화제인식에 의한 단락별 계산방법의 설계

김혜경[1], 이상곤[2]
전주대학교 교육대학원 컴퓨터교육전공[1]
전주대학교 정보기술공학부 컴퓨터공학전공 언어과학실[2]
{syky79[1], samuel[2]}@jj.ac.kr

Design of Passage Calculation Method based on Topic Recognition

Hye-Kyung Kim[1], and Samuel Sangkon Lee[2]
Dept. of Computer Education, Graduate School of Education[1],
Language Science Lab., Dept. of Computer Science & Engineering,
School of Information Technology & Engineering[2],
Jeonju University[1,2]

요 약

화제가 혼합되어 있는 문서에서 각 화제의 단락을 추출하면 사용자의 질의어에 일치하는 정보만을 추출할 수 있다. 정확하고 빠르게 사용자의 검색요구에 일치하는 관련 정보를 추출할 수 있다. 본 논문에서는 문서에서 사용자의 요구에 적합한 단락을 추출하는 기술을 설명한다. 문서에서 분야연상어를 추출하고, 각 문장마다 화제분야의 출현·계속·전환이 어떻게 변화하여 가는지를 추적하여 계산한다. 긴 문서에서 어떤 화제가 출현하는가를 파악하고, 화제가 계속되거나 혹은 전환되는 지점을 인식하여, 분야별 단락을 추출하는 방법을 제안한다.

1. 서론

긴 문서는 여러 화제가 혼합되어 있기 때문에 각 화제의 경계부분을 구분할 수 있는 단락검색을 이용하면 사용자의 검색요구에 적절한 검색을 신속하게 수행할 수 있다. 단락검색은 사용자의 검색요구에 밀접하게 관련 있는 텍스트와 의미적인 실마리를 가장 많이 포함하고 있는 부분적인 텍스트를 검색하여 사용자의 검색요구에 적합한 검색을 수행할 수 있는 기술이다.

본 논문에서는 기존연구에서 진행되어 온 단일 혹은 복합 분야연상어[2-7]를 이용하여 문서에서 출현하는 화제의 범위를 빠르게 파악하고, 화제의 경계부분을 파악하여 각 화제의 범위를 결정한다. 화제흐름의 특징을 조사하여 분야연상어의 연속된 출현율을 토대로 산출된 화제의 계속성과 전환성[1]을 계산하여 사용자의 검색요구에 적합한 의미 있는 단락을 추출하고자 한다.

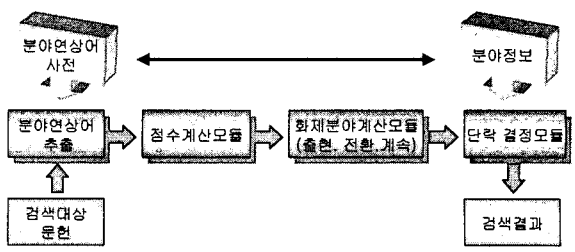
이하, 제 2장에서는 본 논문의 이론적 배경이 되는 분야연상어와 화제분야의 추적엔진의 개요도에 대하여 설명하고, 3장에서는 신문기사에서 수집한 문서를 대상으로 분야연상어를 이용하여 문장마다 화제 분야의 출현·계속·전환이 어떻게 변화하여 가는지를 측정하는 계산법을 중심으로 설명한다. 마지막으로 결론과 향후 과제를 서술한다.

2. 화제분야

분야연상어[4-7]란 문서에서 "스파이크"라는 단어가 출현하면 "배구"의 분야가 연상되듯이 인간이 단어를 보는 것만으로 분야를 직관적으로 연상할 수 있는 단어이다. 분야연상어는 다른 분야의 문서와 구별할 수 있는 어휘레벨의 정보이며, 문서의 분야를 파악하는데 유일한 단서가 되는 단어이다. 이러한 분야연상어를 이용하면 신속하고 정확하게 문서의 분야를 인식할 수 있으며, 화제별 단락을 추출할 수 있다.

2.1 화제분야의 추적

본 논문에서 제안하는 화제분야 추적엔진의 개요를 다음의 (그림 1)에 표시하였다.



(그림 1) 화제분야 추적엔진의 개요

분야연상어 추출은 각 문장에서 존재하는 모든 분야연상어를 AC방법[8]을 이용하여 추출한다. AC방법은 복수의 키워드를 검출할 수 있는 문자열 탐색 알고리즘이다.

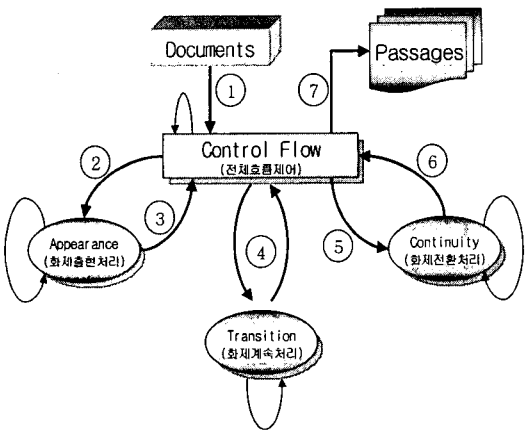
점수계산(score calculation) 모듈은 추출된 분야연상어는 각 수준[5]에 따라 분야를 한정하는 정도가 다르기 때문에 수준 1을 12, 수준 2를 8, 수준 3을 4, 수준 4를 2점으로 부여한다. 다음의 <표 1>과 같이 추출된 분야연상어의 각 수준별 점수를 부여하여 각각의 문장에서 분야별 점수를 합산한다.

<표 1> 점수집계의 예

d		분야별		
		F_1(골프)	F_2(농구)	F_3(배구)
문서	S_1	20	-	-
	S_2	42	-	-
	S_3	54	-	-
	S_4	8	-	-
	S_5	-	36	-
	S_6	-	56	-
	S_7	-	24	-
	S_8	-	-	44
	S_9	-	-	12
	S_{10}	-	-	96

단락의 결정[2]모듈은 각 분야마다 얻어진 점수를 이용하여 전환도·계속도를 계산하고, 이를 이용하여 특정화제의 출현·전환·계속처리를 수행한다. 이는 문서에서 검색요구와 관련이 깊은 분야 혹은 특정 화제별로 단락을 분할하여 검출할 수 있다.

3. 화제변화의 계산방법



(그림 2) 단락결정 알고리즘의 전체흐름도

본 논문에서 제안하는 (그림 2)의 단락결정 알고리즘의 전체흐름도에 따라 어떤 문서(d)가 2절에서 설명하는 방법에 의해 앞의 <표 1>과 같은 점수집계 결과를 가지면 ① 전체흐름제어에서 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-Topicp에는 현재 화제분야 (F_{theme})의 변경을 파악하기 위한 플래그로서 Null값으로 초기화된다.

전체흐름제어 알고리즘에 의해 현재 처리할 문장의 분야(F_1)가 결정되고, (F_{theme})이 "Field-Neutral"이므로 첫 번째 문장의 분야연상어 점수집계(Freq(s_1,F_1))를 전환도($\beta_1(F_1)$)값에 대입한다. ② 화제출현처리로 분기한다. 전환도($\beta_1(F_1)$)의 값이 0이 아니고 전환도의 값이 2개 이상이 아닌 한 개이므로 최대치를 갖는 화제분야(F_1)가 (F_{theme})이 되고 Old-Topic이 된다. 전환도($\beta_1(F_1)$)의 값이 계속도($\alpha_j(F_{theme})$)의 값으로 대입되고 첫 번째 문장(s_1)이 stack에 push 된다. ③ 전체흐름제어로 되돌아와 (F_{theme})이 "Field-Neutral"이 아니므로 계속도($\alpha_2(F_{theme})$)를 다음과 같은 방법으로 계산한다. 계속도($\alpha_2(F_{theme})$)를 계산하기 위해 두 번째 문장에서의 쇠퇴율 Dec_2[1]을 먼저 계산한다.

$$Dec_2 = -1 \times \frac{\left([\sum_{S_k \in C_i} Freq(S_1, F_{theme})] + Freq(S_2, F_{theme})\right)}{num(C_i)+1}$$

$$= -1 \times \frac{[20]+42}{1+1} = -1 \times \frac{62}{2} = -31$$

$$\alpha_2(F_{theme}(=F_1)) = \alpha_1(F_{theme}) + [\rho \times Dec_2] + Freq(S_2, F_{theme})$$
$$= 20 + [0.8 \times (-31)] + 42$$
$$= 20 + [-24.8] + 42$$
$$= 37.2$$

나머지 분야의 전환도 ($\beta_2(F_2)$)와($\beta_3(F_3)$)값을 다음과 같이 계산한다. 계속도와 동일하게 쇠퇴율 (Dec_2)을 먼저 계산하여 그 결과를 전환도 계산식에 대입한다.

$$Dec_2 = -1 \times \frac{\left([\sum_{S_k \in C_i} Freq(S_1, F_2)] + Freq(S_2, F_2)\right)}{num(C_i)+1}$$

$$= -1 \times \frac{[0]+0}{1+1} = -1 \times \frac{0}{2} = 0$$

$$\beta_2(F_2) = \beta_1(F_2) + [\rho \times Dec_2] + Freq(S_2, F_2)$$
$$= 0 + [0.8 \times 0] + 0$$
$$= 0$$

F_3의 경우도 마찬가지로 다음과 같은 방법으로 계산한다.

$$Dec_2 = -1 \times \frac{\left([\sum_{S_k \in C_i} Freq(S_1, F_3)] + Freq(S_2, F_3)\right)}{num(C_i)+1}$$

$$= -1 \times \frac{[0]+0}{1+1} = -1 \times \frac{0}{2} = 0$$

$$\beta_2(F_3) = \beta_1(F_3) + [\rho \times Dec_2] + Freq(S_2, F_3)$$
$$= 0 + [0.8 \times 0] + 0$$
$$= 0$$

<표 2> 화제의 출현 · 전환 · 계속도의 계산 예

Freq(S_j, F_k)				$\beta_j(F_k)$						Old Topic	$\alpha_j(F_{theme})$		F_{theme}	j'	stack	Passage(d, F_k)
s_j	F_1	F_2	F_3	β_1	(F_1)	(F_2)	(F_3)	max	(F_{theme})							
initial					0	0	0			NULL			Field-Neutral			
s_1	20	-	-	β_1	20			F_1	20	F_1	a_1	20	F_1		s_1	
s_2	42	-	-	β_2		0	0				a_2	37.2			s_1,s_2	
s_3	54	-	-	β_3		0	0				a_3	60.32			s_1,s_2,s_3	
s_4	8	-	-	β_4		0	0				a_4	43.52			s_1,s_2,s_3,s_4	
s_5	-	36	-	β_5		30.24	0	F_2			a_5	23.68			s_1,s_2,s_3,s_4	s_1,s_2,s_3,s_4
														5	s_1,s_2,s_3,s_4	
															Clear	
. . .				. . .												
					0	0				F_3					s_8	
s_9	-	-	12	β_9	0	0					a_9				s_8,s_9	
s_{10}	-	-	96	β_{10}	0	0					a_{10}				s_8,s_9,s_{10}	
Final																s_8,s_9,s_{10}

$(\beta_2(F_2))$와$(\beta_3(F_3))$의 전환도 중 최대치(max)가 $(\alpha_2(F_1))$와 비교하여 계속도의 값이 크면 ④ 화제계속(Transition)처리로 분기하고, 전환도의 값이 크면 ⑤ 화제전환(Continuity)처리로 분기하여 수행한다.

본 논문에서 제안하는 단락결정 방법을 이용하여 단락 사이의 경계를 해석하고, 각 분야별 단락을 추출할 수 있다. 실행 예를 <표 2>에 설명하였다(단, $\rho = 0.8, \alpha_{theme} = 0$ 으로 계산, F_k 의 k는 1에서 3으로 가정한다).

4. 결론

본 논문에서 제시하는 방법은 사용자의 질의어에 적합한 단락을 제시하며, 빠르고 정확하게 동작한다. 또한 잘못 검색된 정보를 효과적으로 차단하는 방법이다. 본 논문에서 제시하는 방법을 정보 검색 엔진에 적용하면 사용자에게 원하는 정보의 존재여부를 지시하고, 문서 내 화제의 계속성과 전환성을 계산하여 가공되지 않은 자연어 문장에서 관련 정보를 추출하는 유용한 방법이다.

결론적으로 문서의 화제분야를 대표하는 분야연상어를 이용하였기 때문에 인간의 두뇌와 유사하게 컴퓨터가 텍스트를 읽어감에 따라 텍스트가 어느 분야에 속하는지를 빠르게 판단할 수 있다. 향후에는 제안된 방법을 이용하여 실험하고 그 결과에 대한 정확율과 재현율에 대해 조사하여 텍스트가 분리되는 현상을 방지하고, 복수분야에 속하는 텍스트의 중복을 제거하는 실용적인 시스템을 디자인하고자 한다.

참고문헌

[1] 이상곤, "분야연상어를 이용한 화제분야의 계산방법과 단락검색", 정보처리학회논문지(B), 제 12권, 제 1호, pp. 57-68, 2005.

[2] 이원휘, 김도연, 이상곤, "그래픽컬한 분야인식기의 설계 및 구현", 한국정보과학회 가을 학술발표 논문집, 제 31권, 제 2호, pp. 769-771, 2004.

[3] 이원휘, 최현, 이상곤, "분야연상어 추출방법의 설계와 구현", 한국정보처리학회 2004년도 춘계 학술발표 논문집, 제 11권, 제 1호, pp. 651-654, 2004.

[4] 최현, 황남선, 이상곤, "문서분류용 목적으로 이용할 효율적인 연상정보의 추출방법", 2004년 봄 한국정보과학회 학술발표 논문집(B), 제 31권, 제 1호, pp. 892-894, 2004.

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

[6] 김숙영, 이상곤, "문서분류용 분야연상어의 추출 알고리즘", 2003 한국정보과학회 호남-제주 지부 학술발표 논문집, 제 15권, 제 1호, pp. 120-124, 2003.

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

[8] 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.