분야연상어를 이용한 화제분야의 계산방법과 단락검색

이 상 곤†

요 약

텍스트에 임베디드 되어 있는 부가적인 정보를 이용하여 문서의 실제적인 의미단위인 텍스트를 분리하는 단락검색은 중요한 기술이다. 본 논문에서는 문서의 분야에 적합한 단락만을 분리하여 사용자의 요구에 적합한 단락을 추출하는 기술을 설명한다. 문서에서 분야연상어를 추출하여, 각 문장마다 화제의 분야가 어떻게 커져가고, 줄어들고, 변화하여 가는지를 측정하는 방법을 실험을 통해 설명한다. 긴 문서에서 어떤 화제가 출현하는가를 파악하고, 화제가 계속되거나 혹은 전환되는 지점을 측정하고, 분야별로 단락을 구분하는 방법을 계산한다. 12,500개의 한국어 신문기사를 이용하여 실험한 결과 88%의 정확률과 78%의 재현율을 얻을 수 있었다.

Passage Retrieval and Calculation Method of Topic Field by Using Field-Associated Terms

Samuel-Sangkon Lee†

ABSTRACT

It is important to segment a text, which is independent upon any text-embedded auxiliary information. This paper presents a technique for dividing the text into field-coherent passages. The presented method is based upon extracting field-associated terms from the text measuring how the topics grow, shrink and shift from sentence to sentence. We propose measures of topic continuity and of topic transition and suggest how those could be used to find the boundaries among passages. After collecting 12,500 documents, we obtain 88% for average precision and 78% for recall in Korean training set.

키워드 : 분야연상어(Field-Associated Term), 화제분야의 추적방법(Tracing Topic Field), 화제분야의 계산방법(Calculation Method for Topic Field), 단락검색(Passage Retrieval)

1. 서 론

단락검색은 문서전체를 하나의 검색단위로 하지 않고, 문서 내 별개의 의미단위인 "단락(passage)"을 이용하여 검색요구와 문서와의 유사도를 계산하는 기술이다. 단락이란 일반적으로 문서 중에 존재하는 의미가 있는 연속된 일부분의 텍스트를 말한다. 이외에도 단락검색은 문서에서 사용자의 검색요구와 강하게 관련이 있으며, 의미적인 실마리를 가장 많이 포함하고 있는 텍스트의 일부분을 검색하여 사용자의 검색요구에 적절한 검색을 수행할 수 있는 기술이다.

본 논문에서는 기존연구에서 진행되어 온 분야연상어를 이용하여 문서에서 출현한 화제의 범위를 빠르게 파악하여, 각 화제의 범위를 결정하고, 화제의 경계부분을 파악하는 방법을 제안한다. 분야연상어를 이용한 단락검색[2-8, 10-12, 16-20]은 분야연상어[22]가 나타나는 텍스트는 특정 화제를 지시하는 단락으로 추측이 가능하지만, 분야연상어가 나타나지 않는 부분의 텍스트에서는 세그멘테이션이 발생한다[21]. 따라서 화제흐름의 특징을 조사하여, 분야연상어의 연속된 출현율을 토대로 산출된 화제의 계속성과 전환성[21]을 계산하여 사용자의 검색요구에 의미 있는 단락을 추출하고자 한다. 본 논문에서 제한하는 알고리즘에 의해 인접한 단락간의 구간분할을 명확히 하여 분야의 중복이 없도록 단락을 결정하는 방법을 제안한다. 본 방법에 의해 결정된 단락과 인간에 의해 결정된 단락이 어느 정도 일치하는가를 비교한다.

이하, 제2장과 제3장에서는 그동안 수집한 한국어와 영어의 분야연상어 컬렉션과 분야연상어 사전의 데이터구조에 대하여 간략히 설명하고, 4장에서는 신문기사에서 수집한 문서를 대상으로 본 논문에서 제안하는 단락검색을 적용하여 정확률과 재현율을 계산하고 그 유효성을 평가한다.

※ 본 연구는 2003학년도 학술진흥재단의 지원(KRF 2003 003 D00415)에 의하여 연구되었음.
† 종신회원 : 전주대학교 정보기술공학부 조교수
논문접수 : 2004년 7월 5일, 심사완료 : 2004년 12월 27일

마지막으로 제5장에서는 결론과 향후의 연구과제에 대하여 서술한다.

2. 분야연상어

분야연상어란 인간의 두뇌작용과 유사하게 문서에서 눈에 띄는 단어를 보는 것만으로 분야를 직관적으로 연상할 수 있는 단어[11-12]이다. 이 단어는 문서를 다른 분야의 문서와 구별할 수 있도록 하고, 문서의 분야를 파악하는데 유일하게 단서가 되는 '단어'나 '어구'를 말한다. 문서에서 몇 개의 분야연상어를 인식하여 신속하고 정확하게 동일한 분야의 단락을 생성할 수 있다. 또한, 이들 분야연상어(단일 분야연상어와 복합 분야연상어 모두 해당)는 문서 분야체계에 의해 분야트리에 부착되어 있다[22].

분야에 해당하는 분야연상어의 수준[21]과 분야연상어가 시간의 경과에 의하여 변화하는 것에 주목한다. 예를 들면, 분야 <야구>에 대하여 "투수", "포수" 등은 안정한 분야연상어이지만, 고교야구에서 "우승고"나 "선수명"은 시간의 변화에 따라 변화하는 불안정한 분야연상어이다. 또한, 단일연상어의 길이는 짧고, 개수도 유한하기 때문에 분야연상어의 후보를 사람이 직접 판단하여 선별한다.

단일어로 구성된 분야연상어를 단일 분야연상어, 그리고, 두개이상의 단일 분야연상어로 구성된 복합어를 복합 분야연상어라 정의하고, 기호 "과 "내에 기술한다. 단, 미등록어는 분야연상어 범주에 포함하지 않는다. 각각 한 개의 접사와 명사로 구성되는 일반적인 복합어 "소비세", "핵연료", "온난화" 등은 세분화하면 분야정보를 잃어버리기 쉬게 되므로 이러한 단어는 단일어로 취급한다. 덧붙여, 분야 <야구>에 대하여 고교명 "군산상고"나 과학기술에 관한 분야에서 회사명 "한국통신" 등의 고유명사(인명 이외)도 단일 분야연상어로 간주한다. 인명에 관한 고유명사 예를 들어, "선동렬"과 같이 성명으로 문서 중에 존재하는 경우는 그대로 단일분야연상어라 하고, "김응용감독"과 같이 보통명사 "감독"과 함께 복합분야연상어로 구성되어 있으면 "김응용"과 "감독"을 독립하여 각각 두개의 독립된 단일분야연상어로 취급한다.

분야트리란 각 분야의 상위·하위관계를 나타내는 분야체계를 트리구조로 표현한 것이며, 분야트리의 단말노드에 상응하는 분야를 종단분야, 종단분야 이외는 "중간분야"라 부른다. 본 연구에는 용어사전 [23]을 표본으로 분야트리를 구축하였다. 이 분야트리의 전체분야 수는 200개이며, 대분야수는 10개, 중간분야 수는 18개, 종단분야 수는 172개(깊이 2, 3, 4의 종단분야는 각각 174개, 208개, 11개)이다. 직접적인 상위분야 혹은 하위분야를 각각 부모분야, 자식분야라 부른다. 분야의 지정은 분야명의 패스 <P>로 기술하지만, 뿌리에 상응하는 <전체분야>는 생략하여 기술하는 것을 원칙으로 한다. 특히 모순이 생기지 않는 경우는 전체 패스지정을 생략하고 종단분야만으로 기술한다. 예를 들면, 분야패스 <P>가 <스포츠/구기/배드민턴>이면 <스포츠/구기>의 하위분야 <배드민턴>을 표시한다.

미리 분류한 분야트리에 따라 문서 데이터에 대하여 형태소 해석을 실시하고, 각 문서 내에 명사로 존재하는 분야연상어를 추출한다. 여기서 구해진 분야연상어는 형태소 사전에 등록되어 있는 단어이며, 복합어에 대한 분야연상어는 단일어의 분야계승을 기초로 반자동적으로 구축할 수 있다. 각 분야에 속하는 문서 데이터 내에 출현하는 단어의 집중률[22]을 계산하여 각 분야의 분야연상어 사전을 이용한다. 형태소 해석 결과 미등록어가 되는 단어는 분야연상어의 대상으로 하지 않는다. 분야트리 내에 문서 데이터에서 수집한 단일어의 분야연상어를 저장하였다.

수집한 문서에서 결정한 분야연상어에는 연상되는 분야의 한정범위에 차이가 있다. 단어는 유일한 종단분야나 중간분야를 한정하는 단어 혹은, 복수의 종단분야나 중간분야를 한정한다고 가정하고, 각 분야연상어 w의 수준(레벨)을 이용한다. 분야트리에 따라 각 분야연상어의 수준은 다음과 같다. 수준 1의 완전연상어 "국기원"과 같이 종단분야 <태권도>를 오직 하나의 분야로 한정한다. 수준 2의 준완전연상어 "단식", "복식"은 같은 부모분야 <스포츠>내의 복수의 종단분야 <테니스>, <탁구> 혹은 <배드민턴> 등을 한정한다.

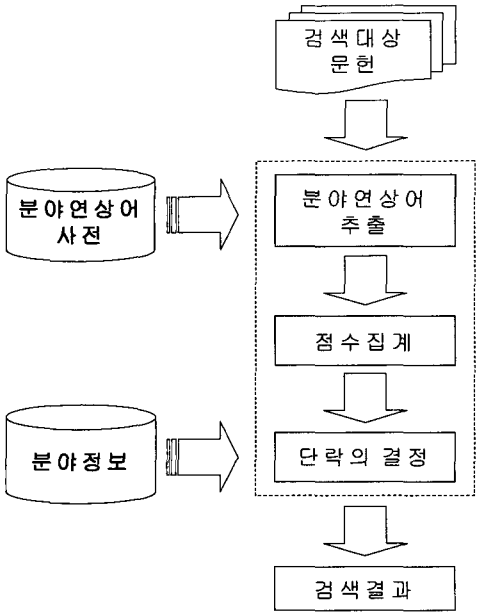
〈표 1〉 점수집계의 예

d		sub script 분 야 별		
		F_1(야구)	F_2(축구)	F_3(테니스)
	s_1	12	-	-
	s_2	20	-	-
	s_3	-	-	-
문	s_4	8	-	-
	s_5	-	12	-
	s_6	12	12	-
서	s_7	-	16	12
	s_8	-	2	-
	s_9	-	-	16
	s_{10}	-	-	20

수준 3의 중간연상어 "시합"은 특정한 종단분야는 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 "경우", "사용" 등의 단어와 같이 어떠한 특정분야도 한정하지 않는 단어는 수준 5의 비연상어라 한다.

본 방법에서는 각 문장에서 출현하는 분야연상어의 수준에 따라 다음과 같이 점수를 부여한다. 수준 1을 12, 수준 2를 8, 수준 3을 4, 수준 4를 2점씩 부여하고, 추출된 분야연상어의 수준에 해당하는 점수를 부여하여 <표 1>과 같

이 각각의 문장에서 분야별 점수를 합산한다.



(그림 1) 화제분야의 추적엔진의 개요

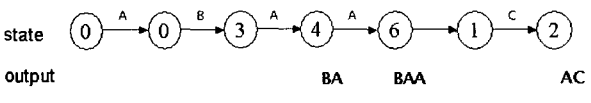
3. 분야연상어 사전의 구축

3.1 시스템 개요

본 논문에서 제안하는 화제분야 추적엔진의 개요를 (그림 1)에 표시하였다. 분야연상어의 추출은 검색대상이 되는 문서에서 한 문장마다 분야연상어를 추출한다. 점수계산(score calculation) 모듈은 분야트리에 부착된 분야연상어를 참조하여 개개의 문장에서 출현하는 모든 분야연상어를 추출하고, 각 분야에 해당하는 점수를 문장별로 누적가산한다. 단락의 결정은 각 분야마다 얻어진 점수를 이용하여 계속도, 전환도를 계산하고, 이를 이용하여 특정화제의 출현, 계속, 전환처리를 수행한다. 이는 문서에서 검색요구와 관련이 깊은 분야 혹은 특정 화제별로 단락을 분할하여 검출한다.

3.2 분야연상어의 추출

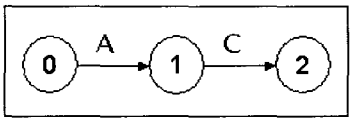
본 연구는 AC 방법을 이용하여 분야연상어가 키워드로 혹은 텍스트의 부분 문자열 조합으로 등록된 "분야연상어 사전"에서 대상 문서 내의 이용 가능한 모든 분야연상어를 탐색하여 추출한다.



(그림 2) PMM의 동작 예

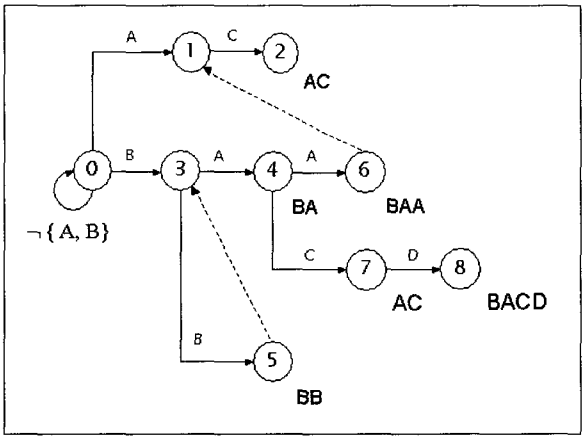
3.2.1 AC(Aho-Corasick) Machine

AC 방법은 문자열 검색 알고리즘의 하나이다. 문서의 처음에서 끝까지 일련의 문자열 탐색으로 복수의 키워드를 모두 검출할 수 있는 알고리즘[1]이다. 최초에 주어진 키워드를 유한오토마타, 다시 말하면 Pattern Matching Machine (PMM)으로 만들어, 이를 문서상에서 탐색하여 모든 키워드를 추출한다.



(그림 3) goto 클래스

PMM은 goto, failure, output 등 세 개의 함수로 구성된다. (그림 2)는 패턴의 집합 Π = {AC, BA, BB, BAA, BACD}에 대한 PMM의 예이다. 여기서, 실선은 goto 함수에 의한 천이, 점선은 failure 함수에 의한 천이, $\neg\{A, B\}$는 A, B 이외의 모든 문자를 표시한다. PMM의 구성방법에 대해 알아보자. 첫 단계로서 초기상태 0으로 된 goto 클래스를 작성한다. 다음에 최초의 패턴 AC에 대하여 상태 1, 2를 도입하고, 레이블 A를 붙여서 상태 1로 천이하고, 상태 1에서 레이블 C가 붙은 상태 2로 천이를 연장한다((그림 4) 참고).



(그림 4) 패턴 조합

이 상태에서 (그림 3)과 같이 goto 변수, output 변수를 각각 goto(0, A) = 1, goto(1, C) = 2, output(2) = {AC}로 정의한다. 이와 같이 하여 모든 키의 PMM 등록을 완료한 후, 초기상태 0에서 천이 되지 않는 문자에 대한 루프를 작성한다. 두 번째 단계에서, goto 변수에 의한 천이가 되지 않는 경우는 failure 변수에 의해 천이를 계속한다. 상태 s에 대한 failure 변수의 수치는 상태 0에서 상태 s에 도달하는 변에 레이블(문자)을 붙여 상태 0에서 도달되는 가장 먼 상태로 정의한다. 이와 같이 작성된 failure 변수가 (그림 4)

의 점선 부분이다.

```
1)  j=k=1, ... 339, αj(Ftheme)=βj(Fk)=0, m = # of sentences in d,
    Ftheme = "Field-Neutral", Old-Topic ← Null
2)   repeat until sm
3)      repeat until Fk (k = 1, ... 339)
4)         if [Ftheme == "Field_Neutral"] then
5)            βj(Fk) ← Freq(sj, Fk)
6)            goto Appearance process
7)            goto Step 2)
8)         else if [Old-Topic != Ftheme]
9)            {   αj(Ftheme) ← Freq(sj, Fk)
10)               Old-Topic ← Ftheme
11)               repeat until Fk
                              except for Ftheme (k=1, ... 339)
12)               βj(Fk) ← Freq(sj, Fk)}
13)        else {
14)               calculate αj(Ftheme)
15)               repeat until Fk except for (Ftheme = Fk)
16)                  calculate βj(Fk) }
17)     if [Fk satisfying (αj(Ftheme) < max(βj(Fk)))]
        then goto Transition process
18)     else goto Continuity process
19)  Passage(d, Ftheme) ← stack
20)  clear stack
```

(그림 5) 전체흐름제어(Control Flow) : 알고리즘①

예를 들어, 위의 PMM을 사용하여 텍스트 "CBAAC"에서 원하는 문자패턴을 검출하는 예를 (그림 4)에 표시한다. 먼저, 초기상태 0에서 시작하여 첫 문자 C에 대한 천이가 없기 때문에 자신으로 상태를 천이 한다. 다음 B에 의하여 상태 3에 천이 한다. 그리고 세 번째 문자 A에 의해 상태 4에 천이하고, 패턴 BA를 검출한다. 다음의 A에 의하여 상태 6에 천이해서 BAA를 검출한다. 최후에 다섯 번째 문자 C에 대한 천이는 상태 6에서 문자 C에 의해 다음 상태가 존재하지 않으므로, failure 변수에 의해 상태 1에 천이 한다. 여기서, C에 의한 천이에 의해 상태 2에서 AC를 검출한다. 이와 같이 진행하여 문자열 CBAAC를 한번의 탐색으로 (그림 2)의 우리가 탐색을 원하는 3가지 문자패턴 BA, BAA, AC를 모두 검색할 수 있다.

```
1)  if [not exist Fk satisfying (βj(Fk) > 0)]
2)     clear stack
3)     terminate Appearance
4)  if [(βj(Fk) ≤ αth) OR (num(max(βj(Fk))) > 2]
5)     insert sj to stack
6)     terminate Appearance
7)  else if [(βj(Fk) > αth) AND (num(max(βj(Fk))) = 1]
8)  {  Ftheme ← Fk
9)     Old-Topic ← Fk
10)    if [stack != empty]
11)       delete sj satisfying βj(Fk) = 0
12)    αj(Ftheme) ← βj(Ftheme)
13)     insert sj to stack }
```

(그림 6) 화제출현판정(Appearance)처리 : 알고리즘②

4. 화제분야의 계산방법

어떤 문서(d)가 2절에서 설명한 방법에 의해 앞의 <표 1>과 같은 점수집계 결과를 가지면, 본 논문에서 제안하는 단락결정 알고리즘을 이용하여 단락 사이의 경계를 해석하고, 각 분야별 단락을 추출하는 실행 예를 <표 2>에 보인다(단, ρ = 0.8, α_{theme} = 0으로 계산, F_k의 k는 1에서 3으로 가정한다). 다음에 <표 2>의 화제 출현 판정과 각 문장에서의 계속도와 전환도 값의 변화를 차근차근 설명한다.

(그림 5)의 전체흐름제어 알고리즘에서 F_k의 k 값은 1에서 3까지이고, s_j는 d를 구성하는 s_1에서 s_{10}까지라 한다. 각 문장별·분야별 전환도($\beta_j(F_k)$)와 계속도($\alpha_j(F_{theme})$)의 값이 초기화되고, 화제분야 F_{theme}이 "분야미정(Field-Neutral)"으로 초기화된다. 변수 Old-Topic은 현재의 화제분야 F_{theme}의 변경을 파악하기 위한 플래그이고, Null로 초기화된다.

```
1)  Passage(d, Ftheme) ← stack
2)  Perform following step at sj
       which is the sentence that α(Ftheme)
       finally decreases,
        and Freq(sj, Ftheme) = 0
        if [βj(Fk)=0 at sj] then j'=j [j' 결정]
3)  clear stack
4)  if [(βj(Fk)≤αth) OR
              (num(max(βj(Fk))) > 2)]
       then Ftheme ← "Field-Neutral"
5)  else if [(βj(Fk) > αth) AND
        (num(max(βj(Fk))) = 1] then Ftheme ← Fk
6)     αj'(Ftheme) ← 0
7)     repeat until Fk except for (Fk = Ftheme)
          βj'(Fk) ← 0
8)  j = j' - 1
9)  goto step 2) in algorithm ③
```

(그림 7) 화제전환(Transition)처리 : 알고리즘③

순서 2)에서 현재 처리할 문장이 s_1이 되고, 순서 3)에서 F_k의 k가 1이 되어, F_k = F_1이 된다. 순서 4)에서 F_{theme}이 "Field-Neutral"(초기화된 값)이므로 순서 5)로 진행하고, $Freq(s_1, F_1)$의 값 12를 $\beta_j(F_1)$에 대입한다. 순서 6)에서 화제 출현판정 처리 알고리즘 ①로 분기한다.

```
1)  if [αj(Ftheme) ≥ αth] then insert sj to stack
2)  else if [αj(Ftheme) < αth]
       then delete s1 ..., sj-1을 만족하는
       Freq(st, Ftheme) ≠ 0 (t = 1, ..., j-1)
       from stack regard st to Passage(d, Ftheme)
       clear stack
3)  goto step 2) in algorithm ①
```

(그림 8) 화제계속(Continuity)처리 : 알고리즘④

화제출현판정처리 알고리즘②에서 $\beta_1(F_1)$(=12) > 0을 만족하는 F_k가 존재하므로 순서 4)로 진행한다.

〈표 2〉 화제의 출현 · 전환 · 계속을 고려한 단계별 알고리즘의 수행 예

$Freq(s_j, F_k)$				$\beta_j(F_k)$						Old Topic	$a_j(F_{theme})$		F_{theme}	j'	Stack	Passage(d, F_k)
s_j	F_1	F_2	F_3	β_j	(F_1)	(F_2)	(F_3)	max	(F_{theme})							
initial					0	0	0			NULL			Field-Neutral			
s_1	12	-	-	β_1	12	0	0	F_1	12		a_1	12	F_1		s_1	
s_2	20	-	-	β_2		0	0				a_2	19.2			s_1, s_2	
s_3	-	-	-	β_3		0	0				a_3	10.68			s_1, s_2, s_3	
s_4	8	-	-	β_4		0	0				a_4	10.68			s_1, s_2, s_3, s_4	
				β_5		10.08	0	F_2			a_5	4.28			s_1, s_2, s_3, s_4	s_1, s_2, s_3, s_4
										F_1				5	s_1, s_2, s_3, s_4	
															clear	
													F_2			
s_5	-	12	-									0				
					0		0									
												12				
										F_2						
					0		0								s_5	
s_6	12	12	-	β_6	7.2		0				a_6	19.2			s_5, s_6	
s_7	-	16	12	β_7	4		8.8				a_7	24.56			s_5, s_6, s_7	
s_8	-	2	-	β_8	1.6		6.4				a_8	18.16			s_5, s_6, s_7, s_8	
				β_9	0		16.32	F_3			a_9	11.44		9	s_5, s_6, s_7, s_8	s_5, s_6, s_7, s_8
s_9	-	-	16		0	0						0	F_3		clear	
					0	0				F_3		16			s_9	
s_{10}	-	-	20	β_{10}	0	0					a_{10}	3.2			s_9, s_{10}	
Final																s_9, s_{10}

$\beta_1(F_1)(=12) \leq a_{threshold}(=0)$을 만족하지 않을 뿐더러 $\max(\beta_1(F_k))$ 값을 갖는 수가 2개 이상이 아니므로 순서 7)로 진행한다. $\beta_1(F_1)(=12) > a_{th}(=0)$이고 최대치를 갖는 F_k의 수가 한 개이므로 순서 8)로 진행하여 F_1이 F_{theme}이 된다.

순서 9)에서 F1이 Old-Topic이 된다. 순서 10)에서 stack이 empty이므로 순서 11)은 실행되지 않는다. 순서 12)에서 $\beta_1(F_{theme})$(즉, $\beta_1(F_1)$)이 $a_j(F_{theme})$의 값으로 대입된다. 순서 13)에서 stack에 s_1이 push 된다.

전체흐름제어 알고리즘 ①으로 되돌아와 순서 7)로부터 순서 2)로 진행하여 j가 2가 된다. 순서 3)에서 F_k의 k가 1이 되어, $F_k=F_1$이 된다. 순서 4)에서 F_{theme}이 Field-Neutral이 아니므로 순서 8)로 진행한다. 순서 8)에서 Old-Topic($=F_1$)과 $F_{theme}(=F_1)$이 같으므로 순서 13)으로 진행한다. 순서 14)에서 다음과 같이 $a_2(F_{theme})$을 계산한다. $a_2(F_{theme})$을 계산하기 위해 아래와 같이 Dec_2(2번째 문장에서의 쇠퇴율(decline)[12])을 먼저 계산한다. 그 결과 다음과 같이 19.2가 대입된다.

$$Dec_2 = -1 \times \left\lfloor \frac{[\sum_{s_k = s_1 \in C_i} Freq(s_1, F_{theme})] + Freq(s_2, F_{theme})}{num(C_i) + 1} \right\rfloor$$
$$= -1 \times \frac{[12] + 20}{1 + 1} = -1 \times \frac{32}{2} = -16$$

$$a_2(F_{theme}(=F_1)) = a_1(F_{theme}) + [\rho \times Dec_2] + Freq(s_2, F_{theme})$$
$$= 12 + [0.8 \times (-16)] + 20$$
$$= 12 + [-12.8] + 20$$
$$= 19.2$$

순서 15)에서 $F_{theme} = F_1$을 제외한 모든 F_k(여기서는 k = 2와 3)의 전환도 $\beta_2(F_2)$와 $\beta_2(F_3)$을 다음과 같이 계산한다. 계속도와 동일하게 Dec_2를 먼저 계산하여 그 결과를 전환도 계산식에 대입한다.

$$Dec_2 = -1 \times \left\lfloor \frac{[\sum_{s_k = s_1 \in C_i} Freq(s_1, F_2)] + Freq(s_2, F_2)}{num(C_i) + 1} \right\rfloor$$
$$= -1 \times \frac{[0] + 0}{1 + 1} = -1 \times \frac{0}{2} = 0$$

$$\beta_2(F_2) = \beta_1(F_2) + [\rho \times Dec_2] + Freq(s_2, F_2)$$
$$= 0 + [0.8 \times 0] + 0$$
$$= 0$$

그 결과 0이 대입된다.

F_3의 경우도 마찬가지로 다음과 같이 계산한다.

$$\beta_2(F_3) = \beta_1(F_3) + [\rho \times Dec_2] + Freq(s_2, F_3)$$
$$= 0 + [0.8 \times 0] + 0$$
$$= 0$$

$$\because Dec_2 = -1 \times \left\lfloor \frac{[\sum_{s_k = s_1 \in C_i} Freq(s_1, F_3)] + Freq(s_2, F_3)}{num(C_i) + 1} \right\rfloor$$
$$= -1 \times \frac{[0] + 0}{1 + 1} = -1 \times \frac{0}{2} = 0$$

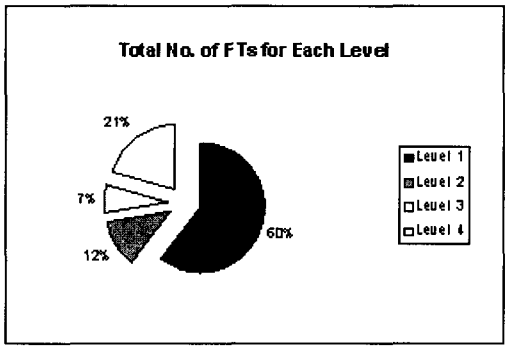
그 결과 <표 2>와 같이 0이 대입된다. 순서 17)에서

$\beta_2(F_2)(=0)$과 $\beta_2(F_3)(=0)$ 중 최대치(max)가 $\alpha_j(F_{theme})(\alpha_2(F_1)=19.2)$과 비교하여 전환도의 값이 크기 때문에 화제전환(Transition)처리로 분기한다. 다음의 순서 18)에서 화제계속(Continuity)처리로 분기하여 수행한다.

5. 실험 및 평가

5.1 분야연상어의 수집

실험을 위해 한국어 3,248개의 분야연상어를 수집하였다. 수집한 문서 데이터는 주로 조선일보 신문기사 2002~2004년 웹사이트에서 수집하여 미리 인간이 분야트리에 의해 분야연상어를 분류하여 각 분야연상어별 수준과 안정성을 사람이 결정하였다. <표 3>에 각 분야별·수준별로 수집한 한국어의 분야연상어 수를 표시하였다. 분야연상어의 수가 200을 넘는 분야에 bullet으로 표시하였다.



(그림 9) 분야연상어의 수준별 퍼센트

<표 3>에서 한국어의 분야연상어 수는 <교육>, <건강&의료>, <산업>, <사회생활> 등의 4 분야는 수집한 분야연상어가 적어서 제외하였다. <표 4>의 영어의 분야연상어는 <문화 & 예술>, <교육>, <산업>, <사회생활>, <학문> 등의 다섯 분야는 같은 이유로 제외하였다. (그림 10) (a)는 <표 3>을 그래프로 표현한 것이며, (그림 10) (b)는 5.3절의 유사한 분야에서의 비교실험을 위해 <스포츠/구기>의 하위분야 중 8개의 종단분야에 대해 수집한 분야연상어의 수를 나타낸다.

(그림 10)은 5.3절에서 다른 방법과의 비교실험을 위해 준비하였다. 분야 <스포츠>에 대한 분야연상어가 다른 분야에 비해 훨씬 많은 이유는 타 분야에 비해 <스포츠>는 인터넷에서 전자화 된 문서를 구하기 쉽고, 문서 내에 분야를 정확히 결정할 수 있는 단어(팀명 혹은 유명 선수명 혹은 감독명)가 다른 분야에 비해 많이 포함되어 있기 때문이다. (그림 9)는 제시한 각 수준별 분야연상어의 수를 퍼센트로 표시하였다. 수준 1의 분야연상어가 60%를 차지하고 준완전연상어를 합하면 전체의 72%를 차지하여 비교적 안정한 분야연상어가 수집되었다.

<표 3> 한국어 Training Set에서 수집한 각 분야별 분야연상어 수

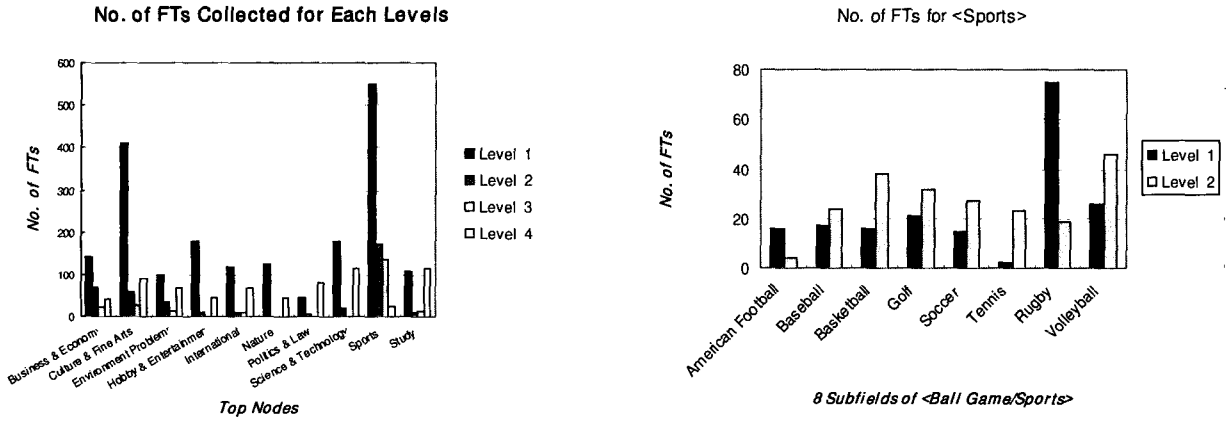
분야별	분야연상어 수준 1	2	3	4		총합
1. 비즈니스&경제	144	69	24	39	●	276
	52.1%	*25.0%*	*8.7%*	*14.2%*		*100%*
2. 문화&예술	410	58	28	89	●	585
	70.0%	*9.9%*	*4.8%*	*15.3%*		
3. 환경문제	100	32	13	66	●	211
	47.4%	*15.2%*	*6.2%*	*31.2%*		
4. 취미&오락	180	9	0	42	●	231
	77.9%	*3.9%*	*0%*	*18.2%*		
5. 국제관계	118	10	9	65	●	202
	58.4%	*5.0%*	*4.5%*	*32.1%*		
6. 자연	126	1	0	43		170
	74.1%	*0.6%*	*0%*	*25.3%*		
7. 정치&법률	43	8	1	79		131
	32.8%	*6.1%*	*0.7%*	*60.4%*		
8. 과학기술	179	21	0	117	●	317
	56.5%	*6.6%*	*0%*	*36.9%*		
9. 스포츠	550	171	135	22	●	878
	62.6%	*19.5%*	*15.4%*	*2.5%*		
10. 학문	109	10	12	116	●	247
	44.1%	*4.0%*	*4.9%*	*47.0%*		
수준별 총합	1,959	389	222	678		3,248
	60.3%	11.9%	6.8%	20.8%		99.8%
평 균	195.9	38.9	22.2	67.8		324.8

<표 4> 영어 Training Set에서 수집한 각 분야별 분야연상어 수

분야별	분야연상어 수준 1	2	3	4		총합
1. 비즈니스&경제	63	10	14	5		92
	68.4%	*10.9%*	*15.2%*	*5.5%*		*100%*
2. 환경문제	33	8	14	5		60
	55.0%	*13.3%*	*23.3%*	*8.4%*		
3. 건강&의료	37	12	12	14		75
	49.3%	*16.0%*	*16.0%*	*18.7%*		
4. 취미&오락	53	10	4	8		75
	70.7%	*13.3%*	*5.3%*	*10.7%*		
5. 국제관계	55	10	5	40		110
	50.0%	*9.1%*	*4.5%*	*36.4%*		
6. 자연	35	12	6	10		53
	66.1%	*22.7%*	*11.3%*	*18.9%*		
7. 정치&법률	83	12	7	18		120
	69.2%	*10.0%*	*5.8%*	*15.0%*		
8. 과학기술	263	7	10	30	●	310
	84.8%	*2.3%*	*3.2%*	*9.7%*		
9. 스포츠	1,210	40	74	81	●	1,405
	86.1%	*2.8%*	*5.3%*	*5.8%*		
수준별 총합	1,832	111	146	211		2,300
	79.7%	4.8%	6.3%	9.2%		100%
평 균	203.6	12.3	16.2	23.4		255.6

〈표 5〉 Training Set에 대한 영어문서의 정확율, 재현율, F-Measure의 값

		Line(L)											
		5			*10*			*15*			*20*		
	평가척도	P	R	F	P	R	F	P	R	F	P	R	F
1.	비즈니스&경제	0.523 *0.323*	0.632 *0.432*	0.572 *0.369*	0.614 *0.414*	0.542 *0.342*	0.578 *0.374*	0.612 *0.412*	0.517 *0.317*	0.561 *0.358*	0.831 *0.631*	0.613 *0.413*	0.705 *0.499*
2.	건강&의료	0.571 *0.371*	0.531 *0.331*	0.669 *0.464*	0.721 *0.521*	0.514 *0.314*	0.602 *0.392*	0.652 *0.452*	0.562 *0.362*	0.604 *0.402*	0.759 *0.559*	0.652 *0.452*	0.702 *0.499*
3.	취미&오락	0.768 *0.568*	0.635 *0.435*	0.695 *0.492*	0.777 *0.577*	0.838 *0.638*	0.806 *0.605*	0.671 *0.471*	0.597 *0.397*	0.632 *0.430*	0.842 *0.642*	0.604 *0.404*	0.704 *0.495*
4.	국제관계	1.000 *0.812*	0.906 *0.706*	0.952 *0.755*	0.911 *0.711*	0.916 *0.716*	0.913 *0.713*	1.000 *0.812*	0.653 *0.453*	0.790 *0.581*	1.000 *0.834*	0.676 *0.476*	0.806 *0.606*
5.	자연	0.612 *0.412*	0.542 *0.342*	0.575 *0.373*	0.813 *0.613*	0.593 *0.393*	0.686 *0.479*	0.783 *0.583*	0.656 *0.456*	0.714 *0.511*	1.000 *0.832*	0.696 *0.496*	0.821 *0.621*
6.	정치&법률	0.751 *0.551*	0.562 *0.362*	0.643 *0.436*	0.972 *0.772*	0.825 *0.625*	0.893 *0.690*	0.773 *0.573*	0.598 *0.398*	0.674 *0.469*	1.000 *0.861*	0.622 *0.422*	0.766 *0.566*
7.	과학기술	0.523 *0.323*	0.435 *0.335*	0.475 *0.328*	0.802 *0.602*	0.652 *0.452*	0.719 *0.516*	0.742 *0.542*	0.571 *0.371*	0.645 *0.440*	1.000 *0.853*	0.846 *0.446*	0.916 *0.585*
8.	스포츠	0.623 *0.423*	0.593 *0.393*	0.607 *0.407*	0.919 *0.719*	0.842 *0.642*	0.878 *0.678*	1.000 *0.821*	0.640 *0.440*	0.780 *0.572*	1.000 *0.791*	0.779 *0.479*	0.877 *0.596*
	평 균	0.672 *0.472*	0.605 *0.417*	0.636 *0.443*	0.816 *0.616*	0.715 *0.515*	0.762 *0.561*	0.779 *0.583*	0.599 *0.399*	0.677 *0.474*	0.929 *0.750*	0.686 *0.448*	0.789 *0.561*



(a) (b)

(그림 10) 수집된 한국어의 분야연상어 수

5.2 최적의 파라미터 결정

실험을 위해 수집한 12,500 문서를 Training Set과 Test Set 등 각각 2가지 부류로 나누었다. Training Set은 분야연상어를 추출하는데 사용한 문서집합이다. Test Set은 분야연상어를 추출하지 않은 문서집합이다. 또한 Training Set에서 다섯 개의 분야를 무작위로 선택하여 L행을 무작위로 추출하였다. L의 값을 5, 10, 15, 20으로 변화시켜 가며 각각 하나의 파일로 혼합하였다. 이 파일의 크기는 약 1~3KB 정도로 제한하여 다음의 정확률과 재현율의 값을 계산하였다.

Training Set에 본 논문의 방법을 적용하여 정확률과 재현율을 계산한다. 정확률과 재현율의 계산은 시스템이 출력한 단락과 인간이 상식지식으로 결정한 단락이 어느 정도 일치하는가의 평가이다. 정확률(*Precision*)과 재현율(*Recall*)을 다음 식 (1)을 이용하여 계산하였다. 시스템이 출력한 단락의 문자수를 P_{output}, 인간이 작성한 정답 단락의 문자수를 P_{answer}, 시스템이 출력된 단락과 인간이 작성한 정답 단락이 일치하는 문자수를 P_{accord}로 계산하였다.

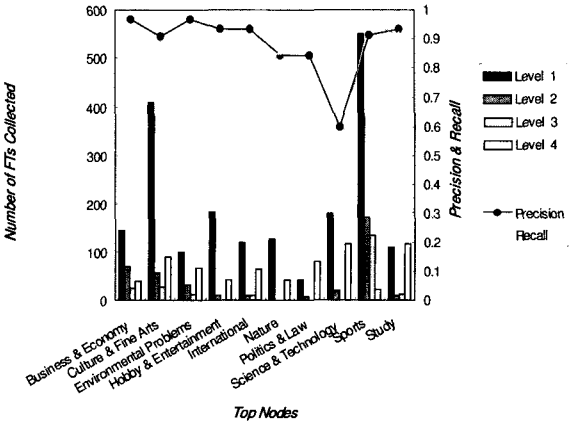
$$Precision = \frac{P_{accord}}{P_{output}},\quad Recall = \frac{P_{acccord}}{P_{answer}} \qquad (1)$$

문서의 종류나 분야에 따라 다음과 같이 다섯 종류의 정확률과 재현율의 평가를 얻을 수 있다.

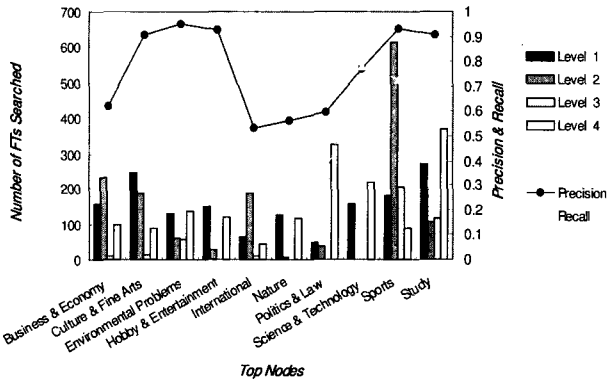
① 정확률 = 재현율 = 1인 경우 : P_{output}과 P_{answer}가 완전일치한다.
② 정확률 = 1, 재현율 < 1인 경우 : P_{output}이 모두 P_{accord}이고, P_{answer}중에는 P_{accord}가 아닌 것이 존재한다.
③ 정확률 < 1, 재현율 = 1인 경우 : P_{answer}가 모두 P_{accord}이지만, P_{output}중에는 P_{accord}가 아닌 것이 존재한다.
④ 정확률 < 1, 재현율 < 1인 경우 : P_{output}중에는 P_{accord}가 아닌 단락이 존재하고, P_{answer}중에는 P_{accord}가 아닌 것이 존재한다.
⑤ 정확률 = 재현율 = 0인 경우 : P_{output}이 전혀 출력되지 않거나, 존재하면 모두 P_{accord}이 아니다. 그리고, P_{accord}가 전혀 존재하지 않는다.

Training Set에 대한 정확률과 재현율을 (그림 11)의 상단에 표시한다. <문화-예술>과 <스포츠> 분야의 정확률과 재현율이 전체분야의 평균정확률과 재현율에 비교하여 높은 이유는 위의 분야에 대하여 수집된 분야연상어의 수가 많기 때문에 화제분야의 출현, 전환, 계속이 효과적으로 이루어졌다고 생각된다. <학문> 분야는 수준 1의 분야연상어 수가 높기 때문에 양호한 결과가 얻어졌다고 생각된다. 반면에, 분야연상어 수가 200을 넘고 있는 <비지니스-경제>와 <취미-오락>의 정확률과 재현율이 저하된 이유는 <비즈니스-경제>의 경우는 수준 2의, <취미-오락>과 <국제관계>는 수준 4의 분야연상어 비율이 높기 때문에 화제분야가 '분야미정(Field Neutral)'이 되기 쉽고, 이는 정확률과 재현율 저하의 원인이 되었다고 생각된다. <자연>은 수준 1의 분야연상어 수는 많지만, 수준 2와 3의 분야연상어의 수가 낮기 때문에 단락 형성시 문장의 중간에서 끊어져 정확률이 낮게 나타났다. <정치-법률>과 <과학기술>에서 정확률과 재현율이 낮은 이유는 수준 4의 분야연상어 수가 많기 때문에 분야결정이 잘 되지 않았다고 생각된다. 그러나 평균정확률은 0.88, 재현율은 0.78가 되어 충분히 실용적이며, 유효성이 입증할 수 있다.

Test Set에 대한 정확률과 재현율을 ((그림 11)의 하단)에 표시하였다. Training Set의 정밀도와 비교해 전체적으로 재현율이 저하하고 있는데, 그 이유는 Training Set에서 수집한 수준 1의 분야연상어 수보다 Test Set에서 검색된 수준 1의 분야연상어 수가 대폭 감소하였기 때문이다. 다른 이유는 <비즈니스-경제>, <정치-법률>에서와 같이 검색된 수준 1이외의 분야연상어 수가 대단히 많은 경우는 같은 부모분야 내에 다른 분야로 오인식 되는 경우가 있어 정확률의 저하를 초래한다. 이것은 6장의 향후 연구에서 논의할 것이지만, 점수집계의 문제와 관련이 있다. 수준 3과 4의 분야연상어 점수를 가볍게 하는 전략이 필요하다고 생각된다. 한편, <스포츠>와 같이 질·양적으로 잘 정돈된 분야연상어가 구축되어 있는 분야에 대해서는 정확률의 차가 크게 보이지 않는다. 결론적으로 수준 1의 분야연상어가 많이 존재하는 분야는 본 논문의 방법으로 단락의 분야를 정확히 추출할 수 있다는 증거이다.



(a) Precision and Recall for Training Set



(b) Precision and Recall for Test Set

(그림 11) 본 방법의 정확율과 재현율(Training and Test Set)

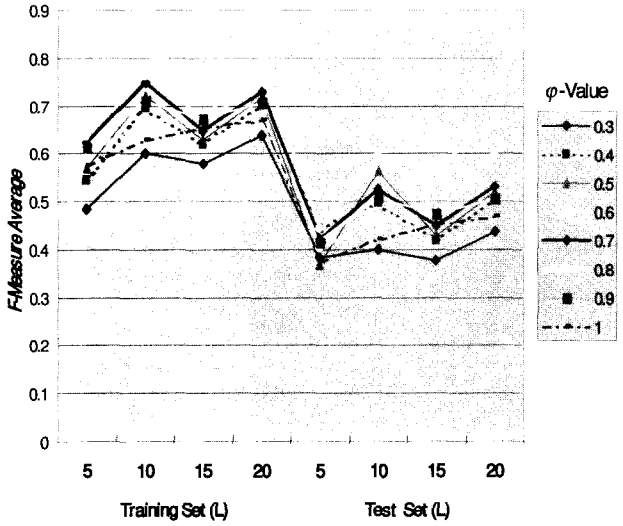
정확률과 재현율의 결과를 토대로 Test Set에 적용할 최적의 파라미터를 찾는다. 다음의 식 (2)를 이용하여 *F-Measure*를 계산하였다. <표 5>는 영어문서의 정확률(P), 재현율(R), F-Measure 값(F)의 결과이다. 각 분야별로 위에 굵은 글씨로 쓴 수치가 Training Set을 이용하여 계산한 수치이고, 밑에 이탤릭으로 쓴 값은 Test Set에 대한 계산결과이다. Training Set의 경우 L = 20인 경우와 Test Set인 경우 L = 10과 20에서 F-Measure 값이 가장 높게 나타났으며, 한국어 문서에 대해서도 동일한 결과를 얻었다.

$$F\text{-}Measure = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \qquad (2)$$

영어문서 Training Set에 대하여 평균 F-Measure의 값에 의해 최적의 ρ의 값을 (그림 12)에 Training과 Test Set

에 적용한 예를 표시하였다. 두 가지 셀 모두 L의 값이 20일 때와 ρ의 값은 0.8일 때, 한국어의 경우 ρ의 값이 0.6~0.8일 때 최적의 결과를 얻었다.

본 논문의 방법은 동일한 화제에 대한 단락을 추출할 때 각 문장에서 분야연상어를 검색하여 분야별 점수계산을 수행하고 그 결과를 이용하여 가능한 모든 분야마다 화제의 출현·전환·계속을 판단하므로 처리시간이 이슈가 될 수 있다. 한국어 Training Set에 대한 처리시간을 <표 6>에 표시하였다. 실험 환경은 Intel Pentium IV, 3.3 GHz CPU, 1GB RAM을 장착한 컴퓨터에서 수행하였으며, 처리시간은 모두 수초이내가 되어 충분히 실용적임을 알 수 있다.



(그림 12) F-Measure 값에 의한 최적의 ρ값의 결정(영어문서, Training & Test Set)

〈표 6〉 Training Set의 처리시간

L	파일 (KB)		처리 시간 (sec.)		
	수	크기	1KB	10KB	파 일
5	32	85.6	0.5899	0.006891	0.068914
10	32	142	0.9587	0.006751	0.067514
15	32	201	1.2431	0.006185	0.061846
20	32	253	1.2233	0.004835	0.048352
Total	**128**	**681.6**	**3.4701**	**0.005091**	**0.050911**

5.3 다른 방법과의 비교실험

본 절에서는 기존의 단락검색 방법과의 비교실험을 수행한다. 참고문헌 [9]의 모지즈키 등이 제시한 어휘적 연쇄(Lexical-chain)를 이용한 방법과 비교한다. 비교실험을 위해 Training Set을 다음과 같이 실험 A와 B로 나누어 선정하였다.

실험 A: 화제의 차이가 큰 문서에서의 비교,
실험 B: 유사한 분야로 된 문서에서의 비교.

실험 A는 정확률과 재현율을 분야트리의 탑노드(화제의 차이가 큰 분야의 문서)에서의 비교이고, 실험 B는 화제가 비슷한 <스포츠/구기>의 8개의 종단분야 중에서 화제분야를 혼합하여 32개의 파일을 작성하였다.

2장에서 서술한 바와 같이 각 분야연상어의 수준에 따라 점수를 일률적으로 부여하였으며, 계산식[참고문헌 [21] 참고]에서 서술한 쇠퇴도(Dec_j)에 영향을 주는 ρ의 값은 0.6으로 실험하였다. 어휘적 연쇄길이[9]의 임계치 = 전체단어수/32 그리고 Gap 길이의 임계치 = 전체단어수/8로 설정하였다. (그림 11(a))는 평가문서에서 수집된 각 수준별 분야연상어 수를 (그림 11(b))는 검색된 분야연상어 수를 나타낸다. 그래프의 검은선은 본 시스템의 정확률과 재현율을 나타내고, 흰선은 어휘적연쇄를 이용한 방법의 정확률과 재현율을 나타낸다.

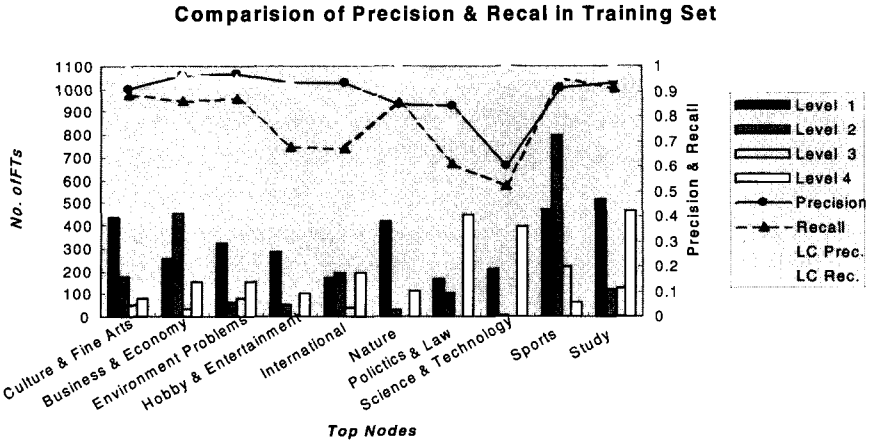
5.3.1 화제의 변화가 큰 문서에서의 비교실험 : 실험 A

(그림 13(a))는 Training Set을 이용한 경우의 비교실험이다. 비교방법(모지즈키의 방법)의 평균 정확률과 재현율이 각각 0.93, 0.93인 것에 비하여, 본 방법은 각각 0.88과 0.78이었다. 비교방법이 평균 정확률과 재현율 모두에서 양호한 결과를 얻고 있다. 이유는 본 논문의 방법은 검색된 분야연상어가 각 분야마다 수준별로 차이가 크게 나타남에도 불구하고 분야연상어의 수준별 점수를 모든 분야에 동일한 점수를 설정하였기 때문이다.

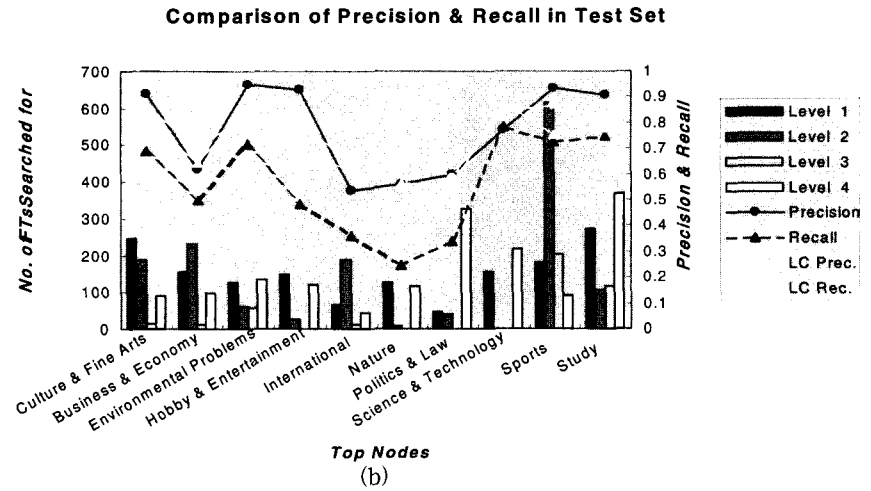
본 방법은 분야연상어 수가 적은 <정치-법률>은 계속도가 쉽게 저하되어 단락 형성 시 중간에서 끊어지는 경향이 많지만, 비교방법은 점수를 고려하고 있지 않기 때문에 이런 현상이 보이지 않는다. <문화-예술>과 <스포츠>와 같이 구축된 분야연상어 수가 많은 분야도 정확률이 낮은 이유는 분야연상어 수가 많아 점수집계의 값이 크게 되고, 화제분야의 계속도가 매우 높아져 분야연상어 수가 상대적으로 적은 분야의 전환도를 흡수하는 경향을 보인다.

(그림 13(b))는 Test Set을 이용한 경우의 비교이다. 비교방법의 평균 정확률과 재현율이 0.77, 0.56에 비하여 본 방법은 각각 0.76, 0.71이다. 마찬가지로, 비교방법의 재현율이 상위한다. 그 이유는 어휘적 연쇄를 이용한 방법은 분야트리의 탑노드에서 수준 2, 3, 4의 분야연상어와 수준 1의 분야연상어가 동등한 역할을 하지만, 본 논문의 방법은 분야의 차이가 큰 문서에서는 수준 2와 3의 분야연상어가 비교적 적게 존재하기 때문에 2절의 마지막과 4절에서 설명한 수준별 점수집계가 별로 중요한 역할을 하지 못하고 있다.

요약하면, 신문기사는 전혀 다른 분야의 문서가 계속되는 것보다는 비교적 유사한 분야의 문서가 계속된다. 따라서, 다음의 [실험 B]와 같이 유사한 분야의 문서에서의 비교가 필요하다.



(a)



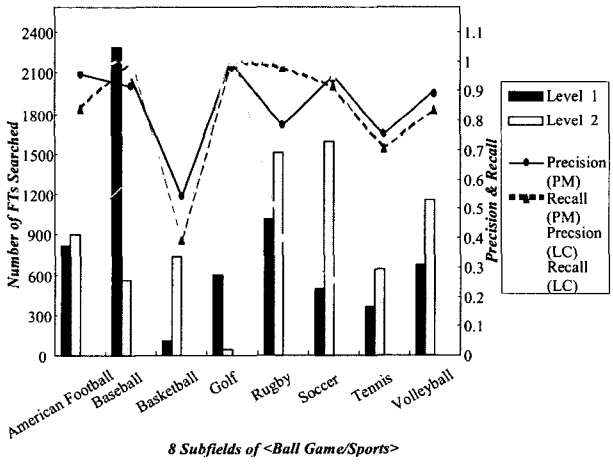
(b)

(그림 13) 다른 방법과의 비교실험

5.3.2 화제가 유사한 문서에서의 비교실험 : 실험 B

유사한 문서에서의 비교실험을 위해 <스포츠/구기>의 하위분야 중 8개의 종단분야에서 추출한 새로운 Training Set을 선정하여 (그림 14)와 같이 분야연상어가 검색되었다. 여기서, 분야연상어의 수준은 1과 2만을 검색하였다. 수준 1의 분야연상어 수가 높은 <야구>는 정확률과 재현율이 각각 0.91과 0.99로 비교방법보다 비교적 높은 수치가 얻어지고 있다. <축구>, <럭비>에 대해서는 수준 2의 분야연상어 추출율이 높고, 높은 정확률과 재현율이 얻어지고 있다. 당연한 결과로 분야 <농구>에서 정확률과 재현율이 낮은 이유는 수준 1의 분야연상어 수가 매우 낮기 때문이다. 수준 1의 분야연상어 추출이 별로 높지 않은 <골프>가 정확률과 재현율 모두 높은 이유는 수준 2의 분야연상어가 거의 검색되지 않았지만, 화제분야의 출현·계속·전환이 비교적 효과적으로 이루어졌기 때문으로 생각된다.

다시, 비교방법과 비교해 보면, 본 방법은 평균정확률과 재현율이 각각 0.67, 0.67인 것에 비해 비교방법은 0.41, 0.79이었다. 본 방법이 재현율은 떨어지지만 정확률은 우세하다. 어휘적연쇄를 이용한 방법은 유사한 분야로 구성된 문서에서는 수준 2의 분야연상어가 많이 존재하기 때문에 단락의 분야가 특정한 하나의 분야를 한정하지 않고, 넓은 분야의 문장집합(단락에 해당)으로 형성되기 때문이다. 그 때문에 재현율은 높아지지만 정확률이 낮아졌다.



(그림 14) 유사한 분야에서의 비교실험

결론적으로 본 논문의 방법은 정확하게 화제분야를 결정하고, 각 문장사이에서 단락의 구간분리가 성공적으로 이루어진다. 특히, 화제의 차이가 큰 문서에서 정확한 화제의 결정이 잘 동작함을 실험을 통해 알 수 있었다.

6. 결 론

단락검색은 사용자의 질의어에 대해 정확하고 빠르게 동작하고, 잘못 검색된 정보를 빠르게 차단하며, 최소한 사용자가 원하는 정보의 존재여부를 지시한다. 문서내 화제의 계속성과 전환성에 기반 한 단락검색은 가공되지 않은 자연어 문장에서 관련된 정보를 추출하는 가장 유용한 방법이다. 본 논문의 방법은 텍스트의 특정 화제분야를 대표하는 실마리로서 분야연상어를 이용하였기 때문에 인간의 두뇌인식과 유사하게 컴퓨터가 텍스트를 읽어감에 따라 텍스트가 어느 분야에 속하는지를 판단하는 방법이며, 화제의 전환성과 계속성을 고려하였기 때문에 텍스트가 분리되는 현상을 방지하고, 복수분야에 속하는 텍스트의 중복을 제거하는 새로운 단락검색 방법이다.

차후에 본 시스템을 개선할 사항에 대하여 서술하면, 첫째로 5장에서 언급한 바와 같이 한국어의 경우 <교육>, <건강&의료>, <산업>, <사회생활> 등과 영어의 경우 <문화&예술>, <교육>, <산업>, <사회생활>, <학문> 등 아직 충분히 분야연상어를 수집하지 못한 분야에 대한 계속적인 수집이다. 복수분야를 연상하는 분야연상어는 각 분야마다 출현빈도가 다르기 때문에 같은 수준의 분야연상어라도 출현빈도에 따라 다른 가중치를 부여하면 재현율을 높일 수 있다.

둘째로, 본 논문에서는 각 수준에 분야연상어의 점수나 파라미터를 하나로 결정하여 연구를 진행하였으나, 대량의 코퍼스에서 자동으로 학습하는 방법도 생각할 수 있다. 아울러, 분야연상어의 상속성에 대해 깊이 연구하고자 한다. 셋째로, 단어사이는 공기관계가 존재하기 때문에 분야연상어에 공기정보 혹은 격 정보를 부여하여 수준 1이외의 분야연상어가 복수 출현한 경우에도 한 가지 뜻으로 분야를 연상시키는 방법에 관한 연구를 생각할 수 있다. 분야연상어 이외에도 문서 내에 존재하는 연결구나, 접속사구, signal 단어에 대하여 연구하고자 한다.

연상어와 (그림 10) (b)에서와 같이 검색된 분야연상어 사이의 비율이 같기 때문에 점수집계의 문제가 발생한다. 따라서 분야별·수준별로 분야연상어의 점수에 차를 주어 정확률의 향상을 도모한다. 또한 어떤 단락은 <취미 & 오락/식도락>이 정확한 분야이지만, <취미 & 오락> 혹은 <스포츠/구기/야구> 등을 연상하는 분야연상어 "베이스"와 수준 2의 5개의 분야연상어(두 번 출현한 분야연상어 "요리", "술", "프랑스요리", "민속음식" 등)에 의해 <취미 & 오락/식도락>과 <취미 & 오락/요리> 등 네분야가 후보분야이지만 점수집계결과 40 포인트를 얻은 후자의 두 분야가 가능하다. 이 단락은 수준 1의 분야연상어가 검색되지 않아 최고점수를 갖는 분야가 두 분야 존재하고, 본 시스템의 출력으로 오직 한 개의 분야를 결과로 출력하도록 하면 '분야미정'으로 출력되어 버린다. 또한 뒤에서 <야구> 이외에 수준 1의 분야연상어가 출현하여도 앞에 <야구>에 대한 텍스트가 길게 계속되어 많은 분야연상어가 출현하였기 때문에 계속도가 너무 높아져 뒤에 <스키>에 관한 텍스트가 출현하여도 <스키>에 대한 전환도 값이 <야구>의 계속도를 능가하지 못하는 경우의 예이다. 이러한 문제점은 미래에 계속 연구되어야 할 사항이다.

다섯째, 앞에서 언급한 바와 같이 현재의 화제분야 F_{theme}의 전환도의 감소폭과 새로운 분야 F_k의 전환도의 증가폭 사이의 관계에 대하여 깊이 연구하고자 한다.

다음에 언급하는 세 가지는 단락검색의 적용에 관한 연구이다. 여섯째, 복합 분야연상어의 조합규칙을 검토해서 두 개 이상의 분야연상어가 서로 짝을 이루어 전혀 다른 분야를 강하게 한정하는 복합 분야연상어로 합성되는 경우에도 이들 모두를 사전에 등록하지 않고 규칙에 의해 합성하여 검색하도록 하였고, 동시에 그 개수를 줄이는 방법이 필요하다. 또한 분야연상어를 이용해 단락을 bottom-up approach로 문서요약에 적용하고자 한다. 일곱째, 본 논문의 실험에서 얻어진 점수 계산결과와 문서의 분야별로 추출된 각 단락의 분포를 조사하여 유사문장 혹은 예제문장 검색에 응용하고자 한다. 마지막으로, 다른 언어에 대해서도 분야연상어를 구축하여 다언어 정보검색(Cross-Language Information Retrieval)에 응용하고자 한다.

참 고 문 헌

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

[2] Cormack, G. V., Clarke, C. L. A., Palmer, C. R., and Kisman, D. I. E., Fast Automatic Passage Ranking (MultiText Experiments for TREC-8). The Eighth Text Retrieval Conference(TREC-8), 1999.

[3] Cormack, G. V., Clarke, C. L. A., Palmer, C. R., and To, S. S. L., Passage-Based Refinement(MultiText Experiments for TREC-6). The Sixth Text Retrieval Conference(TREC-6), 1997.

[4] Cormack, G. V., Clarke, C. L. A., Palmer, C. R., and To, S. S. L., Passage-based Query Refinement (MultiText Experiments for TREC-6). An International Journal of Information Processing and Management. Vol.36, No.1, pp.133-153, 2000.

[5] Daniels, J. J., Retrieval of Passages for Information Reduction. Doctoral Thesis. University of Massachusetts Amherst, MA, USA, 1997.

[6] Kaszkiel, M., Zobel, J., and Sacks-Davis, R., Efficient

Passage Ranking for Document Databases. ACM Transactions on Information Systems. Vol.17, No.4, pp.406-439, 1999.

[7] Knaus, D., Mittendorf, E., and Schauble, P., Improving a Basic Retrieval Method by Links and Passage Level Evidence. The Third Text Retrieval Conference (TREC-3), 1994.

[8] Kretser, O., and Moffat, A., Efficient Document Presentation with a Locality-Based Similarity Heuristic. The Proceedings of the 22nd Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval(SIGIR 1999), 1999.

[9] 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)

[10] Myaeng, S. H, Jang, D. H., Kim, M. S., and Zhoo, Z. C., A Flexible Model for Retrieval of SGML Documents. The Proceedings of the 21st Annual International ACM Special Interest Group Information Retrieval Conference on Research and Development in Information Retrieval(SIGIR 1998), 1998.

[11] Samuel Sangkon Lee, Masami Shishibori, Toru Sumitomo, and Jun-Ichi Aoe, "Extraction of Field-coherent Passages," Information Processing & Management, Vol.38, No.2, pp.173-207, 2002.

[12] Sangkon Lee and Masami Shishibori, "Passage Segmentation based on Topic Matter," International Journal of Computer Processing of Oriental Languages, Vol.15, No.3, pp.305-339, 2002.

[13] Sangkon Lee, Masafumi Koyama, Shoji Mizobuchi, Kyoko Uchibayashi, Fumihiko Kawano, Takahiro Komatsu, and Jun-ichi Aoe, "Cross-language Multi-media Information Retrieval System : BOSS," Proceedings of the Eighteenth International Conference on Computer Processing of Oriental Languages, Vol.1, pp.245-248, 1999.

[14] Sangkon Lee, Masami Shishibori, Kazuhiro Morita, and Jun-ichi Aoe, "Passage Retrieval based on Topic-Matter," The 19th International Conference on Computer Processing of Oriental Languages, Vol.1, pp.193-198, 2001.

[15] Shoji Mizobuchi, Sangkon Lee, Fumihiko Kawano, Tsuyoshi Kobayashi, Takahiro Komatsu, and Jun-ichi Aoe, "Multi-lingual Multi-media Information Retrieval System," NTCIR Workshop I, Vol.1, pp.171-178, 1999.

[16] Wilkinson, R., Effective Retrieval of Structured Documents. The Proceeding of 17th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval Research (SIGIR 1994), 1994.

[17] Williams, M., An Evaluation of Passage-Level Indexing Strategies for a Technical Report Archive. LIBRES : Library and Information Science Research Electronic Journal. Vol.8, No.1, pp.194-218, 1998.

[18] Yamamoto, K, Masuyama, S., and Naito, S., Experimental Study on Paragraphing Japanese Sentences Using Cue Words. The Proceedings of the First Annual Meeting of the Association for Natural Language Processing. Vol.84-9, 1991(in Japanese).

[19] Yang, K., Maglaughlin, K. L., and Newby, G. B., Passage Feedback with IRIS, An International Journal of Information Processing and Management. Vol.37, No.3, pp.521-541, 2001.

[20] Zobel, J., Moffat, A., Wilkinson, R., and Sacks-Davis, R., Efficient Retrieval of Partial Documents. An International Journal of Information Processing and Management. Vol.31, No.3, pp.361-377, 1995.

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

[22] 이상곤, 이완권, "분야연상어의 수집과 추출 알고리즘," 정보처리학회논문지 B, 제10권, 제3호, pp.347-358, 2003.

[23] 자유국민사, 현대용어의 기초지식, 1997(in Japanese).

이 상 곤

e-mail : samuel@jj.ac.kr

1994년 전주대학교 영어영문학과(학사)

1996년 전북대학교 컴퓨터과학과(학사)

1998년 전북대학교 전산통계학과(이학석사)

2001년 일본 도쿠시마대학교 지능정보공학과(공학박사)

2001년 전주 우석대학교 정보통신 및 컴퓨터공학부, 전북대학교 컴퓨터과학과 시간강사

2001년~2002년 원광대학교 음성정보 기술산업 지원센터 연구원

2002년~현재 전주대학교 정보기술공학부 컴퓨터공학 전공 조교수

관심분야 : 자연언어처리, 한국어 정보처리, 이메일문서처리, 한글공학, 정보검색, 문서분류, 문서요약