김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제9호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 62 No. 9 JUCHE105(2016).

중요단어득점값을 리용한 문서요약의 한가지 방법

정만흥, 김경미

선행연구[1]에서는 질문단어와 그 련관단어를 리용한 질문중심문서요약방법을 제기하였으며 선행연구[2]에서는 토대득점값 및 련관득점값의 개념에 기초하여 질문중심다중문서요약방법을 제기하였다.

론문에서는 중요단어들과 련관정도가 큰 단어일수록 보다 가치있는 중요단어로 될수 있다는 가정에 기초하여 단어의 중요도계산을 위한 한가지 방법을 제기하고 그에 기초하여 단일문서요약을 실현하는 방법을 제기하였다.

1. 단어의 중요도계산

요약의 견지에서 높은 중요도를 가지는 단어들과 보다 밀접한 관계를 가지는 단어일수록 요약문장의 요약단어들로 될 가능성이 크다고 하자. 이때 단어 s_i의 중요도를 $v(s_i)$라고 하면 우의 가정으로부터 다음의 관계식을 이끌어낼수 있다.

$$v(s_i)=\alpha\sum_{i=1}^{n}w_{ij}v(s_j)+\beta q(s_i),\quad \alpha,\ \beta\in[0,\ 1],\ \alpha+\beta=1 \tag{1}$$

여기서 w_{ij}는 단어 s_i와 s_j사이의 류사성을 특징짓는 파라메터이며 $q(s_i)$는 단어 s_i의 특징값으로서 단어의 중요도를 결정하는데서 중요한 의의를 가진다.

따라서 우리는 이 두 값들에 대하여 보기로 한다.

우선 파라메터 w_{ij}를 $w_{ij}=\begin{cases}\mathrm{sim}(s_i,\ s_j), & i\neq l\\ 0, & i=l\end{cases}$ 과 같이 계산한다. 여기서 $\mathrm{sim}(s_i,\ s_j)$는 단어 s_i와 s_j사이의 류사도이다.

웃식에서 보는바와 같이 w_{ij}의 계산은 단어들사이의 류사도 $\mathrm{sim}(s_i,\ s_j)$의 계산에 귀착되는데 이를 위해서는 토대득점값을 결정하여야 한다.

R를 요약하려는 문서에 들어있는 명사단어들의 모임이라고 하면 R에 속하는 단어 s에 대하여 단어 s 그자체에 고유한 토대득점값 $v_b(s)$는 다음과 같이 계산된다.

$$v_b(s)=\log(N/n(s))$$

여기서 N은 모임 R에 있는 단어의 전체개수이며 $n(s)$는 단어 s의 원천문서에서의 출현빈도수이다. 분명히 N은 $n(s)$보다 크며 log값은 정인 값을 가지게 된다.

여기로부터 단어들사이의 류사도 $\mathrm{sim}(s_i,\ s_j)$는 다음과 같이 계산된다.

$$\mathrm{sim}(s_1,\ s_2)=v_b(s_1)v_b(s_2)\left(\frac{I(s_1,\ s_2)}{\exp(\alpha\rho(s_1,\ s_2))}\right)$$

여기서 $\alpha \in [0, 1]$이고 $\rho(s_1, s_2)$는 s_1과 s_2사이에 존재하는 단어개수이며 $I(s_1, s_2)$는 단어 s_1과 s_2의 호상정보량이다.

다음 특징값 $q(s_i)$에 대하여 보기로 하자.

특징값 $q(s_i)$는 단어의 중요도가 요약의 의미에서 가치를 가지도록 결정해야 하는데 그 방법에는 다음과 같은 두가지가 있다.

방법 1 $q(s_i) = v_b(s_i)$

방법 2 $q(s_i) = v_b(s_i) + \mathrm{local}v_b(s_i)$

여기서 $\mathrm{local}v_b(s_i)$는 이미 알려진 요약방법으로서 얻어진 요약문서에서의 단어 s_i의 토대득점값이다. 이 값은 주어진 문서요약방법에 의해 구해진 요약을 보다 갱신하려고 할 때 얻을수 있다.

2. 단어의 중요도 $v_b(s_i)$의 계산과 문서요약

① 단어의 중요도 $v_b(s_i)$의 계산

단어의 중요도 $v_b(s_i)$를 계산하자면 식 (1)로 주어지는 반복도식을 풀어야 한다. 반복도식 (1)을 풀기 위하여 우선 행렬 $\boldsymbol{W}$ 그리고 두 벡토르 $\boldsymbol{v}$와 $\boldsymbol{q}$를 정의한다.

$$\boldsymbol{W} = \alpha \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nn} \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} v(s_1) \\ v(s_2) \\ \vdots \\ v(s_n) \end{bmatrix}, \quad \boldsymbol{q} = \beta \begin{bmatrix} q(s_1) \\ q(s_2) \\ \vdots \\ q(s_n) \end{bmatrix}$$

분명히 $\boldsymbol{W}_{ij}$를 i째행, j째렬의 원소로 하는 행렬 $\boldsymbol{W}$를 정의할 때 이 행렬은 대칭행렬이다.

이때 반복도식 (1)을 행렬－벡토르형식으로 쓰면 다음과 같다.

$$\boldsymbol{v} = \boldsymbol{W}\boldsymbol{v} + \boldsymbol{q} \tag{2}$$

$$(\boldsymbol{I} - \boldsymbol{W})\boldsymbol{v} = \boldsymbol{q} \tag{3}$$

방정식 (3)의 풀이가 유일존재하도록 행렬 $\boldsymbol{W}$를 다음과 같이 변경시킨다.

첫째로, 행렬 $\boldsymbol{W}$의 렬 또는 행에 대한 정규화를 진행한다.

둘째로, 행렬 $\boldsymbol{W}$의 매 원소들의 크기를 작게 하기 위해 감쇠인자 $\theta(0 < \theta < 1)$를 $\boldsymbol{W}$의 매 원소에 곱한다.

이리하여 최종적으로 다음의 련립방정식을 얻는다.

$$(\boldsymbol{I} - \theta\boldsymbol{W})\boldsymbol{v} = \boldsymbol{q} \tag{4}$$

우리는 결수행렬 $(\boldsymbol{I} - \theta\boldsymbol{W})$가 강한 대각선우세행렬이 되도록 감쇠인자 θ를 선택하였다. 이때 련립방정식 (4)의 풀이 $\boldsymbol{v}$는 유일존재하며 가우스－자이델법에 의해 단어의 중요도를 계산할수 있다.

② 문서요약

먼저 단어의 중요도에 기초하여 문장의 중요도를 문장속에 들어있는 명사단어들의 중요도합으로 정의한다. 즉 $M(s_k) = \sum_{s_i \in C_k} v(s_i)$.

여기로부터 문서요약알고리듬은 다음과 같다.

주어진 문서에 들어있는 문장들을 $s_1, s_2, \cdots, s_n$이라고 하자.

① 문장 $s_k(k=1, \cdots, n)$에 들어있는 단어 s_i를 리용하여 문장 s_k의 중요도 $M(s_k)$를 계산한다.

② 문장의 중요도가 큰 순서로 문장들을 순서화한다.

③ 정해진 개수의 문장을 순서화의 순위로 선택한다.

④ 원천문서에서의 문장들의 순서관계를 유지하도록 선택된 문장들을 재배치하여 요약문서를 얻는다.

3. 실험결과 및 분석

론문에서 제기한 문서요약방법의 성능을 평가하기 위하여 선행한 방법[3]과 비교실험을 진행하였다. 요약문서자료는 58개의 문장으로 구성된 콤퓨터의 일반상식을 서술한 문서이다. 그리고 요약문서의 문장의 개수는 12로 하였다.

비교실험은 두가지 방법으로 하였다.

방법 1 특징벡토르를 $q(s_i)=v_b(s_i)$에 의해 구한 경우

방법 2 특징벡토르를 $q(s_i)=v_b(s_i)+\mathrm{local}v_b(s_i)$에 의해 구한 경우

표. 실험결과

방법	P	R	F(β=3)
선행한 방법	0.480	0.612	0.510
제안한 방법 1	0.534	0.623	0.522
제안한 방법 2	0.573	0.764	0.625

정답요약문장과 체계가 출력한 요약문장들을 가지고 적중률, 완전률, F값을 리용하여 체계의 성능을 평가한 결과는 표와 같다.

표에서 보는바와 같이 론문에서 제안한 방법 1은 선행한 방법과 F값이 거의 같으나 방법 2는 F값이 1.23배로 높아졌다.

참 고 문 헌

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

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

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

주체105(2016)년 5월 5일 원고접수

A Method for Documents Summarization using Gain-Values of Important Words

Jong Man Hung, Kim Kyong Mi

We discuss a method for documents summarization using gain-values of important words. Our method is the basis for assumption that a word is important in a document if it is heavily linked with many important words in the same document.

Key words: important word, document summarization