한국컴퓨터정보학회 동계학술대회 논문집 제23권 제1호 (2015. 1)

키워드 가중치 방식에 근거한 도서 본문 주제어 추출

안희정°, 최건희*, 김승훈**
°단국대학교 멀티미디어공학과
*단국대학교 소프트웨어학과
**단국대학교 응용컴퓨터공학과
e-mail:dreaminghee90@gmail.com°, rjsgmlgood@naver.com*
,edina@dankook.ac.kr**

Thematic Word Extraction from Book Based on Keyword Weighting Method

Hee-Jeong Ahn°, Gun-Hee Choi*, Seung-Hoon Kim**
°Dept. of Multimedia Engineering, DanKook University
*Dept. of Software Science, DanKook University
**Dept. of Applied Computer Engineering , DanKook University

● 요 약 ●

본 논문에서는 문장 및 문단에서 키워드의 역할에 따른 가중치에 근거하여 도서 본문에서 주제어를 추출하는 방법을 제안한다. 기존의 주제어 추출 방식은 도서 본문이 아닌 신문이나 논문에 대한 방식이므로 도서 본문에서의 주제어 추출에 그대로 적용하기에는 어려움이 있다. 따라서 본 논문에서는 빈도수뿐만 아니라 문장 내 중요 요소에 대한 가중치와 중요 문장에 대한 가중치를 후보 키워드에 부여하는 방식을 제안하였다. 제안한 계산 방식을 비문학 도서에 대하여 실험한 결과, 빈도수만으로 주제어를 추출한 기존 방식보다 본 논문에서 제안한 방식의 주제어 추출 결과의 정확도가 향상되는 것을 확인하였다.

키워드: 도서(book), 텍스트마이닝(text mining), 주제어(thematic word), 추출(extraction)

I. 서 론

인터넷의 발달로 전자문서의 증가와 함께 주제어 추출을 위한 연구가 활발히 진행되고 있다. 도서 정보에 대한 디지털화가 진행되면서 온라인 서점을 통한 전자책의 개수와 그 수요가 증가되고 있다. 하지만 고객들이 도서를 검색 시, 검색 키워드로는 제목과 저자만을 이용하여 검색하고 있으며, 도서 자체를 표현할 수 있는 주제어를 제공하는 추천서비스 또는 검색서비스는 부족한 현실이다. 또한 기존의 주제어 추출 연구들은 대부분 빈도수를 기준으로 하고 있으며, 신문과 논문이 아닌 도서에 대한 주제어 추출 연구는 미흡하다. 따라서 기존의 빈도수를 기반으로 하는 주제어 추출을 보완하는 새로운 방식으로 도서 본문에 적합한 주제어를 추출하는 연구가 필요하다.

본 논문에서는 이러한 문제점을 개선하기 위해 주제어 추출 시, 문장 내의 중요 요소에 대한 가중치와 중요 문장 안에서의 단어에 대한 가중치를 주는 방식을 제안한다.

II. 관련 연구

1. 주제어 추출 기법

문서 내의 주제어 추출에 대한 연구에는 통계적 기법인 단어의 빈도수(TF: Term Frequency)와 역문서 빈도수(IDF: Inverse Document Frequency)를 이용한 키워드 추출 연구[1,2,3]와 언어학적 접근 방법으로 구문 분석을 이용하여 문장 내의 '주어'를 중심으로 주제어를 추출한 연구[4]가 있다. 하지만 문서 내 중요 단어를 포함하고 있는 문장이나 문장 위치 중요도에 대한 내용은 대부분 포함하지 않기 때문에, 본 논문에서는 주제어 추출에 있어 통계적, 언어적 기법에 문장의 중요도도 고려하여 주제어를 추출하는 방식을 제안하였다.

2. 문장 중요도

문장 내의 중요도를 판단할 수 있기 위해서는 문서의 구조를 먼저 파악해야한다. 글의 내용구성은 흔히 국어국문학에서 언급하는 단락의 구성과 연관된다. 두괄식은 단락의 첫 부분에, 미괄식은 단락의 마지막 부분에 핵심적인 내용을 담아낸다[5].

그림 1. 두괄식과 미괄식 문단 구조
Fig. 1. Deductive & Inductive Structure

정보 제공을 우선으로 하는 신문의 경우에도 주제문 추출 시 문서의 앞부분이나 뒷부분에 주제문이 분포하는 것을 알 수 있다[6]. 따라서 본 논문에서는 문단의 기본적 구조인 두괄식과 미괄식에 관한 중요도를 포함한 방식을 제안하였다.

III. 본 론

1. 시스템 구성

본 논문 시스템의 전체적인 구성은 그림 2와 같다. 시스템은 세 단계로 진행한다. 먼저, 도서 본문을 형태소 분석기를 사용하여 명사만을 추출한다. 여기서 추출된 명사는 일반 명사와 고유 명사이다. 두 명사만을 지정한 것은 한글 형태소 품사 태그표 안의 명사 종류에 의존 명사, 수사와 대명사와 같은 불용어가 포함되었으므로 일반 명사와 고유 명사만을 지정하여 추출한다. 두 번째 단계로 각 키워드의 빈도수와 단어, 문장에 따른 가중치 값을 계산한다. 마지막으로 주제어 추출 단계는 앞에서 계산한 빈도수와 가중치 값을 더하여 상위에 위치한 키워드를 주제어로 추출한다.

그림 2. 제안한 시스템 구조
Fig. 2. Proposed System Architecture

2. 빈도수 기반 키워드 추출

도서 본문 내 단어들의 TF(Term Frequency)를 구하는 식은 (식 1)과 같다. 임의 단어 i에 대해 정규화 된 빈도수를 의미하는 TF_i는 단어 i의 출현 빈도수 n_i에 도서 본문 안의 모든 단어의 총 출현 횟수를 나누어 값을 정규화 한다. 이는 도서 본문의 길이에 영향을 받지 않기 위함이다.

$$TF_i = \frac{n_i}{\sum_k n_k} \quad \cdots\cdots\cdots\cdots \text{(식 1)}$$

3. 키워드 가중치 계산

주제어를 추출하기 위한 두 번째 단계로 본 논문에서는 다음과 같은 세 가중치 부여 방식을 제안한다. 먼저, 문장 내의 중요 구성요소, 그리고 중요 문장이라고 생각되는 문단의 처음과 마지막 문장 마지막으로 결론과 요약을 나타내는 부사구가 포함되는 문장 안의 키워드에 가중치를 부여한다.

3.1 문장 내 주요 키워드 추출

한 문장 내에서 문장을 이끌어가는 주어와 목적어 역할을 하는 키워드를 추출한다. 표 1에서 볼 수 있듯이, 주어는 주격조사인 “은”, “는”, “이”, “가”를 기준으로, 목적어는 목적격조사인 “을”, “를”로 구분하였다. 각 조사 앞에 나오는 품사가 일반 명사 또는 고유 명사일 경우 각 키워드마다 개수를 누적한다. 문장 내 키워드가 주어 및 목적어일 정규화 된 빈도수에 가중치를 곱한 값, SO_i는 (식 2)와 같다.

$$SO_i = w_s * \left(\frac{s_i}{\sum_k s_k} + \frac{o_i}{\sum_k o_k}\right) \quad \cdots\cdots \text{(식 2)}$$

단어 i에 대하여 s_i는 단어 i가 주어인 빈도수이며 o_i는 단어 i가 목적어인 빈도수이다. s_i는 도서 본문 내 주어의 총 출현 개수로, o_i는 목적어의 총 출현 개수로 각 값을 정규화 한다. 그리고 두 값을 더해 가중치를 곱해준다. 가중치w_s은 $0 \le w_s \le 1$ 사이의 값이다.

표 1. 한글 형태소 태그표
Table 1. Part Of Speech Tag Table

태그	설명
NNG	일반 명사
NNP	고유 명사
JKS	주격 조사
JKO	목적격 조사

3.2 문단 내 주요 키워드 추출

문단의 구조가 두괄식과 미괄식인 경우의 가중치를 주기위해 첫 번째 문장과 마지막 문장 안의 키워드를 추출한다. 도서 안에서의 문단 구분은 작가가 특정 (1, *, #와 같은)문자로 지정해 놓은 경우도 있지만, 지정해 놓지 않은 경우가 있으므로 문단 구분은 개행 문자의 개수로 구분한다. 주어, 목적어 구분 없이 문장 내의 모든 일반 명사와 고유 명사를 추출하여 각 키워드 마다 빈도수를 계산하고 가중치를 곱해준다. 처음과 마지막 문장 내의 키워드 빈도수에 가중치를 곱한

값, P_i는 (식 3)과 같다. p_i는 단어 i가 문단의 처음과 마지막 문장 안에 출현한 빈도수이며, w_p는 $0 \leq w_p \leq 1$사이의 값이다.

$$P_i = w_p * p_i \quad \cdots\cdots\cdots\cdots \text{ (식 3)}$$

3.3 결론 및 요약 문장 내 키워드 추출

결론과 요약을 나타내는 부사들이 시작되는 문장 안의 키워드를 추출한다. "결론적으로", "결론은", "따라서", "즉" 과 같은 부사가 시작되면 그 다음으로 출현하는 일반 명사와 고유 명사를 추출하여 각 키워드마다 빈도수를 계산하고 가중치를 곱한다. 결론 및 요약을 나타내는 문장 내의 키워드 빈도수에 가중치를 곱한 값, C_i는 (식 4)와 같다. c_i는 단어 i가 결론 및 요약을 나타내는 문장 안에 출현한 빈도수이며, w_a은 앞의 가중치와 마찬가지로 $0 \leq w_a \leq 1$사이의 값이다.

$$C_i = w_a * c_i \quad \cdots\cdots\cdots\cdots \text{ (식 4)}$$

3.4 키워드 가중치 계산

세 방법으로 추출한 값을 더해 총 키워드 가중치 값을 구한다. 키워드의 합 공식은 식 (5)와 같고, 가중치 값의 합은 1이다, $w_s + w_p + w_a = 1$.

$$W_i = w_s * \left(\frac{s_i}{\sum_k s_k} + \frac{o_i}{\sum_k o_k}\right) + w_p * p_i + w_a * c_i \quad \cdots\cdots \text{ (식 5)}$$
$$= SO_i + P_i + C_i$$

4. 주제 키워드 추출

각 키워드의 TF_i값에 키워드 가중치 값, W_i을 더하여 최종적인 값, T_i는 (식 6)과 같다. 여기서 α값은 $0 \leq \alpha \leq 1$사이의 값이다. α값에 따라 기존 방식인 빈도수 기반의 TF_i와 본 논문에서 제안하는 비중치 W_i의 비중이 달라진다.

$$T_i = \alpha\, TF_i + (1-\alpha)\, W_i \quad \cdots\cdots\cdots\cdots \text{ (식 6)}$$

α값이 1이면 기존의 빈도수만을 고려한 주제어 추출이며, α값이 0이면 본 논문에서 제안한 방식으로 키워드 가중치 값만을 고려한 주제어가 추출이 된다.

5. 실험

본 논문에서 제안한 계산 방법으로 비문학 도서 중 김미경 저자의 "아트 스피치"를 분석하였다. 주제어 판단 확인을 위해 본 연구와 무관한 실험자 10명이 선택한 "아트 스피치"의 주제어와 비교한다. 주관적인 주제어가 될 수 있으므로 각자 주제어를 10개씩 선택하여 서로 일치하는 주제어의 개수를 판단 한다.표 2는 실험자 10명이 선택한 "아트 스피치"의 주제어 분포 개수이다. 7개부터 10개까지는 상위 키워드 즉, 가장 주제어가 될 확률이 높은 키워드이다. 4개부터 6개까지는 중위, 그리고 1개를 제외한 2개부터 3개까지는 하위 주제어로 정한다.

표 2. 주제어 분포표
Table 2. Thematic Words Distribution Table

일치개수	키워드	개수
10	청중, 스피치	2
9	콘텐츠	1
8	-	0
7	아트	1
6	설득	1
5	이야기, 스피커	2
4	말, 노하우	1
3	연습, 준비, 기술	3
2	이해, 방법, 공감, 소통, 노력	5
1	대화 외 20	21
총 개수		100

본 논문에서 제안한 방식으로 계산한 결과는 표 3과 같다. 여기서 α값은 0.5로 같은 비율로 빈도수와 키워드 가중치 값을 적용하였다.

표 3. 주제어 추출 결과
Table 3. Result of Thematic Words Extraction

단어	TF	SO	P	C	T
말	0.0356	0.0885	23.6	1.8	25.462
스피치	0.0217	0.0569	20.4	1.8	22.238
청중	0.0207	0.0734	17.2	1.8	19.042
사람	0.0228	0.0499	11.6	1.8	13.437
콘텐츠	0.0060	0.0210	4.4	1.8	6.212
이야기	0.0146	0.0455	4.8	1.2	6.028
스피커	0.0044	0.0168	4.4	0.6	5.009
에피소드	0.0071	0.0249	4.4	0.3	4.714
때	0.0151	0.0251	4.4	0	4.422
강연	0.0106	0.0291	3.6	0	3.619

본 논문에서 제안한 키워드 가중치 방식으로 추출된 주제어 정확도 확인을 위해 주제어 분포표를 기준으로 α값을 1로 지정하여 빈도수만으로 추출한 주제어와 α값을 0.5로 둔, 본 논문에서 제안한 계산식으로 추출한 주제어를 비교하였다.

표 4. 결과표
Table 4. Result Table

중요도	α=1	α=0.5
상	청중, 스피치	청중, 스피치, 콘텐츠
중	이야기, 말	이야기, 스피커, 말
하	-	-

표 4, 결과표를 보면 기존 빈도수만으로 주제어를 추출했을 때의 상위 키워드 일치는 4개 중 2개인 50%의 정확도를 보였고 중위 키워드에서는 5개 중 2개인 40%의 정확도를 보였다. 반면, 본 논문에

서 제안한 키워드 가중치 값을 적용한 결과는 상위 키워드에서는 4개 중 3개가 일치하여 75%의 정확도를 보였고, 중위 키워드에서는 5개 중 3개인 60%의 정확도로 빈도수로 추출한 주제어와 상위와 중위 키워드에서의 정확도 향상을 보였다.

IV. 결 론

본 논문에서는 도서 본문에서의 주제어 추출 시 빈도수로 추출하였을 때보다 정확도를 향상시키기 위해 문장 내 중요 구성 요소에 대한 가중치와 문단 내의 중요 문장에 대한 가중치를 추가하여 주제어를 추출하는 방식을 제안 하였다. 구체적으로는 문장 내의 주어와 목적어에 대한 가중치와 문단에서의 처음과 마지막 문장 안의 키워드에 대한 가중치, 그리고 결론과 요약을 나타내는 문장 안의 키워드에 가중치를 주었다. 본 논문에서 제안한 시스템으로 주제어 추출 실험 결과 정확도가 기존 빈도수로 추출한 주제어보다 중요한 키워드에 대하여 약 20% 이상의 향상을 보였다. 향후 연구로는 다양한 도서 장르에서도 정확도를 향상시킬 수 있는 가중치 값과 가중치 값을 적용할 요소들에 연구를 진행할 예정이다.

Acknowledge

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2014년도 문화기술연구개발지원사업의 연구결과로 수행되었음

참고문헌

[1] Mal-Rey Lee, Hwan-Kuk Bae, "Design of Keyword Extraction System Using TFIDF", Korean Journal of Cognitive Science, Vol. 13, No.1, pp. 1-11, Mar 2002.

[2] Mi-Young Kang, Dae-Wook Kang, "Automatic Document Classification System for Studying Document", Proceedings of Symposium of Korean Institute of communications and Information Sciences pp. 1571-1574, Jul 2006.

[3] Sung-Jick Lee, Han-Joon Kim, Byung-Jeong Lee, Soo-Young Kang, "Concept Network-based User Profile Construction for Personalized Web Search", Korea Information Sceience Society, Vol. 36, No. 1 C. pp 203-208, Jun 2009.

[4] Heui-Kook Ahn, Hi-Young Roh, "Sentence Cohesion & Subject driving Keywords Extraction for Document Classification", Korea Information Science Society, pp. 463-465, Jul 2005.

[5] Seong-geun Hwang, "Rhetorical Approach of Writing Education", Rhetoric Society of Korea, pp. 35-51, Aug 2006.

[6] In-Seok Song, Hyuk-Ro Park, "Text Understanding System for Summarization", The Korean Institute of Information Scientists and Engineers, pp 1-6, Sep 1997.