김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제5호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 62 No. 5 JUCHE105(2016).

조선어질문응답체계에서 질문확장에 의한 문서검색의 한가지 방법

리청한, 정만흥, 김순실

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)

조선어질문응답체계를 구축하는데서 나서는 중요한 문제의 하나는 다량의 문서집합(collection)에서 질문에 적합한 정답이 들어있는 문서를 검색하여 질문응답에 참가하는 문서의 수를 줄이는것이다. 그것은 질문에 적합한 문서를 먼저 검색함으로써 제한된 시간안에 정답을 찾는것이 질문응답체계의 성능평가에서 매우 중요하기때문이다.

이로부터 대규모의 문서집합에서 질문에 적합한 문서를 검색하는 방법들이 많이 제안되였다.[1, 3]

문서검색에서 가장 많이 리용되는 방법은 일반정보검색에서 흔히 리용하는 벡토르모형을 리용한 검색방법, BM25검색모형과 확률모형을 리용한 검색방법들이다.[1, 2]

그러나 이 방법들은 문서의 크기에 비해 질문이 짧은 경우 검색의 정확도가 떨어지는 결함으로 하여 질문응답체계의 문서검색에서는 적합하지 않다.

이로부터 우리는 초기질문을 확장하는 방법과 확장된 질문을 리용하여 문서검색을 진행하는 방법을 제안한다.

1. 질문확장방법

일반적으로 질문응답체계에서는 사용자들로부터 제기되는 질문이 문서의 크기에 비해 매우 짧은것이 특징이다.

그러므로 우리는 질문응답체계의 특성으로부터 사용자질문이 짧은 경우 문서검색의 효률을 높이기 위한 방법을 제안한다.

우선 벡토르모형을 리용하여 초기질문을 가지고 문서검색을 진행하여 문서들을 순위화한다.

다음 순위화된 문서에서 웃준위문서(실험적으로 3개의 문서를 선택하였을 때가 제일 좋다.)의 용어들로 질문벡토르를 확장하고 이 질문벡토르를 리용하여 문서검색을 다음과 같이 한다.

문서 d_j를 t차원벡토르로 표현한다. 즉

$$\vec{d}_j = (\omega_{1j},\ \omega_{2j},\ \cdots,\ \omega_{tj})\ .$$

여기서 ω_{ij}는 문서 d_j에서 용어 tj의 무게이다.

마찬가지로 질문 q는 질문벡토르 $\vec{q} = (\omega_{1q},\ \omega_{2q},\ \cdots,\ \omega_{tq})$로 표현한다.

여기서 ω_{iq}는 질문 q에서 질문용어 tq의 무게이다.

이로부터 질문 q에 대한 문서 d_j의 류사성등급을 문서벡토르 $\vec{d}_j$와 질문벡토르 $\vec{q}$사이 각의 코시누스로 평가한다. 즉

$$\mathrm{sim}(d_j,\ q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} \omega_{i,\ j} \times \omega_{i,\ q}}{\sqrt{\sum_{i=1}^{t} \omega_{i,\ j}^2} \times \sqrt{\sum_{i=1}^{t} \omega_{i,\ q}^2}}\ .$$

여기서 $|\vec{d}_j|$와 $|\vec{q}|$는 각각 문서 및 질문벡토르의 노름이다.

한편 문서벡토르와 질문벡토르에서 용어무게는 여러가지 방법으로 계산할수 있지만 우리는 tf－idf 법을 리용하였다. 즉

$$\omega_{i,\ j} = f_{i,\ j} \times \log\frac{N}{n_i},\ \ \omega_{i,\ q} = \left(0.5 + \frac{0.5\,freq_{i,\ q}}{\max_l freq_{l,\ q}}\right) \times \log\frac{N}{n_i},\ \ f_{i,\ j} = \frac{freq_{i,\ j}}{\max_l freq_{l,\ j}}\ .$$

여기서 N은 문서의 총 개수, n_i는 질문용어 ti가 나타나는 문서의 수이다.

$\omega_{i,\ j} \geq 0$이고 $\omega_{i,\ q} \geq 0$이므로　$\mathrm{sim}(d_j,\ q) \in [0,\ 1]$이다.

이제 E를 질문확장을 위한 다음과 같은 문서벡토르모임이라고 하자.

$$E = \left\{ d_j^+ \,|\, \frac{\mathrm{sim}(d_j,\ q)}{\max\limits_i \mathrm{sim}(d_i,\ q)} \geq \tau \right\}$$

여기서 q는 초기질문벡토르, τ는 류사성턱값이다.(이 류사성턱값은 웃준위 3개의 문서가 선택되도록 정한다.)

E에서 문서벡토르의 합 d_s를

$$d_s = \sum_{d_j^+ \in E} d_j^+$$

라고 하면 이것은 초기질문에 대하여 확장된 정보로 볼수 있다.

따라서 확장된 질문벡토르 q'는 다음과 같다.

$$q' = \frac{q}{\|q\|} + \alpha \frac{d_s}{\|d_s\|}$$

여기서 α는 무게조종을 위한 파라메터로서 α의 값이 대체로 0.3～0.8사이에 놓인다는것을 얻었으며 $\alpha = 0.5$일 때 가장 좋은 결과가 얻어졌다.

최종적으로 확장된 질문으로 류사도 $\mathrm{sim}(d_j,\ q')$를 계산하여 문서들을 다시 순위화한다.

이런 방법으로 검색된 웃준위문서가 중복될 때까지 반복한다.

2. 질문확장에 의한 문서검색알고리듬

```
D[n][t];    // 문서벡토르모임
Q[t];       // 질문벡토르
sim[n];     //문서벡토르와 질문벡토르사이의 류사도
Ds[t];      //턱값을 넘는 문서벡토르들의 합
Qq[t];      // 개선된 질문벡토르
E=0;        //d의 부분문서벡토르
        For(i=1; i<=n; i++)
            sim[i]=calc_sim(d[i], q);
             Max_sim=max(sim);
        For(i=1; i<=n; i++)
            If((sim[i]/max_sim)>=k)
                Insert(d[i],E);
        For(i=1; i<=size(E);i++)
            For(j=1; j<=n; j++)
                Ds[j]+=E[i][j];
        For(i=1;i<=n;i++)
            Qq[i]=q[i]/norum(q)+h*ds[i]/norum(ds);
       For(i=1;i<=n; i++)
            sim[i]=calc_sim(d[i],qq);
      // calc_sim함수
            D:문서벡토르 1차원(크기 t)
            Q:질문벡토르 1차원(크기 t)
        Function  calc_sim(d,q){
            S=0;
          For(i=1; i<=t; i++){
              S+=d[i]*q[i];
               s/=norum(d)*norum(q)}
            Return s;}
        // norum 함수
            V: 1차원배렬(크기 t)
              Function     norum(v){
                 S=0;
                  For(i=1;i<=n; i++){
                        S+=v[i];}
                        Return s;}
```

알고리듬은 문서의 왼쪽에서부터 시작하여 오론쪽으로 탐색을 진행한다. 이때 탐색기의 초기상태는 문서 D에서 나타나는 모든 질문용어의 가장 왼쪽위치목록에 의하여 주어진다. 여기서 insert((I, q), L)은 i의 증가순서로 정렬된 목록 L에 (I, q)를 삽입하는 함수, remove(L)

은 목록 L의 첫 요소를 제거하는 함수, DS는 문서에서 가장 높은 문서의 득점값이다.

질문확장에 의한 문서검색알고리듬의 계산시간은 $O(kn)$이다. 여기서 n은 D에서 질문용어의 총빈도수이고 k는 질문용어의 수이다.

3. 실험결과 및 분석

론문에서는 질문확장에 의한 문서검색의 성능을 평가하기 위하여 대상자료로서 《조선전사》(1～15권)에 기초하여 만든 360개의 표준질문과 응답문서들을 준비하였다. 그리고 질문응답체계에서 문서검색의 성능을 평가할 때 흔히 리용되는 MRR(거꿀순위평균) 평가척도를 가지고 평가하였다.(표)

표. 검색성능평가결과

검색모형	MRR	검색모형	MRR
벡토르검색모형	0.408	확률검색모형	0.502
BM25검색모형	0.504	제안된 방법	0.523

표에서 알수 있는바와 같이 제안된 방법이 선행한 방법보다 우월하다는것을 알수 있다.

맺 는 말

질문확장에 의한 문서검색방법은 사용자질문이 짧게 제기되는 질문응답체계의 특징을 반영한 문서검색방법으로서 이것은 질문응답체계의 문서검색에서 검색의 정확도와 체계의 응답시간을 단축할수 있게 한다.

참 고 문 헌

[1] Wei Xu et al.; Proceedings of the 5th International Joint Conference on Natural Language Processing, **11**, 1046, 2011.

[2] P. Knoth et al.; Proceedings of the 23rd International Conference on Computational Linguistics, **8**, 590, 2010.

[3] N. Foucault et al.; Proceedings of Recent Advances in Natural Language Processing, **9**, 716, 2011.

주체105(2016)년 1월 5일 원고접수

A Method of Document Retrieval using Query Expansion in Korean Question Answering System

Ri Chong Han, *Jong Man Hung* and *Kim Sun Sil*

We discussed the document retrieval method using query expansion, a new method for document retrieval. We retrieved the documents with initial query using vector model and ranked the documents.

Then we expanded the query vector with terms of upper level documents in ranked documents, and retrieved the document again using these query vectors.

Key words: document retrieval, query expansion