한글문서 분류용 분야연상어의 추출 알고리즘

김 숙 영[1], 최 창 원[2], 이 상 곤[1]
전주대학교 정보산업대학원[1]
전주대학교 정보기술컴퓨터공학부[2]
이메일 : sykim@jtc.ac.kr, jaee123@hanmail.net, samuel@jeonju.ac.kr

Extraction Algorithm of Field-Associated Terms for Korean Document Classification

Sukyoung Kim, Changwon Choi, and Sangkon Lee
Dept. of Computer Science and Engineering,
Graduate School of Information and Industrial Engineering,
Jeonju University

인간은 문서에서 대표적인 단어를 보는 것만으로 정치나 스포츠 등의 분야를 정확히 인지할 수 있다. 문서전체를 대상으로 하지 않고 부분적인 텍스트에서 출현하는 소수의 단어정보에서 문서의 분야를 정확히 결정하기 위해 분야연상어의 구축은 중요한 연구과제이다. 인간이 미리 분야체계를 정의하고, 각 분야에 해당하는 문서를 인터넷이나 서적을 통해 수집하고, 수집문서의 분야를 정확히 지시하는 분야연상어를 수집하는 방법을 제안한다. 문서의 분야결정 시점을 고려하여 분야연상어의 수준과 안정성랭크에 대하여 논의한다. 학습데이터에서 분야연상어 후보의 각 수준을 자동으로 결정하고, 컴퓨터가 제시하는 분야연상어의 수준, 안정성랭크, 집중률, 빈도정보를 이용하여 단일어로 된 분야연상어를 추출하는 방법을 제안한다.

1. 서론

전자화된 문서 증가로 문서의 자동분류에 관한 연구개발이 활발한 가운데 문서전체의 정보를 이용하여 유사도를 계산하는 확률모델[1]이나 벡터모델[2] 등의 기술이 확립되어 있으나, 실제로 문서에는 복수의 화제나 분야가 혼합되어 있으며, 사용자가 검색을 원하는 내용은 문서 일부분에 존재하는 경우가 대부분이다. 따라서 문서 단편내의 소수의 단어정보를 이용하여 분야를 정확하게 결정하기 위한 분야연상어의 구축은 중요한 연구과제[4]이다.

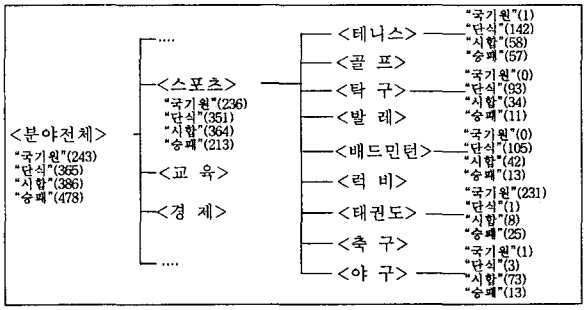
문서의 일부분에서 몇 개의 단어정보를 이용하여 문서가 포함되는 분야를 정확하게 결정할 수 있는 단어를 "분야연상어"라 정의하고, 상식적인 분야연상어의 구축, 유사문서(문장)검색, 문서요약 등의 기초연구를 수행한다. 본 논문에서 제안하는 방법은 문서의 처음부분에서 잘못된 분야연상어가 추출되어도 이후에 나타나는 올바른 분야연상어에 의해 정확한 문서의 분야를 추적할 수 있는 장점을 가진다.

인간이 결정한 분야체계와 수집된 학습데이터를 이용하여 분야연상어를 추출하는 알고리즘을 제시한다. 이 알고리즘을 이용하면 잘못된 분야연상어가 추출되는 비율이 몇 퍼센트 이하가 되는 추출 방법을 제안한다.

먼저, 제 2장에서는 분야연상어를 단일과 복합 분야연상어로 나누어 정의하고, 형태소 사전과의 관계를 설명한다. 제 3장에서는 학습데이터에서 분야연상어 후보와 수준을 자동적으로 결정하는 알고리즘을 제안한다. 4장에서 결론과 향후 연구과제에 대하여 논의한다.

2. 분야연상어

2.1 단일 분야연상어



(그림 1) 분야트리와 분야연상어의 예

더 이상 분할이 불가능한 의미를 가진 최소단위를 단어라 하고, 형태소 사전에 등록되어 있는 단어를 "단일어"라 부른다. 두 단어 이상의 단일어로 구성되는 단어를 "복합어"라 부른다. 이들을 (그림 1)에 표시한 바와 같이 기호 "과 "내에 기술한다. 단일어와 복합어로 구성된 분야연상어를 각각 단일 분야연상어와 복합 분야연상어라

기술한다.
단 미등록어는 분야연상어의 대상으로 하지 않는다.

2.2 분야트리

본 논문에서는 "분야체계(이후, "분야트리"라 부른다)를 사용한다. 분야트리의 단말노드에 상당하는 부분을 "종단분야"라 부르고, 종단분야 이외는 모두 "중간분야"라 부른다. 분야트리의 전체 분야수는 180개이며 중간 분야수는 22개, 종단분야 158개(깊이 2와 3의 종단분야는 각각 122개, 36개)이다. 어떤 분야와 직접 인접한 상위 혹은 하위분야를 각각 "부모분야"와 "자식분야"라 부른다. 분야의 지정은 분야명의 패스(<Path>)로 기술하나, 루트에 상당하는 <전체분야>는 생략하고 종단분야 만으로 기술하는 것을 원칙으로 한다.

2.3 분야연상어의 수준별 랭크([4])

<표 1> 분야연상어의 연상분야와 수준

분야연상어	연 상 분 야	수 준
국기원	<스포츠/태권도>	1
단식,복식	<스포츠/테니스> <스포츠/탁구>	2
시 합	<스포츠>	3
승 패	<취미-오락/장기> <정치/선거> <스포츠>	4
경우,사용	-	5

수준 1의 완전 분야연상어 "국기원"은 종단분야 <태권도>를 오직 하나의 뜻으로 한정한다. 수준 2의 준완전 분야연상어 "단식", "복식"은 부모분야 <스포츠>내의 복수의 종단분야 <테니스>, <탁구>, <배드민턴> 등을 한정한다. 수준 3의 중간 분야연상어 "시합"은 어떠한 종단분야도 한정하지 않으나, 한 개의 중간분야 <스포츠>를 한정한다. 수준 4의 다분야연상어 "승패"는 복수의 종단분야 <취미-오락/장기>, <정치/선거> 등과 중간분야 <스포츠>에 속하는 복수의 종단분야를 한정한다. 마지막으로, 수준 5의 비연상어는 "경우", "사용"과 같이 분야트리 내의 어떠한 특정분야도 한정하지 않는 단어를 비연상어로 정의(<표 1> 참조)한다.

2.4 분야연상어의 안정성랭크

안정성랭크(Stability Rank)는 랭크의 순위가 높은 순으로 보통명사를 a로, 인명이외의 고유명사를 b로, 인명에 해당하는 고유명사를 c로 할당한다. 다음 <표 2>에 수준 1의 분야연상어와 안정성랭크의 예를 표시하였다.

단일 분야연상어의 길이는 의미를 형성하는 가장 짧은 길이(최단길이)이며, 그 개수도 유한하기 때문에 분야연상어 후보를 사람의 상식지식을 이용하여 선별한다. 무한히 만들어지는 복합 분야연상어는 자동적으로 단어의 길이를 짧게 하고 그 수를 줄이는 방법이 필요하다.

<표 2> 분야연상어 후보와 안정성랭크

수 준	안정성랭크	분야연상어 후보	분 야
1	b	해태타이거스	<스포츠/야 구>
1	a	투 수	<스포츠/야 구>
1	b	자 이 언 트	<스포츠/야 구>
1	a	야 구	<스포츠/야 구>
1	a	홈 런	<스포츠/야 구>

3. 분야연상어의 결정

3.1 수준 결정 알고리즘

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 식 (1)과 같이 정규화 된 빈도 *Normalization*(w, <T>)를 정의한다.

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

여기서 *r*(정수) 값에 대하여 설명하면, 스포츠 분야 전체에서 출현하는 분야연상어 "야구"의 전체 합계빈도 *Total_Frequency*("야구")는 약 1,000 단어 정도 출현한다. 이것은 빈도가 높은 단어의 정규화 된 값이 매우 적음을 의미한다. 따라서 적당한 값 *r*을 계산하여 출현빈도를 조정한다. 이후에는 부록 A에서 제시하는 전체 분야트리에서 출현한 빈도정보를 이용하여 계산한 $r = 10^5$의 값으로 논의를 진행한다.

어떤 분야 <P>를 <C>의 부모분야라 할 때, 분야 <C>에 대한 분야연상어 w의 집중률 *Concentration*(w, <C>)는 중간분야 <P>에 대한 정규화 된 빈도 *Normalization*(w, <P>)를 <P>의 하위에 존재하는 종단분야 <C>의 *Normalization*(w, <C>)로 나눈 다음 식 (2)와 같이 정의한다.

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

다음에 분야연상어를 각 수준에 따라 적당한 후보로 결정하기 위한 알고리즘을 표시하였다.

- **분야연상어의 수준결정 알고리즘**

  입력 : ① 분야연상어의 후보어 w
  ② 분야 <C>를 연상한다고 추측되는 단어 w의 *Normalization*(w, <C>)값
  ③ 분야트리

  출력 : 연상되는 분야와 수준

(순서 A1) [완전 분야연상어의 결정]

분야트리의 루트 <P>에서 그 자식분야 <P/C>에 대하여 단어 w가 집중되어 있는가 혹은 그렇지 않은가를 다음의 조건식

$Concentration(w, <P>) \geq \alpha$ .....(3)

으로 판정하고, 조건식을 만족하면 <P>에 자식분야 <C>를 연결하여 <P/C>로 바꾸고, 다시 하위의 자식분야에 대해 동일한 판정을 수행한다. α는 0.5보다 큰 값을 채택하였기 때문에 이 조건식을 만족하는 분야연상어 w는 한 개 미만이다. 반복처리 결과 <P/C>의 종단분야가 되면, w를 분야 <P/C>의 "완전 분야연상어(수준 1)"로 결정한다. 만약 조건식을 만족하는 <P>의 자식분야 <P/C>가 존재하지 않으면 다음으로 진행하여 수준 2, 3, 4의 분야연상어 후보 판정을 진행한다.

(순서 A2) [준완전 혹은 중간 분야연상어의 결정]

분야연상어 w가 수준 1로 결정되지 않은 경우에 분야 <P>는 종단분야까지 도달하지 않았다는 것을 의미한다. 따라서 이 분야 <P>는 반드시 중간분야이며, 최소한 m(≥2)개의 자식분야들을 가지고 있다. 수준 2는 <P>의 복수 개의 자식분야에 집중하는 분야를 탐색하는 것이 목적이므로 집중률의 총합이 α이상이 되는 자식분야를 복수 개 가진다. 단, 복수의 자식분야 수를 분야 <P>의 모든 자식분야의 수 m에 가깝게 하면 수준 3의 중간 분야연상어의 빈도는 자매(혹은 형제) 분야 중 평균빈도 이상의 분야를 선정한다.

$$Concentration(w, <P>) \geq \left\{ \frac{Normalization(w, <C>)}{m} \right\} \quad .....(4)$$

식 (4)를 만족하는 자식분야 <P/C>를 추출하고, $Concentration(w, <P/C>)$의 값이 큰 순서부터 누적가산하고, 누적 가산치가 최초에 α를 넘으면, k(1 < m)개의 자식분야를 결정한다. 이 때 k개의 자식분야 <P/C>가 모두 종단분야이면, w를 분야 <P/C>의 "준완전 분야연상어(수준 2)"로 결정한다. 모두 종단분야가 아니면, 다음 순서로 진행하여 다분야연상어 판정을 수행한다. 차례차례 누적 가산한 누적치가 α를 초과하지 않으면, w를 <P>의 "중간 분야연상어(수준 3)"로 결정한다.

(순서 A3) [다분야연상어의 결정]

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 (순서 A1)과 (순서 A2)를 다시 실행하면 복수 개의 중간분야와 종단분야가 얻어진다. w를 분야 <P>의 "다분야연상어(수준 4)"로 결정한다.

이상의 알고리즘을 통하여 수준1에 해당하는 <스포츠/야구> 분야의 분야연상어는 <표 2>와 같이 추출되었고, <표 3>은 <스포츠> 분야의 수준 2와 3을 나타내며, 안정성과 연상되는 분야를 출력한 실험 결과를 나타낸다. 실제 예는 논문 발표 시 제시를 할 계획이다.

<표 3> <스포츠>의 수준별 단일 분야연상어의 예

수준	안정성	분야연상어후보	분야A	분야B	분야C
2	a	선 발	<스포츠/야구>	<스포츠/축구>	-
2	a	선제공격	<스포츠/야구>	<스포츠/축구>	<스포츠/럭비>
2	a	승 점	<스포츠/야구>	<스포츠/축구>	<스포츠/요트>
2	c	김 병 연	<스포츠/야구>	<스포츠/축구>	-
3	a	시 합	<스포츠/야구>	<스포츠/축구>	<스포츠/배구>
3	a	리딩히터	<스포츠/축구>	<스포츠/축구>	<스포츠/요트>
3	b	바르셀로나	<스포츠/야구>	<스포츠/검도>	<스포츠/육상>
3	a	선 수	<스포츠/축구>	<스포츠/야구>	<스포츠/배구>

4. 결론

본 논문에서는 분야연상어를 정의하고 단일어에 대한 분야연상어 정보를 이용하여 일상생활에서 끊임없이 생성되는 복합 분야연상어를 효율적으로 결정하는 방법을 제안하여 180분야의 학습데이터를 토대로 그 유효성을 평가하였다.

학습데이터에서 분야연상어의 후보와 그 수준을 자동적으로 결정하는 알고리즘을 제안하였다. 학습데이터의 불균형성에 대해서는 상대빈도를 이용하여 빈도를 정규화하고, 분야연상어가 특정분야에 집중하는 기준율 α를 정의하였다.

감사의 글

이 논문은 2003년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2003-003-D00415). 재단의 연구비 지원에 감사 드립니다.

참 고 문 헌

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

[2] 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), p235, 1999.

[3] 남영신, 우리말 분류 사전, 성안당, 2001.

[4] 이상곤, "분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할 방법", 정보처리학회논문지B, 제10권 제1호, pp.57-66, 2003.

[5] 이상곤, 이완권, "분야연상어의 수집과 추출 알고리즘", 정보처리학회논문지B, 제10권, 제3호, p.347-358, 2003.