김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제12호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 62 No. 12 JUCHE105(2016).

# 잠재의미색인화모형에 의한 패쎄지검색방법

**리 청 한**

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)

질문응답체계에서 중요한 문제의 하나는 다량의 문서집합에서 질문에 대한 정답이 들어있는 패쎄지를 검색하는것이다. 그것은 질문에 적합한 정답이 들어있는 패쎄지를 먼저 검색함으로써 제한된 시간안에 패쎄지속에 포함된 정답을 찾는것이 질문응답체계의 성능평가에서 매우 중요하기때문이다.[4] 이로부터 대규모의 문서집합에서 질문에 적합한 패쎄지를 검색하는 방법들이 많이 제안되였다.[1－3]

패쎄지검색에서 가장 많이 리용되는 방법으로서는 일반정보검색에서 흔히 리용하는 벡토르모형을 리용한 검색방법[1], BM25검색모형과 확률모형을 리용한 검색방법[2]이 있다. 그러나 이 방법들은 문서의 크기에 비해 질문이 짧은 경우 검색의 정확도가 떨어지는 결함으로 하여 질문응답체계의 패쎄지검색에서는 적합하지 않다.

론문에서는 사용자의 질문이 짧은 경우 패쎄지검색에 효과적인 방법인 잠재의미색인화모형에 의한 패쎄지검색방법을 제기한다.

## 1. 잠재의미색인화모형에 의한 패쎄지검색

패쎄지집합 $P$를 $P=(\hat{p}_1,\ \hat{p}_2,\ \cdots,\ \hat{p}_n)$에 의하여 정의된 패쎄지행렬이라고 하자.
여기서

$$\hat{p}_j=\frac{p_j}{\|p_j\|}$$

이다.

이때 $P$를 특이값분해하면 $P$는 3개의 특징을 가지는 행렬로 분해된다. 즉

$$P=USV^T.$$

여기서 $U$와 $V$는 각각 크기가 $m\times r$, $n\times r(r=\mathrm{rank}(P))$인 행렬이다. 그리고 $S$는 특이값 $\sigma_i(\sigma_i\geq\sigma_j,\ i\leq j)$를 가지는 대각선행렬로서 다음과 같이 표시된다.

$$S=\mathrm{diag}(\sigma_1,\ \sigma_2,\ \cdots,\ \sigma_r)$$

한편 $U$와 $V$에서 대응하는 렬과 함께 $S$에 들어있는 가장 큰 특이값 $k(<r)$를 유지하면 $P$는 $P_k=U_kS_kV_k^T$에 의하여 근사되는데 이 근사값은 용어들사이의 잠재의미관계를 나타낸다. 여기서 $U_k$, $S_k$, $V_k$는 각각 크기가 $m\times k$, $k\times k$, $n\times k$인 행렬이다.

여기에 기초하여 패쎄지와 질문사이의 류사성을 계산해보자. 이를 위해 $V_k=(v_{ji})$ 에서 $v_j=(v_{j1}, \cdots, v_{jk})$ $(1\le j\le n,\ 1\le i\le k)$ 를 행벡토르라고 하자. 그러면 $k$ 차원공간에서 패쎄지 $p_j$ 는 $p_j^*=S_kv_j^T$ 로 표현되고 초기질문 역시 $q^*=U_k^Tq$ 로 표현된다.

결국 질문과 패쎄지의 류사성은 $\mathrm{sim}(p_j^*,\ q^*)$ 에 의하여 얻어진다.

## 2. 잠재의미색인화모형에 의한 패쎄지검색실현

패쎄지에서 $P_0\subset P_1$ 이고 $Q(P_0)=Q(P_1)$ 이면 $PS_{\mathrm{DIDF}}(P_0)>PS_{\mathrm{DIDF}}(P_1)$ 로 된다.

정의 패쎄지 $P$ 에 대하여 $P'\subset P$ 이고 $Q(P')=Q(P)$ 를 만족시키는 $P'$ 가 없다면 $P$ 를 질문용어최소패쎄지($Q$-minimal passage)라고 한다.

이러한 정의에 기초하여 [1, $|P|$]에서 질문용어최소패쎄지를 검색하는 알고리듬은 다음과 같다.

```
scanner=[];
size=0;
PS=0;
for each q in Q([1, N])  {
Insert((PL[q][0], q), scanner);
size=size+1;
}
while(size>0)  {
      (L, q1)=scanner[0];
      idfsum=0;
      for (r=0; r<size; r=r+1)  {
            (R, qr)=scanner[r];
            idfsum=idfsum+idf[qr];
            PS=idfsum*exp(－β*(R－L));
}
 Remove(scanner);
 size=size－1;
 remove(PL[q1]);
 if (PL[q1] exist) {
insert((PL[q][0], q1), scanner);
size=size+1;  }  }
```

여기서 insert((I, q), L)은 i의 증가순서로 정렬된 목록 L에 (I, q)를 삽입하는 함수, remove(L)은 L의 첫 요소를 제거하는 함수이며 PS는 득점값이 가장 높은 패쎄지를 나타낸다. 그리고 매 질문용어 $q\in Q(1,\ N)$ 에 대하여 PL[q]는 위치목록, PL[q][h]는 D에서 q의 h번째 위치를 나타낸다.

탐색기의 초기상태는 패쎄지 $P$에서 나타나는 모든 질문용어의 가장 왼쪽 위치목록에 의하여 주어진다.

한편 잠재의미색인화모형에 의한 패쎄지검색알고리듬의 계산시간은 $O(kn)$이다. 여기서 $n$은 $P$에서 질문용어의 총빈도수이고 $k$는 질문용어의 수이다.

## 3. 실험결과 및 분석

론문에서는 잠재의미색인화모형에 의한 패쎄지검색의 성능을 평가하기 위하여 대상자료로서 《조선전사》(1～15권)에 기초하여 만든 360개의 표준질문과 응답패쎄지들을 준비하였다. 여기에 기초하여 질문응답체계에서 패쎄지검색성능을 평가할 때 흔히 리용되는 MRR(거꿀순위평균)평가척도를 가지고 평가하였는데 그 결과는 표와 같다. 표에서 보는바와 같이 제안된 방법이 선행한 방법들보다 검색정확도가 높다는것을 알수 있다.

표. 검색성능평가결과

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

## 맺 는 말

잠재의미색인화모형에 의한 패쎄지검색방법은 사용자질문이 짧게 제기되는 질문응답체계의 특징을 반영한 패쎄지검색방법으로서 질문응답체계의 패쎄지검색에서 검색의 정확도와 체계의 응답시간을 단축할수 있다.

## 참 고 문 헌

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

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

[3] Baoxun Wang et al.; Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 1230, 2010.

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

주체105(2016)년 8월 5일 원고접수

## A Method of Passage Retrieval by using Latent Semantic Indexing

*Ri Chong Han*

We propose a passage retrieval method by using latent semantic indexing model to improve the well-known VSM and experimentally show that the retrieval on latent semantic indexing is also advantageous for dealing with short queries on condition that documents are long.

Key words: passage retrieval, question answering system, model