김일성종합대학학보
(자연과학)
주체106(2017)년 제63권 제4호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 63 No. 4 JUCHE106(2017).

# 련관단어를 리용한 질문중심다중문서요약의 한가지 방법

**정만흥, 리청한**

사실형질문《전화기는 언제 발명되였는가?》 또는 정의형질문《프락탈이란 무엇인가?》와 같은 질문에 대답은 사실형 또는 정의형질문응답체계에 의해 얻을수 있다.[2, 3] 그러나《물시금치는 어떤 남새인가?》와 같이 어떤 대상에 대한 일정한 지식을 요구하는 서술형질문에 대한 대답은 얻을수 없다.

서술형질문에 대한 대답은 문서요약기술을 리용하여 얻을수 있으며 이와 같은 문서요약기술을 질문중심문서요약이라고 부른다. 일반적으로 질문중심문서요약체계는 4개의 부분 즉 적합문서추출부분, 특징추출부분, 문장선택부분 그리고 요약생성부분으로 구성된다.[4]

질문중심문서요약은 선행연구[1, 5]에서 제기되였다. 선행연구[1]에서는 온톨로지에 의한 사실형질문설계방법과 일반문서요약방법을 결합한 서술형질문응답체계를 정의하고 실현하였다. 선행연구[5]에서는 질문눈덩이개념에 기초한 질문응답의 다중문서요약방법을 제기하였다.

론문에서는 득점값에 의한 련관단어추출의 한가지 방법을 제기하고 련관단어－문장행렬에 대한 비부값행렬분해(NMF)법[3]에 의하여 적합문장들을 평가 및 선택하여 질문중심문서요약을 실현하는 한가지 방법을 제기하였다.

## 1. 련관단어추출과 다중문서요약

### 1) 련관단어모임

여기서는 질문문장에 들어있는 질문단어와 요약하려는 다중문서모임속에 들어있는 단어들사이의 련관도를 평가하는 방법에 대하여 론의한다.

우선 $R$를 질문단어 $q$와 문장준위에서 동시출현하는 단어들의 모임이라고 할 때 $R$에 속하는 단어 $w$에 대하여 먼저 질문단어 $q$와 독립인 단어 $w$에 고유한 단어중요도값을 다음과 같이 계산하는데 이 단어중요도값을 간단히 토대득점값이라고 부르고 $s_b(w)$로 표시한다.[5]

$$s_b(w) = \log\left(\frac{N}{n(w)}\right)$$

여기서 $N$은 다중문서모임에 들어있는 문서의 총개수이며 $n(w)$는 단어 $w$의 문서빈도수이다.

다음으로 질문단어 $q$와의 련관관계를 나타내는 단어 $w$에 대한 련관득점값 $s_r(w)$를 계산한다. 득점값은 단어 $w$의 토대득점값과 단어 $w$와 질문단어 $q$와의 호상정보량에 기초하여 계산된다.

$$s_r(w) = s_b(w) \cdot s_b(q) \cdot I(q,\ w)\exp(-\alpha\rho(q,\ w))$$

여기서 $\alpha \in [0,\ 1]$이고 $\rho(q, w)$는 $q$와 $w$사이에 존재하는 단어개수이며 $I(q, w)$는 질문단어 $q$와 단어 $w$의 호상정보량이다. 만일 질문단어 $q$와 단어 $w$의 문서빈도수확률이 각각 $p(q)$와 $p(w)$라면 호상정보량은 다음과 같다.

$$I(q, w) = \log \frac{p(q, w)}{p(q)p(w)}$$

여기서 $p(q, w)$는 질문단어 $q$와 단어 $w$의 문서준위동시출현확률 즉 $q$와 $w$가 동시출현하는 문서개수를 다중원천문서모임의 총문서개수 $N$으로 나눈 값이다.

득점값 $s_r(w)$의 정의로부터 알수 있는것처럼 련관관계를 나타내는 득점값은 매개 단어의 토대득점값, 단어들의 호상정보량 그리고 두 단어사이의 근접성정도에 의존하도록 정의하였다. 호상정보량은 일반적으로 단어련관성에 대한 확률모형으로 자주 리용된다.

단어의 련관득점값에 따라 질문단어들과의 련관단어모임을 다음과 같이 얻는다.

① 질문문장으로부터 련관단어모임 $S$를 질문단어 $q_1, q_2, \cdots, q_m$들로 초기화한다. 즉 $S=\{q_1, q_2, \cdots, q_m\}$으로 놓는다.

다음 $k=1, 2, \cdots, m$에 대하여 ②－⑤를 수행한다.

② 질문단어 $q_k$의 동시출현단어 $w \in R$들의 득점값 $s_r(w)$를 계산한다.

③ 득점값 $s_r(w)$들에 대한 평균값을 계산한다. 즉

$$AverageScore = \frac{1}{|R|}\sum_{w \in R} s_r(w)$$

여기서 $|R|$는 동시출현단어 $w$들의 총개수이다.

④ $s_r(w) > AverageScore$이면 단어 $w$를 질문단어 $q_k$의 1차련관단어로 정의하고 련관단어모임 $S$에 포함시킨다.

⑤ 1차련관단어들을 질문단어로 간주하고 걸음 ②－⑤를 반복한다. 이때 얻어지는 련관단어를 질문단어 $q_k$의 2차련관단어로 정의한다.

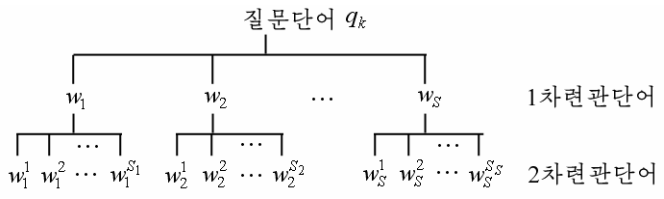


그림 1. 1차 및 2차련관단어사이관계

결국 련관단어모임 $S$는 다음과 같이 정의된다.

$S=\{$질문단어모임$\} \cup \{$1차 및 2차 련관단어$\}$

질문단어 $q_k$의 1차 및 2차련관단어관계는 그림 1과 같다.

2) 단일문서요약

련관단어모임 $S$에 토대하여 크기 $m \times n$ 인 단어－문장행렬 $A$를 문서별로 만든다. 여기서 $m$은 련관단어모임 $S$에 속하는 해당 문서속의 단어개수이며 $n$은 문서에서의 문장수이다. 행렬 $A$의 $i$째 행 $j$째 렬에 놓이는 원소의 값은 련관단어모임 $S$에 속하는 $i$째 단어가 원천문서의 $j$째 문장에 출현하는 빈도수이다.

NMF알고리듬[3]을 리용하여 단어－문장행렬 $A$를 $m \times r$ 비부값의미특징행렬 $W$와 $r \times n$ 비부값의미변수행렬 $H$의 적으로 분해한다. 즉 $A=WH$. 여기서 $r$는 보통 $m$이나 $n$보다 작게 선택되여 $W$와 $H$의 총크기는 행렬 $A$의 크기보다 더 작아지게 한다.

행렬분해알고리듬 NMF는 Frobenius노름을 최소화하는 원리에 기초하여 얻는다.

Frobenius노름은 다음과 같다.

$$\Theta_E(W, H) \equiv \|A - WH\|_F^2 \equiv \sum_{j=1}^{m}\sum_{i=1}^{n}\left(X_{ji} - \sum_{l=1}^{r} W_{jl}H_{lj}\right)^2$$

분명히 이 값이 0이기 위해서는 $A=WH$ 일것이 필요하고 충분하다. $W$와 $H$는 $\Theta_E(W, H)$

가 미리 정의된 턱값이하로 수렴하거나 반복수를 초과할 때까지 반복적으로 갱신된다. 갱신규칙들은 다음과 같다.

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu}\frac{(W^TA)_{\alpha\mu}}{(W^TWH)_{\alpha\mu}}\ ,\ W_{i\alpha} \leftarrow W_{i\alpha}\frac{(AH^T)_{i\alpha}}{(WHH^T)_{i\alpha}}$$

행렬분해후 문장의 일반적적합성을 계산한다.[5]

$$j\text{째 문장의 일반적적합성} = \sum_{i=1}^{r}(H_{ij}\cdot \text{weight}(H_{i*}))$$

여기서 $\text{weight}(H_{i*}) = \sum_{q=1}^{n} H_{iq} \Big/ \sum_{p=1}^{r}\sum_{q-1}^{n} H_{pq}$ .

다음 주어진 턱값보다 큰 일반적합성값을 가지는 $k_p$개의 문장을 선택한다. 여기서 첨수 $p$는 다중문서모임에서 문서번호이다.

### 3) 다중문서요약

론문에서는 다중문서모임에 속하는 개별적인 문서들에 대한 요약문서들을 얻은 다음 거기에 들어있는 문장들의 문장순위화를 진행하여 다중문서요약을 실현하였다.

두 문서 $d_1$과 $d_2$에 대한 요약문장과 그것의 전후관계문장실례는 그림 2와 같다.

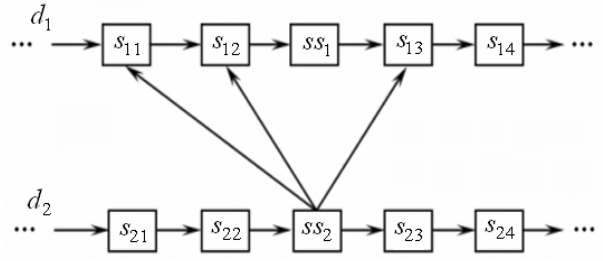


그림 2. 요약문장과 전후관계문장

그림 2에서 $ss_1$과 $ss_2$는 2개의 요약문장이며 $s_{1i}$와 $s_{2i}$들은 각각 요약문장 $ss_1$과 $ss_2$의 전후관계문장들이다. 요약문장과 그것의 전후관계문장들은 동일한 문서에 속한다.

만일 $ss_2$가 $s_{11}$과 류사하다면 $ss_2$와 $s_{11}$은 린접이라고 간주할수 있는 확률이 크며 문장 $s_{11}$이 $ss_1$의 앞에 있으므로 $ss_2$가 $ss_1$ 의 앞에 있을 확률은 크다고 할수 있다. 분명히 문장 $ss_2$와 문장 $s_{11}$사이의 류사성이 높을수록 $ss_2$가 $ss_1$의 앞에 있을 확률이 높다. 만일 $ss_2$가 $s_{11}$과 류사하지 않고 $s_{12}$와 류사하다고 해도 $ss_2$가 $ss_1$의 앞에 있을것이라고 가정할수 있다. 만일 $ss_2$가 $s_{13}$과 류사하다면 $ss_1$이 $ss_2$보다 앞에 놓일수 있다고 볼수 있다.

두 문장의 린접믿음도를 계산한 다음 요약문장들의 순서를 나타내기 위해 무게를 고려한 방향그라프를 리용한다. 그라프의 정점은 문장을 의미한다. 만일 두 문장이 린접하고있다면 두 문장을 련결하는 릉이 존재한다. 릉의 무게는 두 문장의 린접믿음도이다.

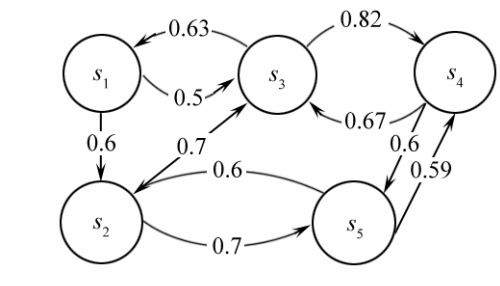


그림 3. 요약문장들과 순서관계그라프

요약문장들사이의 순서관계를 보여주는 실례그라프는 그림 3과 같다.

그림 3에서 보는바와 같이 정점 $s_1$로부터 정점 $s_3$에로의 릉이 존재하며 무게는 0.5이다. 이것은 문장 $s_1$과 문장 $s_3$이 린접해있으며 린접믿음도가 0.5라는것을 의미한다. 그림 3의 방향그라프로부터 최대무게를 가지며 모든 정점

을 포함하는 경로를 구할수 있으며 이 경로의 순서를 요약문장들의 가장 론리적인 순서로 간주할수 있다. 그러나 그라프의 최량경로를 찾는 문제는 전형적인 NP－난문제이다.

그러므로 론문에서는 근사적인 최량경로를 찾는 알고리듬을 제기한다.

이를 위해 문서모임 $D=\{d_1, d_2, \cdots, d_n\}$과 문서 $d_j=\{s_{j1}, s_{j2}, \cdots, s_{jm}\}$을 가정한다. 여기서 $s_{ji}$는 문서 $d_j$의 $i$째 문장이다. 그리고 요약문장모임 $SS=\{ss_1, ss_2, \cdots, ss_k\}$와 빈 그라프 $G$=NULL을 가정하면 $j$번째 문서 $d_j$의 $i$번째 문장 $ss_i$에 대하여 그것의 전후관계문장모임은 $\{d_j - ss_i\}$로 된다.

알고리듬은 다음과 같다.

걸음 1 요약문장모임 $SS$의 매개 요약문장 $ss_i$들을 그라프 $G$의 정점으로 추가한다.

걸음 2 문서모임 $D$의 $j$째 문서 $d_j$의 요약문장 $ss_i$에 대하여 류사도 $\mathrm{sim}(ss_k, s_{j,p})$들을 계산한다. 여기서 $ss_k = \{SS - ss_i\}$, $s_{j,p} \in \{d_j - ss_i\}$ 이다. 만일 $\mathrm{sim}(ss_k, s_{j,p}) > \theta$ (턱값)인 $s_{j,p}$가 존재하면 $ss_i$와 $ss_k$는 방향있는 이웃이다. 이때 $s_{j,p}$가 $ss_i$의 앞에 있으면 $ss_k$가 $ss_i$의 앞에 있을 믿음성무게는 다음과 같이 계산한다.

$$w(ss_i, ss_k) = \alpha \mathrm{sim}(ss_k, ss_{j,p}) + (1-\alpha) r(s_{j,p}) / r(ss_i)$$

여기서 $r(ss_i)$는 문장 $ss_i$가 속하는 문서에서 $ss_i$의 순위(실례로 문서의 두번째 문장의 순위는 2), $\alpha$는 믿음성에 대한 문장류사성의 기여률이다.

만일 $ss_k$로부터 $ss_i$에로의 릉이 존재하지 않는다면 그라프 $G$에 릉을 추가한다. 릉이 존재하는 경우 $w(ss_i, ss_k)$가 릉에 이미 할당된 무게보다 크다면 릉에 할당된 값을 갱신한다. $ss_i$가 $s_{j,i}$의 앞에 있으면 $ss_k$가 $ss_i$의 앞에 있을 믿음성무게는 다음과 같이 계산된다.

$$w(ss_i, ss_k) = \alpha \mathrm{sim}(ss_k, ss_{j,p}) + (1-\alpha)(L(d_j) - r(s_{j,p}))/(L(d_j) - r(ss_i))$$

여기서 $L(d_j)$는 문서 $d_j$에 포함되여있는 문장의 총개수이다.

만일 $ss_i$로부터 $ss_k$에로의 릉이 존재하지 않는다면 그라프 $G$에 릉을 추가한다. 릉이 존재하는 경우 $w(ss_i, ss_k)$가 릉에 이미 할당된 무게보다 크다면 릉에 할당된 값을 갱신한다. 이때 문장 $s_1$과 $s_2$사이의 류사성은 TF-IDF 혹은 WordNet에 기초하여 계산할수 있다.

론문에서는 두 문장사이의 류사성을 TF-IDF의 코시누스류사성을 리용하여 계산하였다.

걸음 3 문장순위화의 첫문장을 결정한다.

먼저 요약문장모임 $SS=\{ss_1, ss_2, \cdots, ss_k\}$에 속하는 문장들사이의 TF-IDF의 코시누스류사성을 리용하여 매개 요약문서 $ST_1, ST_2, \cdots, ST_d$들의 첫번째 문장 $ss_{t1}$들의 류사성특징을 다음과 같이 계산한다.

$$\mathrm{feature}(ss_{t1}) = \frac{1}{K}\sum_{j=1}^{K} \mathrm{sim}(ss_{t1}, ss_j),\ t = 1, 2, \cdots, d$$

여기서 $d$는 요약문서의 개수이고 $ss_{t1}$은 $t$번째 요약문서에 들어있는 첫번째 문장이다.

다음 $\mathrm{feature}(ss_{t1})$값이 최소로 되는 문장

$$ss_{p1} = \arg\min\{\mathrm{feature}(ss_{t1}),\ t = 1, 2, 3\}$$

을 문장순위화에서 첫 문장으로 한다.

걸음 4 순위화목록 $P$를 문장 $ss_{p1}$로 초기화한 다음 $P$의 문장들을 차례로 얻는다.

요약문장들의 모임 $SS$의 문장들중에서 이미 순위화된 문장렬 $ss_1$, $ss_2$, …, $ss_i$가 주어졌다고 하자. 이때 순위화의 다음 문장 $ss_{i+1}$을 다음과 같이 구할수 있다.

$$ss_{i+1} = \arg\max\{w(ss_i, ss_j), ss_j \in SS - P\}$$

걸음 5 만일 $P$가 모든 요약문장들을 포함한다면 알고리듬을 끝내고 그렇지 않으면 걸음 4로 이행한다.

## 2. 실 험 결 과

론문에서 제기한 방법의 성능을 평가하기 위하여 《광명》홈페지에 구축되여있는 《생물대사전》문서를 리용하였다. 실험에 리용된 질문의 개수는 10개로 하였으며 이 질문들에 대한 정확한 요약문서들은 수동으로 작성하였다.

정확한 요약문서를 구성하고있는 정답문장들과 체계가 출력시킨 요약문서를 구성하고있는 대답문장들을 가지고 적중률, 완전률, $F$－값을 계산하여 체계의 성능을 평가하였다. 여기서 적중률, 완전률, $F$－값들은 10개의 질문들에 대한 평균값들이다.

적중률, 완전률, $F$－값은 다음과 같이 계산된다.

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

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

$$F\text{－값: } F = \frac{(\beta^2 + 1)\cdot PR}{\beta^2 P + R}$$

여기서 $S_h$는 정답문장모임, $S_s$는 체계가 출력시킨 대답문장모임, $\beta$는 적중률과 완전률의 중요도를 조절하는 상수로서 $\beta$=2로 설정하였다.

실험은 호상정보량과 단어의 가까운 정도를 고려하여 련관단어를 정의한 선행연구결과[1]와 비교하는 방법으로 진행하였다.(표)

표. 실험결과

| 방법 | $R$ | $P$ | $F$ |
|---|---|---|---|
| 선행한 방법[1] | 0.401 4 | 0.536 4 | 0.386 3 |
| 제안한 방법 | 0.690 9 | 0.540 0 | 0.654 3 |

표에서 보는바와 같이 제안한 방법이 선행한 방법과 거의 같은 적중률을 보장하면서도 선행한 방법에 비해 완전률을 1.72배, $F$－값을 1.69배로 높였다.

## 맺 는 말

토대득점값들과 호상정보량을 고려한 새로운 련관득점값에 기초하여 련관단어모임을 정의하였으며 련관단어－문장행렬의 비부값행렬분해알고리듬과 문장순위화과정을 통해 질문중심다중문서요약을 실현하는 질문중심다중문서요약체계를 설계하고 그 효과성을 검증하였다.

## 참 고 문 헌

[1] 정만흥 등; 정보과학, 1, 28, 주체104(2015).

[2] MAO Xianling et al.; Journal of Frontiers of Computer Science and Technology, 6, 3, 193, 2012.

[3] S. Park et al.; Proc. of Knowledge-based Intelligent Information and Engineering Systems, 84～89, 2006.

[4] Jimmy Lin et al.; The 2010 Annual Conference of the North American Chapter of the ACL, 305, 2010.

[5] Hajime Morita et al.; Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Shortpapers, 223, 2011.

주체105(2016)년 12월 5일 원고접수

## A Method of Query-based Multiple Documents Summarization using Associative Words

*Jong Man Hung, Ri Chong Han*

We discuss a method for query-based multiple documents summarization using associative words. We conduct the document summarization using the method of selecting sentences with a high degree of appropriacy by non-negative matrix factorization of the word-sentence matrix based on query word and associative word set.

Key words: query-based document summarization, associative word