제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

연상정보를 이용한 단락분할 방법

홍 성 옥, 이 상 곤
전주대학교 교육대학원 컴퓨터교육전공
이메일 : {mydream, samuel}@jeonju.ac.kr

A Passage Retrieval Method by Using Field-Associated Information

Sung-Og Hong and Samuel Sangkon Lee
Dept. of Computer Education,
Graduate School of Education,
Jeonju University

요 약

문서에 여러가지 화제가 혼합되어 있는 문서에서 화제의 실마리 부분을 특정화하여 각 화제별 단락을 추출하는 기술은 정보검색 분야에서 중요한 역할을 담당하는 기술이다. 잘 정의된 분야체계에 따라 구축된 분야연상어를 이용하여 단락분할을 시도한다. 분야연상어는 특정한 분야를 정확하게 연상할 수 있는 단어로서 잘 분류된 문서 컬렉션에서 구축할 수 있다. 이 분야연상어를 이용하여 문서를 관련된 분야별로 추출하여 의미기반 단락추출 방법을 제안한다. 화제의 계속성에 주목하여 분야연상어의 수준(범위)이나 연속출현성에 의해 계산된 계속도에 의해 화제의 실마리를 추적하고, 화제의 전환성을 고려한 방법을 제안한다. 문서 내 각 화제의 단락 구분을 명확히 하여, 단락을 화제분야별로 추출하는 방법을 제안한다. 50문서를 실험한 결과 82%의 정확율과 63%의 재현율을 얻어 실용성을 기대할 수 있다.

1. 서론

종래의 정보검색 방법은 문서전체를 하나의 객체로 생각하여 검색요구에 적당한 문서를 검색하여 왔다. 대량의 문서를 특정한 기준에 따라 자동으로 분류하는 연구분야는 각 단락간의 유사도를 계산하여 유사도가 높은 순으로 문서를 분류하는 단락별 분류 방법이 사용되고 있다. 또한 문서에서 특정한 정보를 추출하는 문서추출 분야에서도 문서의 화제분야 추정은 유용한 실마리가 된다.

본 논문에서는 분야연상어[4, 6]를 이용하여 검색요구에 일치하는 단락 추출을 하고, 범위에 해당하는 단락의 분야를 결정하는 방법을 제안한다. 분야연상어가 나타나는 텍스트 주변부분은 특정화제를 묘사하는 화제분야로 추정할 수 있다. 그러나, 분야연상어가 나타나지 않는 단락에 대해서는 문서 내 화제를 파악하고, 화제의 계속도를 계산하고 동시에, 화제의 전환성을 고려하여 문장 간의 구간분리를 명확히 하여 분야의 중복이 없는 단락을 추출한다.

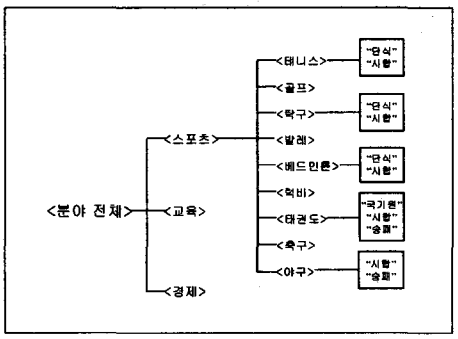
제 2장에서는 미리 정의된 분야체계에 따라 각 분야를 지시하는 분야연상어에 대하여 설명한다. 제 3장에서는 일반적인 문서에서 화제흐름의 특징을 정의하고, 문서 내에서 동일분야의 단락을 결정하는 방법을 설명한다. 제 4장에서는 예제 문서를 사용하여 실험을 하고, 본 방법의 유용성을 평가한다.

2. 분야연상어

2.1 분야체계

분야체계란 각 분야의 상위·하위관계를 트리구조로 표현한 분야별 체계를 말한다. 이를 "분야트리"라 정의하고, 분야트리의 잎에 상응하는 분야를 "종단분야", 종단분야 이외는 "중간분야"라 부른다. 본 연구는 분류사전[2]을 전자화하여 다음의 (그림1)

과 같이 분야트리를 구축하였다.



(그림 1) 분야트리와 분야연상어의 예

분야의 지정은 분야명의 패스 <P>로 기술하지만, 뿌리에 상응하는 <전체분야>는 생략하여 기술하는 것을 원칙으로 한다. 특히 모순이 생기지 않는 경우는 전체패스명을 생략하고, 종단분야만으로 설명한다.

2.2 분야연상어의 수준

수준 1의 완전연상어[1]에서 '국기원'은 종단분야 <태권도>를 오직 하나의 분야로 한정한다. 수준 2의 '단식'과 '복식'은 준완전연상어인데, 부모분야 <스포츠>내에서 복수의 종단분야 <테니스>, <탁구> 혹은 <배드민턴> 등을 한정한다. 수준 3의 중간연상어 '시합'은 어떠한 종단분야도 한정하지 않으나, 한 개의 중간분야 <스포츠>를 한정한다. 또한, 수준 4의 다분야연상어 '승패'는 중간분야 <스포츠> 혹은 복수의 종단분야 <취미·오락/장기>, <정치/선거> 등 복수의 분야를 한정할 수 있는 분야연상어이다. 마지막으로 수준 5의 비연상어는 '경우', '사용'과 같이 어떤 특정분야도 한정하지 않는 단어이다(<표 1> 참조).

〈표 1〉 각 수준별 분야연상어의 예

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

3. 단락의 결정

본 방법에서는 문서의 각 문장마다 처리를 진행해 분야별 단락을 추출한다. 이하 설명에서 사용되는 각각의 변수를 정의한다. 먼저, 처리대상 문서 $d_i = \{s_{i,1}, s_{i,2}, s_{i,3}, \cdots, s_{i,j}, \cdots, s_{i,m}\}$이다. 단, $s_{i,j}$는 문서 d_i 내의 j번째 문을 표시한다. F는 분야트리 전체집합을 의미하며, $\{F_1, F_2, F_3, \cdots, F_k,\}$으로 구성되어 있다.

$Frequency(s_{i,j}, F_k)$는 문서 d_i의 한 문장 s_j 내에 존재하는 분야 F_k의 분야연상어의 점수이다.

$Passage(F_k) = \{P_{k,1}, P_{k,2}, \cdots, P_{k,p}, \cdots\}$는 문서 d_i 내에 존재하는 분야 F_k의 단락의 집합으로 정의한다. 단, $P_{k,p}$은 문서 d_i 내에 존재하는 분야 F_k의 p번째 단락 집합을 표시한다.

3.1 분야연상어의 점수집계

본 방법에서는 문서 내에 존재하는 분야연상어를 각 문장에서 추출한다. 추출할 때 복수 키워드에 대한 고속 문자열 조합법으로 알려진 AC법[3]을 이용한다. 미리 인간이 자신의 상식지식으로 구축한 분야연상어를 AC법을 이용해 AC사전으로 저장하여 두고, 각 문장에 존재하는 모든 분야연상어를 추출한다. 추출된 분야연상어는 각 수준에 따라 분야를 한정하는 정도가 다르기 때문에 동일분야에 대한 각 수준별 점수를 합산한다. 각 수준의 점수로서 수준 1을 10, 수준 2를 5, 수준 3을 3, 수준 4를 2점으로 각각 설정한다.

아래의 예제문장(각 문장의 $s_{i,1}$~$s_{i,4}$가 <야구>, $s_{i,5}$~$s_{i,6}$이 <축구>의 화제에 대하여 쓰여진 문서)에 대하여 점수집계를 하면 이탤릭·볼드체 단어가 분야연상어로 추출된다. 여기서 각 분야연상어의 오른쪽 위에 기술한 위첨자는 분야연상어의 각 수준을 표시한다.

◎ 예제 문서

오늘 체육시간에 ***시합***[3]이 있었다. $s_{i,1}$

나는 ***선수***[3]로 출전한다. $s_{i,2}$

비가 우려되었으나 아무 문제없었다. $s_{i,3}$

시합[3]은 굿바이 ***홈런***[1]으로 이겼다. $s_{i,4}$

다음 주에는 ***축구***[1]가 있다. $s_{i,5}$

나는 ***MF***[1]로 출전할 예정이다. $s_{i,6}$

〈표 2〉 예제 문서의 점수집계 결과

문장 번호	야 구	테니스	축 구
$s_{i,1}$	3	3	3
$s_{i,2}$	3	0	0
$s_{i,3}$	0	0	0
$s_{i,4}$	13	3	3
$s_{i,5}$	0	0	10
$s_{i,6}$	0	0	10

각 문을 가장 점수가 높은 순으로 분야별 단락을 추출하면 Passage(<야구>) = $\{\{s_{i,1}, s_{i,2}\}, \{s_{i,4}\}\}$,

Passage(<테니스>) = $\{\{s_{i,1}\}\}$[1], Passage(<축구>)

1) 분야연상어 '시합'은 중간분야 <스포츠>를 한정하기 때문에 분야 <테니스>에 관한 분야연상어도 가능하다.

$=(\{s_{i,1}\},\{s_{i,5}, s_{i,6}\})$의 단락이 형성된다.

그러나, 단락분할 결과 다음과 같은 두 가지 문제점이 있다.

① 분야연상어가 각 문장에서 계속하여 출현하지 않는다면, 특정화제의 단락은 분리되어 화제의 실마리가 끊어진다.

② 한 개의 문에서 복수분야의 분야연상어가 같은 점수로 출현하면 여러 분야에 속하는 단락이 형성되어 단락 들 사이에 분야별 중복이 발생한다.

3.2 화제의 계속성과 전환성

신문이나 잡지기사 등의 일반적인 문서 내에 기술되는 화제의 흐름에는 다음의 두 가지 특징이 있다. 첫째, 일련의 화제는 계속성을 가지며, 실마리를 형성하고 있기 때문에 한 개의 화제가 산발적으로 진행되는 일은 없다. 이를 화제의 계속성이라하고 화제의 계속성을 "계속도"라 부르고 α로 명시한다. 둘째, 화제의 흐름에는 전환점이 있으며, 복수의 화제가 병행적으로 동시 진행되거나 중복되는 일은 없다. 이를 "전환성"이라는 관점에서 단락간 화제구간을 명확하게 한다[1]. 전환성을 측정하는 척도를 "전환도"라 정의하며, β로 표시한다. 다시 이들 특징을 이용하여 한 개의 문은 한 개 이하의 화제에 대응하기 때문에 처리대상의 문이 특정화제로 되는 분야를 "화제분야"라 정의하고 F_{theme}으로 표시한다.

3.3 계속도와 전환도의 계산

본 방법에서는 계속도 α를 산출할 때에 화제의 계속성이 쇠퇴하는 비율을 "쇠퇴도"로 정의한다. 분야연상어의 연속출현성을 고려한 쇠퇴율(Decline)을 이하의 계산식으로 정의한다.

$$Dec_{i,j} = \left[\frac{\sum_{s_{i,k} \in C_i} (Freq(s_{i,k}, F_{theme}))}{num(C_i)+1} + \frac{Freq(s_{i,j}, F_{theme})}{num(C_i)+1} \right]$$

단, $C_i = \{s_{i,j-n}, \cdots, s_{i,k}, \cdots, s_{i,j-1}\}$은 문 $s_{i,j-1}$에서 거꾸로 진행하여 화제분야 F_{theme}의 분야연상어가 연속으로 출현하고 있는 문의 집합이다. 위의 식에 의해 산출된 쇠퇴율을 사용하여 문 $s_{i,j}$에 대한 계속도 α_i를 계산한다. 본 방법에서는 문 $s_{i,i-1}$에서 문 $s_{i,i}$로(해석이 새로운 문으로) 진행할 때, 화제분야 F_{theme}의 계속도가 쇠퇴하고, 문 $s_{i,j}$에서 화제분야 F_{theme}의 분야연상어가 출현하면 화제의 계속성이 상승하였다고 생각한다. 계속도 α_i의 계산 방법을 표시한다.

[계속도(α)의 계산 알고리즘]

① $\alpha_j = \alpha_{j-1} + \rho \times Decline_j$

(단, $\alpha_j < 0$의 경우는 $\alpha_j = 0$이 된다.)

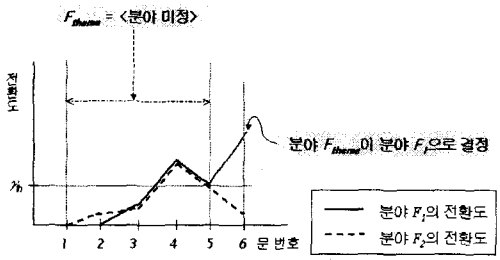
② $\alpha_j = \alpha_j + Frequency(s_i, F_{theme})$ [순서 종료]

3.4 단락의 결정

본 논문의 방법은 각 문장마다 단락분할을 판단한다. 단락결정에 대한 처리를 화제의 출현 판정처리, 전환처리, 계속처리로 나누어 설명한다. 먼저, 화제분야 F_{theme}이 한 가지로 정해지지 않으면 <분야미정>으로 정의한다. 각 문에 대하여 $\beta(F_k)$를 산출해 화제의 출현 판정처리를 수행한다. F_{theme}이 특정한 분야로 결정되어 있으면, α와 $\beta(F_k)$를 계산하여, $\alpha < \beta(F_k)$이면 화제 전환처리를 한다. 반대로 $\alpha > \beta(F_k)$이면 화제 계속처리를 행한다.

3.4.1 화제출현

본 처리에서는 각 분야의 전환도 $\beta(F_k)$로부터 어느 분야가 화제분야 F_{theme}이 되기 쉬운가를 판정한다. 먼저 $\beta(F_k)$가 γ_{th}(임계값)를 넘지 않거나 또는 $\beta(F_k)$가 최대가 되는 분야 F_k가 두 분야 이상 존재하는 경우, F_{theme}=<분야미정>으로 한다. 해석 문을 단락의 후보로 선정하여 스택에 저장한다. 반대로, $\beta(F_k)$가 특정 임계값을 초과하고, 최대가 되는 F_k가 한 분야로 모아지는 경우, F_k를 F_{theme}이라 하고, 단락형성의 후보에서 $Freq(s_{i,j}, F_{theme}) = 0$인 문장 s_j를 선택하여 제외한다. (그림 2)에 표시한 예와 같이 문 s_6에서 $F_{theme}=F_1$이 되고, 문 $s_{i,1}$와 $s_{i,2}$를 선택하여 제외한 $s_{i,3} \sim s_{i,6}$을 분야 F_{theme}의 단락구성 후보문장으로 형성된다



(그림 2) 분야미정으로 스택에 저장된 문장에서 화제분야의 결정

3.4.2 화제전환

화제의 전환이 일어난 경우 인접하는 문장사이에서 구간을 분리할 필요가 있다. 예를 들면, 전환이 일어난 문 $s_{i,j}$에서 새로운 화제가 나타나면, $s_{i,j-1}$을 구간분리위치로 하는 것이 바람직하다. 그러나 화제가 전환되기 이전의 문 $s_{i,j}$까지는 화제가 계속될 가능성이 있기 때문에 스택내의 문장집합에서 처리를 거꾸로 진행하여 구간 분리위치 $s_{i,j}$'을 설정한다.

계속도가 감소하기 시작한 부분과 전환도가 증가하기 시작한 부분 중 어느 하나를 우선하여 한 단락으로 결정하는가는 보다 깊은 논의가 필요하지만 이번 연구에서는 전환도의 증가를 우선하기로 하였다.

3.4.3 화제계속

화제 계속처리는 계속도 α가 특정 임계값 γ_{th} 이

상의 경우에는 화제가 계속되고 있다고 판단하여, 문 $s_{i,j}$를 단락구성의 문장 후보에 추가한다. 만약 α가 α_{th}보다 낮을 경우는 화제가 종료했다고 판단하여 단락 후보에서 Freq ($s_{i,j}$, F_{theme}) = 0의 문 $s_{i,j}$를 제외한 나머지 문을 Passage (F_{theme})에 추가한다.

4. 실험 및 평가

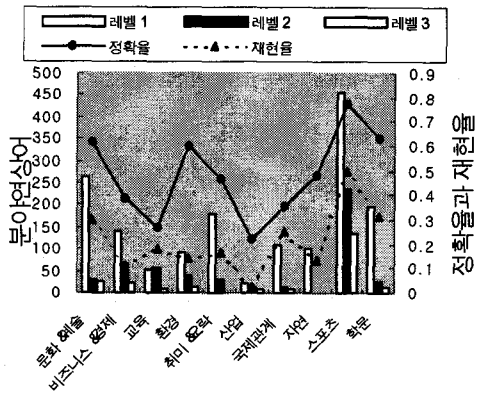
실험을 위해 각 분야별·수준별로 구축된 3,248개의 분야연상어를 구축하였다. 본 시스템이 출력한 단락과 인간이 직접 자신의 상식지식으로 결정한 단락이 어느 정도 일치하는가를 정확율과 재현율을 통해 비교한다. 정확율(P)과 재현율(R)을 아래의 식을 이용하여 계산하였다.

$$P = \frac{P_{correct}}{P_{output}}, \quad R = \frac{P_{correct}}{P_{answer}}$$

여기서, $P_{correct}$은 출력된 단락과 정답 단락이 일치하는 문자수를, P_{output}은 출력 단락의 문자수, P_{answer}는 정답 단락의 문자수를 나타낸다.

실험 데이터 준비는 분야연상어를 구축할 때 사용한 문서에서 작성한 데이터셋(Training Set)과 분야연상어의 구축에 쓰이지 않는 문서에서 작성한 데이터셋(Test Set)등 두 종류로 나누어 실험 데이터를 준비하였다.

(그림 3)에서 보는 바와 같이 각 분야의 평균 정확율은 약 0.82, 재현율은 0.63이 되어 충분히 실용적이며, 본 방법의 유효성이 입증되었다.



(그림 3) 분야 연상어 수와 Training Set의 정확율과 재현율

Training Set의 정밀도와 비교해 전체적으로 재현율이 저하되어 있다. 이것은 검출된 분야연상어의 수가 극히 감소하였기 때문이다. 그러나 <스포츠>와 같이 질적 혹은 양적으로 잘 정돈된 분야연상어가 구축되어 있는 분야에 관해서는 정밀도의 저하가 크게 보이지 않아 충분히 실용화 할 수 있다. 따라서, 분야연상어가 구축되기 쉬운 분야나 수준 1의 분야연상어가 많이 존재하는 분야에 대해서는 본 방법은 상당히 높은 정밀도의 단락을 추출할 수 있다.

5. 결론

본 논문에서 제시하는 방법론은 문서내 화제의 계속성과 전환성에 기반한 검색방법이므로 가공되지 않은 자연어문장 형태의 정보를 추출하는 유용한 방법이다. 결론적으로 본 논문의 방법은 텍스트의 특정 화제분야를 대표하는 실마리로서 분야연상어를 이용하였기 때문에 인간의 두뇌 혹은 인지작용과 유사하게 컴퓨터가 텍스트를 읽어감에 따라 텍스트가 어느 분야에 속하는지 빠르게 판단한다. 또한 단락검색시 화제의 전환성과 계속성을 고려하였기 때문에 동일 분야의 텍스트가 분리되는 현상을 방지하고, 복수분야에 속하는 텍스트의 중복을 제거하는 새로운 단락검색법이다.

참고문헌

[1] 이상곤, "분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할 방법", 정보처리학회논문지 B, 제 10권, 제 1호, 2003.

[2] 남영신 공저, 새로운 우리말 분류대사전, 성안당, 2002.

[3] Aho, A. V., & Corasick, M. J., "Efficient String MatchingAn Aid to Bibliographic Search," Communications of the ACM, Vol. 18, No. 6, pp. 333-340, 1975.

[4] Fuketa, M., Lee, S., Tsuji, T., Okada, M., & Aoe, J., "A Document Classification Method by Using Field Association Words," An International Journal of Information Sciences, Elsevier Science, Vol. 126, No. 1-4, pp. 57-70, 2000.

[5] Lee, S., Koyama, M., Mizobuchi, S., Uchibayashi, K., Kawano, F., Komatsu, T., & Aoe, J., "Cross-Language Multi-Media Information Retrieval System: BOSS," Paper Presented at the 18th International Conference on Computer Processing of Oriental Languages(ICCPOL '99), 1999.

[6] 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.