langle \text{스포츠} \rangle)}{m} = \frac{364}{19} \fallingdotseq 19.16$$

이므로 이 수치보다 높은 빈도를 갖는 자식분야 <테니스>, <탁구>, <배드민턴>, <야구>의 빈도율을 구하면 다음과 같다.

$$Concentration(\text{“시합”}, \langle \text{테니스|배드민턴|탁구|야구} \rangle) = \frac{Normalization(\text{“시합”}, \langle \text{테니스|배드민턴|탁구|야구} \rangle)}{Normalization(\text{“시합”}, \langle \text{스포츠} \rangle)} = \frac{59|42|34|73}{364} \fallingdotseq 0.16|0.12|0.09|0.2$$

이들의 모든 수치를 가산하여도 0.57이 되어 α(0.92)를 초과하지 않는다. 따라서 네 개의 자식분야 중 어떤 분야에도 집중적으로 출현하지 않는다고 판정하고, “시합”은 중간 분야연상어(수준 3)로 결정된다.

다른 예로, “승패”는 <P>=<분야전체>의 어느 하위분야에도 기준치 α(=0.92) 이상으로 집중되지 않으므로 수식(4)에 의해 복수분야가 결정된다. <전체분야>의 자식분야 수 m은 13이라고 제한하였으므로 수식(4)에 의해 다음과 같이 계산된다.

$$\frac{Normalization(\text{“승패”}, \langle \text{전체분야} \rangle)}{m} = \frac{478}{13} \fallingdotseq 36.77$$

이 값 이상의 빈도를 갖는 자식분야는 <스포츠>(213), <취미-오락>(161), <정치>(94)라 할 때, 이들 각각의 집중률은 다음과 같이 얻을 수 있다.

$$Concentration(\text{“승패”}, \langle \text{스포츠} \rangle) = \frac{213}{478} \fallingdotseq 0.45$$

$$Concentration(\text{“승패”}, \langle \text{취미}-\text{오락} \rangle) = \frac{161}{478} \fallingdotseq 0.34$$

$$Concentration(\text{“승패”}, \langle \text{정치} \rangle) = \frac{94}{478} \fallingdotseq 0.19$$

여기서, 이들 세 분야의 누적 가산치 0.98(≒0.45+0.34+0.19)은 기준 집중률 α를 초과하지만, 분야 <P/C>=<스포츠>, <취미-오락>, <정치>는 모두 종단분야가 아니다. 따라서, “승패”는 수준 2의 준완전 분야연상어가 될 수 없으며, <C>=<스포츠>, <취미-오락>, <정치> 등을 부분트리로 하여 수식(3)과 수식(4)를 반복 실행하면 “승패”는 분야 <스포츠>에 집중하지만, 그 하위의 자식분야에는 집중하지 않는다. 분야 <취미-오락>, <정치>에 대해서는 그 자식분야 <취미-오락/장기>, <정치/선거>에 집중함을 알 수 있다. 따라서, “승패”는 중간분야 <스포츠>와 종단분야 <취미-오락/장기>, <정치/선거> 등에 해당하는 다분야연상어(수준 4)가 된다.

4. 분야연상어 추출 시스템의 설계 및 구현

본 장에서는 3장에서 논의된 알고리즘을 기반으로 실제 빈도수를 추출하고 이 빈도수에 기반 한 빈도율 및 분야연상어의 수준을 구한다.

4.1 구현환경

분야연상어 추출 시스템은 다음의 [표 1]에서 예시한

환경하에서 구현되었다.

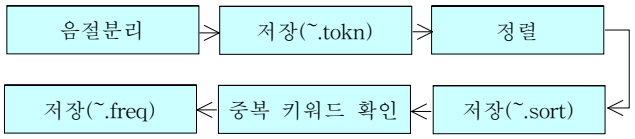
4.2 구현방법

분야연상어 자동 추출 시스템은 JAVA를 이용하여 프

CPU : PentiumⅣ 2.6 GHz 메모리 : 256MB 언어 : JAVA 데이터 구조 : 디렉토리 구조 파일 구성 : *.txt : 원본파일 *.tokn :음절분리 후 저장파일 *.sort : tokn파일을 정렬 후 저장파일 *.freq : 빈도수추출 후 저장파일

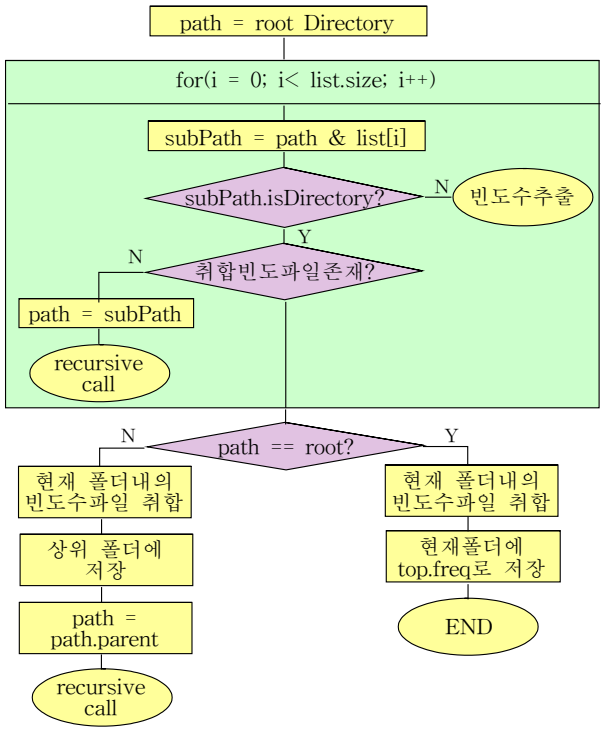
[표 1] 구현환경

로그래밍하였다. 본 시스템의 핵심부분은 빈도수 추출과 추출된 빈도수를 취합하여 상위분야로 보내는 것이다. 즉, 종단분야 <야구>에는 여러 개의 파일이 존재하게 된다. 이들 파일을 각각 빈도수를 계산한다.



[그림 3] 빈도수 추출에 대한 흐름도

각 빈도수파일은 빈도수파일 집계 모듈에 의하여 <야구>폴더 내에 있는 모든 빈도수 파일들의 빈도수를 모아 <야구>폴더의 부모분야, 즉 <스포츠>의 하위에 "tot_야구.freq"라는 파일로 저장되게 된다. <스포츠> 폴더 내에 있는 파일들 또한 빈도수를 추출한 후, <스포츠>의 상위 폴더에 "tot_스포츠.freq"라는 파일로 저장된다. 이렇게 반복수행하여 최종적으로 분야트리의 루트(root)에 이르면 루트내의 빈도수 파일들을 집계하여 "top.freq"라는 파일을 생성하게 된다.



[그림 4] 빈도수 집계의 흐름도

[그림 5] 빈도수 추출 후의 실행화면

[그림 6] 빈도수 집계 후의 실행화면

5 결론

본 논문에서는 분야연상어를 정의하고, 주어진 문서에서 빈도수를 추출하였다. 또한 이 빈도수를 이용하여 특정 키워드가 어떤 분야를 연상할 수 있는지를 자동 판단하여 평가하였다. 또한 본 연구에서는 분야체계를 미리 정의한다고 하였으나, 분야연상어 구축은 어떠한 분야체계에서도 손쉽게 적용될 수 있으므로 보편성은 충분하다고 생각된다[2, 3].

감사의 글

본 연구는 한국과학재단 목적기초연구(과제번호 : R05-2003-000-10690-0) 지원으로 수행되었습니다.

참고문헌

[1]남영신, 우리말 분류 사전, 성안당, 2001.발표 논문집(Ⅰ), 제 30권, 제 2호, pp. 544-546, 2003.

[2]이상곤, "분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할 방법", 정보처리학회 논문지B, 제 10권, 제 1호, pp. 57-66, 2003.

[3]홍성옥, 이상곤, "연상정보를 이용한 단락분할 방법", 2003년도 한국정보처리학회 춘계 학술발표 논문집(상), 제 10권, 제 1호, pp. 497-500, 2003.

[4]이상곤, 이완권, "분야연상어의 수집과 추출 알고리즘", 정보처리학회 논문지B, 제 10권, 제 3호, pp. 347-358, 2003.

[5]김숙영, 최창원, 이상곤, "한글문서분류용 분야연상어의 추출 알고리즘", 한국정보과학회 2003 가을 학술발표 논문집(Ⅰ), 제 30권, 제 2호, pp. 544-546, 2003.