김일성종합대학학보
정보과학

주체110(2021)년
제67권 제2호

문장중요도에 의한 다중문서요약의 한가지 방법

김 예 화

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《과학기술정보사업을 강화하여야 합니다. 과학기술정보사업을 잘하여야 적은 밑천과 품을 들여 과학기술발전에 절실히 요구되는 귀중한 자료들을 얻을수 있습니다.》(《**김정일**선집》 증보판 제15권 501페지)

최근에 정보검색분야에서 중요한 기술의 하나인 다중문서요약에 대한 연구가 심화되고있다. 특히 중요문장을 추출하여 다중문서요약을 실현하는 방법[1－3]들이 많이 제기되고있다.

빈도수에 토대한 중요문장추출에서 중요단어들은 문서에서 다른 단어들에 비해 여러번 반복된다는것을 가정하고있으며 특징에 토대한 방법은 제목/표제단어, 문장위치, 문장길이, 단어무게, 고유실체와 같은 특징들을 문장의 적합성을 결정하는데 리용하였다. 무리짓기방법에 의한 중요문장선택방법은 다중기사들의 다양성과 풍부성을 표현하는 과제에서 우월하나 무리의 최량개수를 결정하기가 어려운 결함을 가지고있다. 또한 그라프에 기초한 방법은 문서모임에 대한 그라프에서 많은 다른 문장들과 강하게 련결된 문장은 중요하다고 보고있다.

론문에서는 다중문서요약을 위한 중요문장을 추출하기 위하여 정보리득률과 *tf-idf*방법을 결합한 문장중요도계산식과 다중문서요약문을 생성하는 한가지 방법을 제안하였다.

1. 정보리득률에 기초한 문장중요도

론문에서는 정보검색결과에 얻어진 문서들을 리용하였다.

문장중요도를 계산하는 과정은 다음과 같다.

먼저 정보검색결과에 얻어진 문서모임에 대하여 계층적무리짓기(집계법)를 진행한다. 매 무리에 대하여 단어의 확률적분포와 일치하는 정도를 고려하여 매 단어의 중요도를 얻는다.

다음으로 얻어진 무리들로부터 매 단어에 대한 정보리득률을 계산한다.

정보검색결과에 얻어진 문서모임을 C라고 하고 C_i를 C의 i번째 부분무리라고 할 때 C에서 단어 w의 확률분포의 정보리득률(IGR: Information Gain Ratio)은 다음과 같이 정의된다.

$$IGR(w,\ C) = \frac{I(w,\ C) - \sum_i \frac{|C_i|}{|C|} I(w,\ C_i)}{split\_E(C)} \tag{1}$$

여기서 $I(w, C)$는 검색결과모임 C에서 단어 w의 정보량이다.

$$I(w,\ C) = -p(w|C)\log_2 p(w) - (1 - p(w|C))\log_2(1 - p(w|C))$$

$$split\_E(C) = -\sum_i \frac{|C_i|}{|C|}\log_2 \frac{|C_i|}{|C|}$$

요약되여야 할 문서들의 모임이 정보검색결과일 때 검색결과모임과 검색되지 않은 문서모임의 대조는 단어무게에 의하여 차이난다고 볼수 있다. 즉 단어들은 검색되지 않은 문서모임보다 검색된 문서모임과 더 련관이 있으므로 검색결과모임을 리용한다.

식 (1)에 의하여 무리들로부터 매 단어에 대한 하나의 무게값을 얻게 된다. 매 단어에 대한 무게모임을 하나로 통합하여 주어진 문서에서의 단어의 득점을 결정한다. 즉 문서 D로부터 검색결과모임(C)까지의 경로에 나타나는 IGR값들의 평균을 문서 D에서 단어 w의 무게로 하고 그것을 $IGR_avg(w, D)$로 표시한다.(그림)

$$IGR\_avg(w,\ D) = \frac{1}{n+1}\left(\sum_{i \in D\_path} IGR(w,\ C_i) + IGR(w,\ D)\right) \tag{2}$$

여기서 D_path는 계층적무리짓기에서 문서 D로부터 검색결과모임 C까지의 경로에 놓인 무리모임이며 그 개수를 n으로 표시한다.

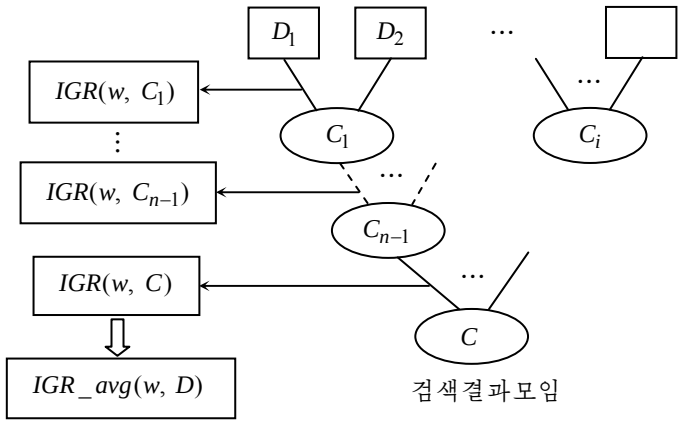

그림. 정보리득률에 기초한 단어무게 $IGR\_avg(w, D)$

다음으로 정보리득률에 기초한 문장의 중요도를 계산한다.

정보리득률에 기초한 문장 S_i의 중요도 $\mathrm{Im}\, p_{IGR}(S_i)$는 문장에서 체언들의 평균무게로 정의한다.

$$\mathrm{Im}\, p_{IGR}(S_i) = \frac{\sum_{w \in Nominal(S_i)} IGR\_avg(w,\ D)}{|\ Nominal(S_i)\ |} \tag{3}$$

우리는 절대적인 득점값보다도 득점값들의 순위를 중요시하므로 다음의 식에서 제시한 T-값을 리용하여 중요도값을 문서에서 정규화하였다.

$$T(x,\ D) = \frac{x\_avg(D)}{standard\_deciation(D)} \tag{4}$$

여기서 x는 정규화를 위한 득점값이고 D는 x를 포함하는 득점들의 평균값이며 $standard_deciation(D)$는 D의 표준편차이다.

마지막으로 문장 S_i의 중요도 $\mathrm{Im}\, p_n(S_i)$는 *tf-idf*에 의한 문장중요도와 정보리득률에 의한 문장중요도의 선형결합으로 정의한다.

$$\mathrm{Im}\, p(S_i) = \alpha \times \mathrm{Im}\, p_{tf-idf}(S_i) + (1-\alpha)\, \mathrm{Im}\, p_{IGR}(S_i) \tag{5}$$

여기서 α는 혼합결수이다.

$$\mathrm{Im}\, p_{tf-idf}(S_i) = \frac{\sum_{w \in Nominal(S_i)} tf(w,\ D) \times idf(w)}{|\ Nominal(S_i)\ |} \tag{6}$$

2. 요약문의 생성

요약문생성에서는 중요문장의 추출과 추출된 중요문장들의 순위화를 진행한다.

$\mathrm{Im}\, p(S_i)$를 요약문장을 생성하기 위한 문장중요도로 보고 $\mathrm{Im}\, p(S_i)$가 어떤 턱값 τ보다 크면 중요문장으로 출력한다. 일반적으로 턱값 τ는 중요문장의 수가 전체 문장수의 17%정도가 되도록 설정한다.

중요문장들의 순위화는 무리수준의 순위화를 진행한 다음 무리내에서의 순위화를 진행한다.

무리수준의 순위화를 위하여 먼저 추출된 중요문장들을 *k*-평균무리짓기의 2진분할법에 의하여 무리짓기를 진행한다. 그다음 무리의 날자를 정하고 날자순위에 의하여 무리들을 순위화한다. 매 무리의 날자는 그 무리에 속하는 문서들중에서 제일먼저 출판된 문서의 날자로 정한다.

마찬가지방법으로 무리내에서의 순위화도 중요문장이 속한 문서의 출판순위와 중요문장들이 같은 문서에 출현한 경우에는 그 문서내에서의 출현순위를 고려하여 순위화를 진행한다.

3. 실험 및 결과분석

중요문장추출을 위하여 사용한 평가척도로는 적중률, 완전률, *F*-척도이다.

적중률, 완전률, *F*-척도는 다음과 같이 계산된다.

적중률: $P = \dfrac{|\ S_s \cap S_h\ |}{|\ S_s\ |}$

완전률: $R = \dfrac{|\ S_s \cap S_h\ |}{|\ S_h\ |}$

F-척도: $F = \dfrac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R}$

여기서 S_h는 정답요약문장모임, S_s는 체계가 출력한 추출요약문장모임이고 β는 적중률과 완전률의 중요도를 조절하는 상수로서 $\beta = 1$로 설정하였다.

*tf-idf*방법(식 (6))과 제안한 방법(식 (5))을 비교하였다.(표 1)

표 1. 중요문장추출방법의 성능비교

비교방법	*P*	*R*	*F*
*tf-idf*방법	0.418	0.321	0.363 134
제안한 방법	0.531	0.367	0.434 024

표 1에서 보여주는바와 같이 제안한 방법이 *tf-idf*방법에 비하여 우월하다는것을 알수 있다.

중요문장순위화를 위한 척도로는 τ거리척도를 리용한다. 이 척도를 일명 Kendall의

척도라고도 한다.

τ 거리척도는 다음과 같이 계산된다.

$$\tau = 1 - \frac{4m}{N(N-1)}$$

여기서 N은 중요문장의 개수, m은 순위화된 중요문장렬을 참고순위의 문장렬로 변환하기 위해 린접한 문장들끼리 순서를 바꾸는 총 회수(참고순위는 정답순위를 의미)이다.

τ의 값은 -1부터 1까지 변한다. 여기서 $\tau = 1$은 $m = 0$인 경우로서 중요문장들의 순서와 참고문장들의 순서가 일치하는 경우이고 $\tau = -1$은 중요문장들의 순서와 참고문장들의 순서가 완전히 거꿀로 되는 경우로서 최대로 나쁜 경우이다.

그러므로 우연적인 순서는 보통 평균값으로서 $\tau = 0$인 경우이다.

중요문장순위화의 실험결과는 표 2와 같다.

표 2. 중요문장순위화의 실험결과

방 법	τ 거리
선행방법[3]	0.657 3
제안한 방법	0.692 86

표 2에서 보는바와 같이 론문에서 제안한 방법이 선행방법에 비하여 순위화가 개선되였다는것을 알수 있다.

맺 는 말

문서요약을 위한 중요문장을 추출하기 위하여 문서에서 매 단어의 정보리득률계산식과 *tf-idf*법을 결합한 문장의 중요도식과 중요문장들의 순위화방법을 제안하고 실험을 통하여 제안한 방법의 성능을 평가하였다.

참 고 문 헌

[1] **김일성**종합대학학보 정보과학, 65, 2, 3, 주체108(2019).

[2] **김일성**종합대학학보(자연과학), 60, 12, 25, 주체103(2014).

[3] Yogan Iaya Kumar et al.; Journal of Computer Science, 12, 4, 178, 2016.

주체110(2021)년 2월 5일 원고접수

A Method for Multi-Document Summarization Using Sentence Importance

Kim Ye Hwa

This paper proposed calculation method of sentence importance for multi-document summarization. The retrieved document set was hierarchical clustering, the sentence importance was calculated by combining information gain and tf-idf, and the summarization sentence was generated by ranking selected important sentences.

Keywords: information gain ratio, multi-document summarization, sentence importance