김일성종합대학학보
(정보과학)
주체107(2018)년
제64권 제1호

조밀분포에 의한 가변길이 패쎄지검색의 한가지 방법

한영진, 전진혁, 리청한

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《현시대는 과학과 기술의 시대이며 이르는 곳마다에서 요구하는것은 기술입니다. 기술을 몰라가지고서는 경제조직사업과 생산지휘를 바로할수 없으며 사회주의건설에 적극 이바지할수 없습니다.》(《**김정일**전집》 제2권 499～500페지)

사용자질문에 적합한 정답이 들어있는 패쎄지를 정확히 검색하는것은 질문응답체계의 성능을 높이는데서 중요한 문제로 나서고있다.

론문에서는 문서내에서 출현하는 질문용어들의 무게를 고찰한데 기초하여 패쎄지의 길이를 고려한 가변길이 패쎄지검색방법을 제기하고 그 효과성을 평가하였다.

질문처리, 문서검색, 패쎄지검색, 응답추출의 흐름으로 이루어진 질문응답체계에서 패쎄지검색은 순위화된 문서모임으로부터 질문에 대한 정답이 들어있는 일정한 크기의 문장모임인 패쎄지들을 추출하고 정확한 응답을 포함하는 정도에 따라 패쎄지들을 순위화하는 작업이다.[1]

질문에 대하여 그 어떤 후보패쎄지도 검색하지 못하거나 혹은 아주 많은 후보패쎄지들을 검색하면 질문응답체계의 전체적인 성능에 부정적영향을 미친다.

그러므로 패쎄지검색은 질문응답체계에서 가장 중요한 부분이며 패쎄지검색을 정확히 하는것은 전체적인 질문응답체계의 성능에 큰 영향을 미친다.

이로부터 대규모의 문서집합에서 질문에 적합한 패쎄지를 검색하는 방법들이 많이 제기되였다.[1－3]

패쎄지검색에서 가장 많이 리용되는 방법은 벡토르모형을 리용한 검색방법이다.[1, 2]

벡토르모형에서는 질문 q에 대한 패쎄지 p_j의 류사성등급을 패쎄지벡토르 $\boldsymbol{p}_j$와 질문벡토르 $\boldsymbol{q}$사이 각의 코시누스로 평가한다.

$$sim(p_j,\ q)=\frac{\boldsymbol{p}_j\cdot\boldsymbol{q}}{|\boldsymbol{p}_j|\times|\boldsymbol{q}|}=\frac{\sum_{i=1}^{t}\omega_{i,j}\times\omega_{i,q}}{\sqrt{\sum_{i=1}^{t}\omega_{i,j}^2}\times\sqrt{\sum_{i=1}^{t}\omega_{i,q}^2}}$$

이 방법의 부족점은 우선 패쎄지검색에서 고정된 크기(보통 300개 단어 혹은 3개 문장)의 패쎄지들을 미리 설정해놓고 사용자질문에 적합한 패쎄지를 검색하므로 사용자질문에 적합한 패쎄지가 서로 다른 패쎄지에 분할되는 경우가 있는것이다. 또한 해당한 패쎄지에 질문용어들의 분포정도를 고려함이 없이 질문용어들이 해당한 패쎄지내에 모두 있으면 그 패쎄지는 질문에 적합한것으로 판정되므로 이 경우 질문용어들이 조밀하게 분포될수록 사용자질문에 적합한 패쎄지로 될수 있다는 실험적사실에 부정적영향을 미치는것이다.

벡토르모형에 의한 패쎄지검색의 결함을 극복하기 위하여 득점함수에 의한 패쎄지검

색방법이 제기되였다.[4]

이 방법에서는 패쎄지 P의 득점최대값을 질문에 적합한 패쎄지로 정의한다.

$$PS_{PIDF(\beta)}(P) = \max_{p \in P} PS_{PIDF(\beta)}(p)$$

웃식에서 매 개별적인 패쎄지의 득점 $PS_{PIDF(\beta)}(p)$는 다음과 같다.

$$PS_{PIDF(\beta)}(p) = \exp(-\beta(r-l)) \sum_{q \in Q([l,\, r])} idf[q]$$

이 방법은 패쎄지의 크기를 반영한 방법(즉 질문용어들의 분포정도를 고려하였다.)으로서 사용자질문에 적합한 패쎄지검색에서 매우 효과적이다. 그러나 이 방법에서는 패쎄지의 크기관계와 질문용어의 거꿀무게빈도수만 고찰하고 매 질문용어의 무게를 고려하지 못한 부족점을 가지고있다.

1. 조밀분포에 의한 패쎄지검색

1) 패쎄지결정

질문에서 출현하는 질문용어들의 모임을 Q라고 하자.

그러면

$$Q = \{q_1, q_2, \cdots, q_t\}$$

이다.

문서 d_j에서 질문용어 q_i가 출현하는 가능한 단어위치모임을 H_i라고 하면

$$H_i = \{h_{i1}, h_{i2}, \cdots, h_{in_i}\}$$

이다. 여기서 h_{ik}는 질문용어 q_i가 문서 d_j에서의 k번째 단어위치, n_i는 H_i의 크기이다.

이로부터 질문용어모임 $Q = \{q_1, q_2, \cdots, q_i\}$가 문서 d_j에서 모두 출현하는 가능한 패쎄지들의 모임 Ω는 다음과 같다.

$$\Omega = \{(h_1, h_2, \cdots, h_t) \mid h_i \in H_i \ , \ i = \overline{1,\ t}\}$$

론문에서는 질문용어들을 모두 포함하는 문장 또는 단락들의 모임을 패쎄지로 결정한다.

패쎄지의 크기는 문서내에서 질문용어들의 분포정도에 따라 가변으로 주어진다.

2) 패쎄지에서 질문용어의 무게화된 분포

$a_j(l)(1 \le l \le L_j)$을 패쎄지 p_j에서 l번째 위치의 용어라고 하자. 여기서 L_j는 단어로 계산된 패쎄지 p_j의 길이이다.

질문 q에서 용어의 무게화된 분포 $b_j(l)$은 다음과 같이 정의된다.

$$b_j(l) = \begin{cases} w_{iq} \cdot idf_i, & a_j(l) = t_{iq}\text{일 때} \\ 0, & \text{기타 경우} \end{cases}$$

여기서 w_{iq}는 질문 q에 있는 i번째 용어의 무게, idf_i는 패쎄지 p_j에서 i번째 용어의 거꿀패쎄지빈도수, t_{iq}는 질문 q에 있는 i번째 용어이다.

w_{iq}와 idf_i는 $tf-idf$ 무게계산공식에 의하여 다음과 같이 계산된다.

$$\omega_{i,q}=\left(0.5+\frac{0.5freq_{i,q}}{\max_l freq_{l,q}}\right)\times idf_i$$

$$idf_i=\lg\frac{N}{n_i}$$

여기서 N은 패쎄지의 총개수, n_i는 질문용어 t_{iq}가 나타나는 패쎄지의 개수이다.

3) 패쎄지의 득점계산

패쎄지 p_j의 득점은 다음과 같이 계산한다.

$$p(j)=\left(\sum_{x=r}^{q}b_j(x)\right)\times\exp^{-\beta(q-r)}$$

여기서 r는 j번째 패쎄지의 첫 단어의 위치, q는 j번째 패쎄지의 마지막단어의 위치, $b_j(x)$는 질문용어의 무게화된 분포이다.

질문 q에 적합한 패쎄지는 다음과 같이 결정한다.

$$score(p,q)=\max\{p(j)\}$$

2. 성 능 평 가

론문에서는 조밀분포에 의한 가변길이 패쎄지검색의 성능을 평가하기 위하여 대상자료로서 《조선전사》(1～15권)에 기초하여 만든 720개의 표준질문과 응답패쎄지들을 준비하였다. 그리고 질문응답체계에서 패쎄지검색의 성능을 평가할 때 흔히 리용되는 MRR(거꿀순위평균)평가척도를 가지고 평가하였다.(표)

표. 성능평가

패쎄지검색방법	MRR
벡토르모형	0.403
거짓－반결합모형	0.413
득점함수에 의한 방법	0.422
제안된 방법	0.516

성능평가에서 보여주는바와 같이 질문용어의 분포밀도와 패쎄지의 크기를 고려한 방법의 MRR가 0.516으로서 성능이 가장 높다는것을 알수 있다.

맺 는 말

론문에서는 사용자질문에 따라 패쎄지의 크기를 동적으로 결정하고 패쎄지내에서 질문용어와 일치하는 매 용어의 무게를 계산한 다음 패쎄지의 크기를 고려한 패쎄지검색방법을 제안하고 그 성능을 평가하였다. 이 방법은 사용자질문이 보통 짧게 제기되는 질문응답체계의 패쎄지검색의 성능을 개선하는데서 매우 효과적이다.

참 고 문 헌

[1] Wei Xu et al.; Proceedings of the 5th International Joint Conference on Natural Language Processing, 1046, 2011.

[2] P. Knoth et al.; Proceedings of the 23rd International Conference on Computational Linguistics, 590, 2010.

[3] B. Wang, X. Wang; Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 1230, 2013.

[4] N. Foucault et al.; Proceedings of Recent Advances in Natural Language Processing, 716, 2011.

주체106(2017)년 11월 5일 원고접수

A Method of Variable Length Passage Retrieval based on Density Distribution

Han Yong Jin, *Jon Jin Hyok* and *Ri Chong Han*

In this paper, we proposed a method of variable length passage retrieval based on density distribution. This method is very effective for higer precision of passage retrieval of Korean question answering system.

Key words: passage, question answering, density distribution