단어들사이의 련관관계에 의한 중요단어추출의 한가지 방법

설룡의, 정만흥

본문정보검색체계, 질문응답체계, 문서요약체계 등 문서정보검색체계에서 벡토르모형을 리용하는 경우 문장들은 모두 단어－벡토르로 모형화된다. 단어－벡토르모형에서 단어의 중요도값을 고려하는것은 문서정보검색체계의 정확도를 높이는 중요한 문제의 하나로 된다.

일반적으로 단어의 중요도는 단어의 출현빈도수와 거꿀출현빈도수에 의해 계산된다. 선행연구들에서는 두 단어의 동시출현성을 고려한 호상정보량에 의해 질문단어와의 련관성을 계산하고 련관성크기에 따라 질문단어와 련관되는 중요단어를 추출하는 방법[1]을 고찰하였으며 단어의 토대득점값 및 동시출현빈도수, 최소의존성거리개념에 의해 단어의 중요득점값을 계산하는 방법[2]을 제기하였다.

론문에서는 단어의 출현빈도수와 거꿀문서출현빈도수의 리용과 함께 중요단어들과의 련관성이 큰 단어일수록 중요하다고 보는 가정밑에서 단어의 중요도값을 계산하는 한가지 수학적모형을 제기하고 실험을 통하여 그 타당성을 론증하였다.

1. 단어중요도값계산모형

문서를 구성하는 단어들이 중요하다는것은 단어가 담고있는 정보의 중요성이다. 그러므로 단어의 중요성을 론의하는데서 가장 중요한 인자는 그 단어가 해당 문서에서 얼마나 많이 출현하는가 하는것이다.

그러나 여러개의 문서를 동시에 론의하는 검색과 요약에서 단어의 출현빈도수만을 가지고 그 단어가 정보적중요성이 큰 단어라고 말할수 없다. 여기로부터 단어의 출현빈도수와 함께 단어의 거꿀문서출현빈도수를 론의하게 된다.

이와 함께 론문에서는 단어들과 많은 련관성을 가지는 단어일수록 중요한 단어로 될 가능성이 크다는것을 가정하고 이것을 단어의 정보적중요성을 평가하는 또 다른 하나의 인자로 보았다.

이러한 고찰로부터 단어 w_i의 중요도값을 $e(w_i)$라고 할 때 이 중요도값을 다음과 같은 수학적모형식을 리용하여 계산하였다.

$$e(w_i)=\alpha\sum_{j=1}^{n}A(w_i,w_j)e(w_j)+\beta b(w_i),\quad 1\le i\le n$$
$$b(w_i)=f_i\times\log_2\frac{N}{n_i},\quad \alpha,\beta\in[0,1],\ \alpha+\beta=1 \tag{1}$$

여기서 $A(w_i, w_j)$는 단어 w_i와 w_j의 련관성을 특징짓는 특징값이며 $b(w_i)$는 단어 w_i의

*tf-idf*값 즉 단어의 출현빈도수와 거꿀문서출현빈도수를 고려한 값이다. 그리고 n은 단어의 총수이다.

단어들사이의 련관관계를 나타내는 특징값 $A(w_i, w_j)$는 다음과 같이 계산하였다.

$$A(w_i,\ w_j)=\begin{cases} I(w_i,\ w_j)\cdot\exp[-\lambda\rho(w_i, w_j)],\ i\neq j \\ 0, \qquad\qquad\qquad\qquad\quad i=j \end{cases} \tag{2}$$

여기서 $I(w_1, w_2)$는 단어 w_i와 w_j의 호상정보량을 특징짓는 량이다.

$$I(w_1, w_2)=\log_2\left[1+\frac{p(w_1, w_2)}{p(w_1)p(w_2)}\right]$$

그리고 $p(w_i)$와 $p(w_j)$는 각각 문서내에서의 단어 w_i와 w_j들의 출현확률이며 $p(w_1, w_2)$는 문장단위의 동시출현확률이다. $\rho(w_i, w_j)$는 w_i와 w_j사이에 놓이는 단어수로서 단어들사이의 거리를 특징짓는 량이다.

련립방정식 (1)의 풀이 $e(w_i)$는 단어 w_i의 중요도값으로서 이 값이 큰 단어일수록 중요한 단어로 간주한다. 분명히 식 (1)의 비동차항 $b(w_i)$는 단어 w_i의 고유한 중요도값이며 앞의 항 $\sum_{j=1}^{n} A(w_i, w_j)e(w_j)$는 단어들사이의 련관성을 고려한 중요도값이다.

2. 중요단어추출알고리듬

1) 단어의 중요도값 $e(w_i)$의 계산

$e(w_i)$를 계산하자면 식 (1)로 주어지는 방정식을 풀어야 한다.

식 (1)을 행렬로 표현하기 위해 먼저 행렬 W 그리고 2개의 벡토르 $\boldsymbol{e}$와 $\boldsymbol{b}$를 각각 정의한다.

$$W=\alpha\cdot\begin{bmatrix} A(w_1,w_1) & A(w_1,w_2) & \dots & A(w_1,w_n) \\ A(w_2,w_1) & A(w_2,w_2) & \dots & A(w_2,w_n) \\ \vdots & \vdots & \ddots & \vdots \\ A(w_n,w_1) & A(w_n,w_2) & \dots & A(w_n,w_n) \end{bmatrix},$$

$$\boldsymbol{e}=\begin{bmatrix} e(w_1) \\ e(w_2) \\ \vdots \\ e(w_n) \end{bmatrix}, \qquad \boldsymbol{b}=\beta\cdot\begin{bmatrix} b(w_1) \\ b(w_2) \\ \vdots \\ b(w_n) \end{bmatrix}$$

이때 방정식 (1)을 행렬－벡토르형식으로 쓰면 다음과 같다.

$$\boldsymbol{e}=W\boldsymbol{e}+\boldsymbol{b} \tag{3}$$

$$(I-W)\boldsymbol{e}=\boldsymbol{b} \tag{4}$$

행렬 I는 대각선성분들이 모두 1인 단위행렬이다. 이때 방정식 (4)의 결수행렬 $(I-W)$는 대각선상의 원소들은 모두 1이고 비대각선상의 원소들은 모두 1보다 갈거나 작은 부 아닌 실수값들이다.

행렬 W의 매 원소들에 감쇠인자 $0<\theta<1$을 곱하며 비동차벡토르 $\boldsymbol{b}$의 원소들의 값이 최대로 1이 되도록 정규화하여 다음의 련립방정식을 얻는다.

$$(I-\theta\cdot W)\boldsymbol{e}=\hat{\boldsymbol{b}} \tag{5}$$

론문에서는 결수행렬 $(I-\theta\cdot W)$ 가 강한 대각선우세행렬로 되도록 감쇠인자 θ=0.6을 설정하였다. $\hat{\boldsymbol{b}}$ 은 비동차벡토르 $\boldsymbol{b}$를 정규화하여 얻은 벡토르이다.

$$\hat{\boldsymbol{b}}=\beta\cdot\begin{bmatrix}\hat{b}(w_1)\\ \hat{b}(w_2)\\ \vdots\\ \hat{b}(w_n)\end{bmatrix},\quad \hat{\boldsymbol{b}}(w_i)=\frac{\boldsymbol{b}(w_i)}{\sum_j \boldsymbol{b}(w_i)}$$

2) 중요단어추출알고리듬

① 입력본문문서모임 S를 구성하는 단어들에 대한 방정식 (5)를 작성한다.

② 가우스자이델반복법을 리용하여 주어진 정확도를 가지는 방정식 (5)의 근사풀이 $\hat{e}(w_i)$ 를 구한다.

③ 주어진 턱값 T에 대하여 $T\le\hat{e}(w_i)$ 를 만족시키는 단어 w_i들을 중요단어로 설정한다.

3. 실험결과 및 분석

론문에서 제기한 중요단어추출방법의 성능을 평가하기 위하여 선행연구[3]와 비교실험을 진행하였다. 비교실험은 두가지 방법으로 하였다.

방법 1 모형식의 비동차벡토르를 단어의 출현빈도수와 거꿀문서출현빈도수만을 리용하여 작성한 경우

방법 2 모형식의 비동차벡토르에서 단어의 출현빈도수, 거꿀문서출현빈도수와 고유실체단어에 대응하는 단어에 대하여 그것의 2배값을 고려한 경우

실험에 리용한 문서모임은 신문에 실린 기사자료 72건으로 된 다중문서모임이다.

전문가에 의해 작성된 정답중요단어모임과 체계가 출력시킨 중요단어모임들을 가지고 적중률, 완전률, F—척도를 리용하여 체계의 성능을 평가하였다.

적중률, 완전률 그리고 F—척도들은 다음과 같이 계산된다.

적중률: $P=\dfrac{|S_s\cap S_h|}{|S_s|}$

완전률: $R=\dfrac{|S_s\cap S_h|}{|S_h|}$

여기서 S_h는 전문가에 의해 작성된 정답중요단어모임, S_s는 체계에 의해 출력된 중요단어들의 모임, $|S|$는 모임 S에 들어있는 중요단어의 개수이다.

F—척도: $F=\dfrac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}$ (β=1 혹은 3)

실험결과는 표와 같다.

표에서 보여주는바와 같이 론문에서 제안한 방법 1과 2의 결과는 선행방법[3]의 F—값에 비하여 각각 1.06과 1.10배로 높아졌다.

표. 실 험 결 과

방 법	P	R	$F(\beta=3)$
선행방법[3]	0.534	0.623	0.522
제안방법 1	0.576	0.653	0.554
제안방법 2	0.583	0.664	0.576

맺 는 말

론문에서는 단어의 출현빈도수와 거꿀문서출현빈도수의 리용과 중요단어들과의 련관성이 큰 단어일수록 중요하다는 가정밑에서 단어의 중요도값을 계산하는 한가지 련립방정식수학모형을 제기하였다.

실험결과는 단어의 출현빈도수와 거꿀문서출현빈도수만을 리용하여 중요단어를 추출하는것에 비해 보다 좋은 결과를 주었으며 식 (1)의 비동차항에 고유실체단어의 중요성을 고려하면 고유실체단어와 련관관계가 있는 단어들이 중요단어로 더 많이 추출된다는것을 알수 있다. 중요단어의 중요도값을 문장벡토르모형에 반영한다면 문서요약을 비롯한 본문정보검색체계들에서 보다 좋은 결과를 줄수 있다.

참 고 문 헌

[1] JOURNAL OF **KIM IL SUNG** UNIVERSITY(Natural Science), 1, 4, 51 Juche101(2012).

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

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

주체107(2018)년 5월 5일 원고접수

A Method of Important Word Extraction by Association between Words

Sol Ryong Ui, *Jong Man Hung*

In this paper we suggested a mathematic model of computing importance value of word with occurrence frequency and inverse document occurrence frequency of word under assumption that the word is of importance if it is heavily linked with many important words.

Key words: word extraction, mutual information content, document summarizer