김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제10호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 62 No. 10 JUCHE105(2016).

다중문서요약기술에 의한 질문중심문서 요약실현의 한가지 방법

정만흥, 리청한

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《정보산업을 빨리 발전시키고 인민경제의 모든 부문을 정보화하여야 합니다.》

(《**김정일**선집》 증보판 제20권 380페지)

《마우스는 언제 발명되였는가?》 또는 《프락탈이란 무엇인가?》와 같은 사실형 또는 정의형질문에 대한 대답은 사실형 또는 정의형질문응답체계를 리용하여 얻을수 있다.[2] 그러나 《마우스의 개념에 대하여》와 같은 질문에 대한 대답의 실현은 사용자질문중심문서요약을 통해 실현할수 있다.

선행연구[1]에서는 령역온톨로지개념련관구조와 호상정보에 기초한 련관용어추출 및 련관용어－문장행렬의 비부값행렬분해풀이를 리용한 사용자질문중심문서요약방법이 제기되였으며 선행연구[3]에서는 질문분해에 의한 적합성반결합의 본문요약방법이 론의되였다. 이 방법은 초기질문을 여러개의 부분들로 쪼개여 질문을 확장하는 원리에 기초하고있는것으로 하여 질문분해에 대한 정보가 충분하지 못한 경우에 질이 낮은 요약문서가 얻어지는 부족점이 있다. 선행연구[4]에서는 질문과 의미특징들사이에 코시누스류사도를 리용하여 요약문장들을 추출하였는데 이 방법은 초기사용자의 질문이 사용자의 요구를 반영하지 못하는 경우에 질이 낮은 요약문서를 산생할수 있다.

선행연구[5]에서는 비부값행렬분해와 적합성반결합 그리고 비부값행렬분해와 의사적합성반결합에 기초한 사용자질문중심의 문서요약방법들을 제안하였다.

론문에서는 다중문서요약기술에 기초하여 다중문서모임으로부터의 사용자질문중심문서요약을 실현하기 위한 새로운 방법을 론의한다.

1. 단일문서요약에서의 다중문서요약기술의 응용

다중문서요약이란 동일한 대상에 대한 정보적내용을 소개한 여러건의 기사들의 모임으로부터 정보적가치가 있는 문장들을 추출하여 순서화하는 기술이다. 이런 의미에서 다중문서요약을 추출요약이라고도 한다. 다시말하여 다중문서요약은 같은 내용을 서술한 기사내용들에서 정보적내용이 담겨져있는 문장들을 추출하는것을 추출요약이라고 말할수 있다. 그러므로 다중문서에서의 요약의 의미는 여러 문서들로부터 정보적가치가 큰 하나의 요약문서를 추출한다는 의미를 담고있다.

일반적으로 문서요약에는 단일문서요약과 다중문서요약이 있으며 특히 단일문서요약에는 일반요약과 질문중심요약이 있다.

문서요약이라고 할 때에는 문서의 의미를 기본적으로 담고있으면서 크기가 전체 문서크기의 15～20%정도 되는 요약문서를 얻는것을 의미하며 질문중심의 요약은 질문자가

관심을 두고있는 보다 구체적인 정보내용을 담고있는 문장들을 얻는것을 의미한다.

단일문서요약에서 다중문서요약기술을 응용하기 위한 절차는 다음과 같다.

① 단일문서로부터 다중요약문서들을 얻는다.

단일문서 T가 주어지는 경우 현재 존재하는 n개의 요약방법을 리용하여 n개의 요약문서 ST_1, ST_2, …, ST_n 들을 얻는다.

같은 내용을 서술한 서로다른 n개의 단일문서들이 주어지는 경우에는 n개의 문서들에 동일한 혹은 서로다른 요약방법을 리용하여 역시 n개의 요약문서 ST_1, ST_2, …, ST_n 들을 얻는다.

만일 같은 내용을 서술한 문서들은 아니지만 문서들에 같은 내용을 포함하고있는 서로다른 n개의 단일문서들이 주어지는 경우에는 이 n개의 단일문서들에 질문중심의 요약기술을 적용하여 우에서와 같이 n개의 요약문서 ST_1, ST_2, …, ST_n 들을 얻는다.

이때 얻어지는 요약문서들은 동일한 내용을 담고있는 짧은 문서들이다.

② 얻어진 n개의 요약문서들을 다중문서로 하는 문서모임에 다중문서요약기술을 적용하여 하나의 요약문서를 얻는다.

2. 질문중심문서요약방법

여기에서는 같은 내용을 서술한 본문내용을 포함하고있는 여러개의 서로 다른 문서들에서 질문중심의 요약을 실현하여 문서요약자에게 필요한 정보내용을 제공하는 문서요약방법을 고찰한다.

실례로 《마우스에 대하여》라는 질문이 주어질 때 콤퓨터일반에 대한 지식을 서술한 여러 도서들로부터 마우스에 대한 지식의 내용을 담은 본문내용을 추출하려는 경우에 다중문서요약기술을 적용하면 좋은 결과를 얻을수 있다. 여기로부터 우리는 다음과 같은 사용자질문중심문서요약방법을 제기한다.

① 입력된 질문문장으로부터 질문용어 q_1, q_2, …, q_t 들을 추출한다. 실례로 《마우스에 대하여》라는 질문이 입력되면 질문용어는 1개로서 용어 《마우스》가 추출된다.

② 질문용어들과 련관이 있는 단어들을 주어진 문서들로부터 추출한 다음 질문용어들과 련관단어들로 만들어지는 용어－문장행렬을 얻는다.

③ 용어－문장행렬에 대한 비부값행렬분해(NMF)법[6]을 리용하여 매 문서들에 대한 요약문서 ST_1, ST_2, …, ST_n 들을 얻는다.

④ 요약문서 ST_1, ST_2 그리고 ST_3들이 각각 다음과 같은 4, 5, 3개의 문장 즉

$$ST_1 = \{s_{11}, s_{12}, s_{13}, s_{14}\}$$
$$ST_2 = \{s_{21}, s_{22}, s_{23}, s_{24}, s_{25}\}$$
$$ST_3 = \{s_{31}, s_{32}, s_{33}\}$$

과 같이 구성되였다고 하면 이때 최종적으로 4, 5, 3의 평균값 4를 크기로 하는 요약문서를 얻는다. 즉 요약문서를 이루는 4개의 문장은 총 12개의 문장중에서 문장득점값이 큰 순서로 선택된다.

만일 문장득점값의 크기순서가 s_{11}, s_{23}, s_{32}, s_{25}이라면 요약문서 S는 다음과 같이 표시할수 있다.

$$S=<s_{11}, s_{23}, s_{32}, s_{25}>$$

3. 질문련관단어와 문장득점평가방법

1) 질문련관단어

질문련관단어를 추출하기 위하여 먼저 문서속에 들어있는 단어들사이의 린접정도와 호상정보량을 리용하여 단어들사이의 련관도를 다음과 같이 계산한다.

$$A(w_1,\ w_2) = I(w_1,\ w_2)\cdot\exp(-\alpha\rho(w_1,\ w_2))$$

여기서 α는 구간 [0, 1]사이의 값이고 $\rho(w_1,\ w_2)$는 동일한 문장속에 들어있는 단어 w_1과 w_2들사이의 린접성정보를 나타내는 량으로서 두 단어사이에 놓이는 단어개수를 num이라고 할 때 다음과 같이 계산된다.

$$\rho(w_1,\ w_2) = \min\{\ \mathrm{num}_k\ |k\text{: 문장번호}\}+1$$

그리고 $I(w_1,\ w_2)$는 단어 w_1과 w_2에 대한 호상정보량으로서 다음과 같이 계산된다.

$$I(w_1,\ w_2) = \log\frac{p(w_1,\ w_2)}{p(w_1)p(w_2)}$$

여기서 $p(w_1,\ w_2)$는 동일한 문장들에서 단어 w_1과 w_2의 동시출현확률이며 $p(w_1)$과 $p(w_2)$는 각각 단어 w_1과 w_2의 출현확률이다.

질문용어 q가 들어있는 문장속의 단어 w들은 $A(q,\ w)$의 값이 작아지는 순서로 순위화하고 여기서 t개의 단어들을 앞순위의 순서로 선택한다. 이때 선택된 t개의 단어들을 질문용어 q의 1차련관단어라고 부른다.

마찬가지로 1차련관단어들의 모임 $\{w_1,\ w_2, \cdots, w_t\}$에 속하는 매개 단어들에 대한 1차련관단어들을 모두 질문용어 q의 2차련관단어라고 부른다.

질문용어 q와 그것의 1차련관단어 및 2차련관단어들로 이루어진 단어들의 모임을 질문련관단어모임으로 정의하고

$$\{w_1,\ w_2, \cdots, w_n\}$$

으로 표시한다.

2) 질문련관단어에 의한 문장의 득점값계산

문장의 득점값계산방법은 다음과 같다.

① 질문련관단어모임에 속하는 련관단어 w_i와 w_j들의 쌍 $\langle w_i,\ w_j\rangle$의 무게를 다음과 같이 결정한다.

$$W = \begin{cases} P_{ij} : w_i\text{와 } w_j\text{가 둘 다 } q\text{의 1차련관단어인 경우} \\ 0.7\times P_{ij} : w_i\text{와 } w_j\text{중 하나만이 } q\text{의 1차련관단어인 경우} \\ 0.3\times P_{ij} : w_i\text{와 } w_j\text{가 둘 다 } q\text{의 2차련관단어인 경우} \end{cases}$$

여기서 P_{ij}는 단어 w_i와 w_j들을 동시에 포함하는 문장의 개수이다.

② 무게가 큰 련관단어의 쌍들을 많이 포함하는 문장일수록 중요한 문장이라고 보고 매 문장에 대하여 그 문장에 포함되여있는 가능한 련관단어쌍 $\langle w_i,\ w_j\rangle$들의 무게합을 그 문장의 득점값으로 설정한다. 즉 문장 S의 득점값 $E(S)$는 다음과 같이 계산한다.

$$E(S) = \sum_{w_i,\ w_j\in S} W$$

이때 련관단어들의 쌍을 하나도 포함하지 않는 문장의 득점값은 0이다.

문장득점값을 계산한 후 득점값이 류사한 서로 다른 요약문서들에 들어있는 두 문장에 대해서는 최종적인 요약문서에서 의미가 같은 문장들의 중복관계를 피하기 위하여 두 문장의 코시누스류사도평가를 진행하여 요약문서에서의 제거처리를 진행한다.

4. 실험결과분석

론문에서 사용자질문중심문서요약체계의 성능을 평가하기 위하여 교과서《콤퓨터기초》를 비롯한 콤퓨터관련도서 5종을 준비하여 실험을 진행하였다.

우리는 1개의 문서《콤퓨터기초》에 질문응답요약체계[2]를 적용한 요약문서결과와 5종의 콤퓨터관련도서 모두에 론문의 방법을 적용하여 얻은 요약문서결과를 놓고 5명의 도서집필자들이 점수를 매기는 방법으로 제안한 방법의 효과성을 검증하였다. 효과성검증결과는 표와 같다.

표. 실험결과

방법	1	2	3	4	5	평균점수
선행한 방법[2]	4.5	4.0	4.1	4.1	4.0	4.14
제안한 방법	4.6	4.4	4.5	4.4	4.3	4.44

표에서 수자 1, 2, 3, 4, 5는 5명의 평가자들을 의미하며 그밑의 수자들은 평가자들의 평가점수를 의미한다. 여기서 평가자 1은 교과서《콤퓨터기초》의 집필자이다.

표에서 보여주는바와 같이 5명의 평가자모두는 론문에서 제기한 방법이 선행한 방법[2]에서 제기한 방법에 비해 보다 효과적인 방법이라고 평가하였다.

맺 는 말

질문중심의 요약체계에서 다중문서요약기술의 원리를 응용하여 요약의 정확도를 개선하기 위한 한가지 방법을 제기하고 실현하였다. 이 방법은 여러 문서를 리용하여 사용자질문에 대한 요약문서를 얻는것으로 하여 요약문서의 질을 향상시킬수 있다.

참 고 문 헌

[1] 정만흥 등; 정보과학, 1, 28, 주체104(2015).

[2] Journal of **KIM IL SUNG** University(Natural Science), 1, 4, 51, Juche101(2012).

[3] D. H. Bea et al.; Proc. of International Workshop on Information Retrieval with Asia Languages, 201, 2000.

[4] C. M. Ahn et al.; Proc. of Knowledge-Based Intelligent Information and Engineering Systems, 84, 2006.

[5] S. Park; 2009 International Conference on Computer Engineering and Applications IPCSIT, 2, 101, 2011.

주체105(2016)년 6월 5일 원고접수

A Method of Implication for Query-Focused Summarization by Multi-Document Summarization

Jong Man Hung, Ri Chong Han

We propose an approach for the query-focused multi-document summarization that extracts needful information for question from different documents including text of similar content.

Key words: document summarization, question answering