김일성종합대학학보
정보과학

주체109(2020)년
제66권 제4호

전통적인 문서검색방법과 패쎄지에 기초한 문서검색방법의 실험적평가

리 청 한

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《과학기술을 발전시키는것은 나라의 경제를 빨리 발전시키기 위한 중요한 담보입니다.》(《**김정일**선집》 증보판 제11권 133페지)

일반적으로 문서의 크기에 비해 검색질문의 길이가 짧은 경우 사용자의 의도에 맞는 정확한 문서를 검색하는것은 힘든 일이다.[1]

왜냐하면 문서의 크기가 크면 클수록 그 문서에는 사용자의 질문에 적합한 부분뿐아니라 여러가지 의미를 담은 다른 내용도 포함되여있기때문이다.[2]

이로부터 패쎄지검색이라고 부르는 새로운 검색체계를 사용하여 이 문제를 해결하려는 연구[3]가 본격적으로 진행되고있다.

론문에서는 문서의 길이에 비해 사용자의 질문이 짧은 경우 조밀분포에 기초한 패쎄지검색방법을 리용하여 문서검색의 성능을 개선하기 위한 방법을 제기하고 선행한 문서검색방법들인 벡토르모형(VSM), 잠재의미색인작성(LSI)방법과 거짓－반결합모형(Pseudo-feedback)에 의한 문서검색방법과의 성능비교를 통하여 패쎄지에 의한 문서검색방법의 우점을 검증하였다.

이 방법은 문서의 크기에 비해 질문이 짧은 질문응답체계의 문서검색에서 매우 효과적이다.

1. 전통적인 문서검색방법

문서검색의 기본목적은 문서집합(collection)으로부터 사용자의 질문에 적합한 순위화된 문서를 검색하여 사용자에게 제공하는것이다.

1) VSM모형에 의한 문서검색

VSM모형은 문서검색을 위한 가장 대표적인 검색모형이다.

VSM모형에서 문서 d_j는 다음과 같이 m차원벡토르로 표현된다.

$$d_j = (w_{1j}, w_{2j}, \cdots, w_{mj})^{\mathrm{T}}$$

여기서 T는 전위행렬, w_{ij}는 문서 d_j에서 용어 t_i의 무게값이다. 그리고 질문 q는 다음과 같이 표현된다.

$$q = (w_{1q}, w_{2q}, \cdots, w_{mq})$$

여기서 w_{iq}는 질문 q에서 용어 t_i의 무게값이다.

론문에서는 w_{ij}를 계산하기 위하여 다음과 같이 정의된 *tf-idf*용어무게달기규칙을 리용한다.

$$w_{ij} = tf_{ij} \cdot idf_i$$

여기서 tf_{ij}는 용어빈도 f_{ij}를 사용하여 계산된 무게값이고 idf_i는 거꿀문서빈도 n_i(용어 t_i를 포함하는 문서의 수)를 사용하여 계산된 무게값이다.

tf_{ij}와 idf_i계산은 보통 다음과 같이 한다.

$$tf_{ij} = \sqrt{f_{ij}}\ ,\ \ idf_i = \log(n / n_i)$$

여기서 n은 전체 문서의 개수이다.

마찬가지방법으로 무게 w_{iq}는 $w_{iq} = \sqrt{f_{iq}}$로 정의된다. 여기서 f_{iq}는 질문 q에서 용어 t_i가 출현하는 개수이다.

문서 d_j와 질문 q사이의 류사성 $sim(d_j, q)$는 문서 d_j와 질문 q사이의 각의 코시누스로 측정된다.

$$sim(d_j,\ q) = \frac{d_j^{\mathrm{T}} \cdot q}{\| d_j \| \times \| q \|}$$

2) Pseudo-feedback에 의한 문서검색

VSM모형에 의한 문서검색에서 제기되는 기본문제점은 사용자의 질문이 짧은 경우 문서의 순위화에 부정적인 영향을 준다는것이다.

이 문제를 해결하기 위하여 초기질문을 검색된 가장 웃준위문서의 용어들로 질문을 확장하는 Pseudo-feedback방법이 제안되였다.

이 방법은 우선 초기질문으로 검색된 문서들을 순위화한다.

그다음 가장 웃준위문서내에 있는 용어들로 초기질문을 확장한다. 문서들은 확장된 질문에 의하여 다시 순위화된다.

방법은 다음과 같다.

E를 다음과 같이 계산된 확장을 위한 문서벡토르모임이라고 하자.

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

여기서 q는 초기질문벡토르이고 τ는 류사성턱값이다.

모임 E에 있는 문서벡토르 d_j^{+}들의 합 d_s는 다음과 같다.

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

이로부터 확장된 질문벡토르 q'는 다음과 같이 얻어진다.

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

여기서 α는 무게조종을 위한 파라메터이다.

최종적으로 문서는 확장된 질문벡토르를 포함하는 류사성척도 $sim(d_j, q')$에 의하여 다시 순위화된다.

3) LSI에 의한 문서검색

LSI는 VSM에 의한 문서검색방법을 개선하기 위하여 리용되는 잘 알려진 방법이다.

행렬 D는 $D=(\hat{d}_1, \cdots, \hat{d}_n)$에 의하여 정의된 문서행렬이라고 하자. 여기서

$$\hat{d}_j = \frac{d_j}{\| d_j \|}$$

이다.

특이값분해를 적용하면 D는 3개의 특징행렬로 분해된다. 즉

$$D = USV^{\mathrm{T}}$$

이다. 여기서 U와 V는 $m \times r$ 와 $n \times r$ ($r = \mathrm{rank}(D)$) 크기의 행렬이고 $S = \mathrm{diag}(\sigma_1, \cdots, \sigma_n)$ 은 특이값 σ_i(만일 $i \le j$이라면 $\sigma_i \ge \sigma_j$) 를 가지는 대각선행렬이다. U(V)에서 매 행벡토르는 용어를 표현하는 r차원벡토르이다.

U와 V에서 대응하는 렬과 함께 S에 있는 가장 큰 특이값 D는 $D_k = U_k S_k V_k^{\mathrm{T}}$ 에 의하여 근사된다. 여기서 U_k, S_k, V_k 는 각각 $m \times k, k \times k, n \times k$ 크기의 행렬이다.

이 근사값은 문서들과 마찬가지로 용어들사이에 잠재의미관계를 나타내도록 한다.

문서와 질문사이의 류사성은 다음과 같이 측정한다.

$v_j = (v_{j1}, \cdots, v_{jk})$ 를 $V_k = (v_{ji})$ $(1 \le j \le n, 1 \le i \le k)$ 에 있는 행벡토르라고 하자

그러면 k 차원공간에서 문서 d_j 는 $d_j^* = S_k v_j^{\mathrm{T}}$ 로 표현된다.

초기질문 역시 $q^* = U_k^{\mathrm{T}} q$ 와 같이 k 차원공간에서 표현된다. 즉 질문과 문서사이의 류사성은 $sim(d_j^*, q^*)$ 에 의하여 얻어진다.

2. 패쎄지에 기초한 문서검색

패쎄지에 기초한 문서검색에서 패쎄지의 크기를 결정하는 방법에는 보통 문맥에 의한 방법, 의미에 기초한 방법, 창문에 기초한 방법이 있다.

론문에서는 조밀분포(Density Distribution)라고 불리우는 창문패쎄지를 리용하였다.

조밀분포의 기본의미는 질문에 있는 용어를 조밀하게 포함하는 문서가 질문에 적합하다는것이다.

그림에 조밀분포의 실례를 보여주었다.

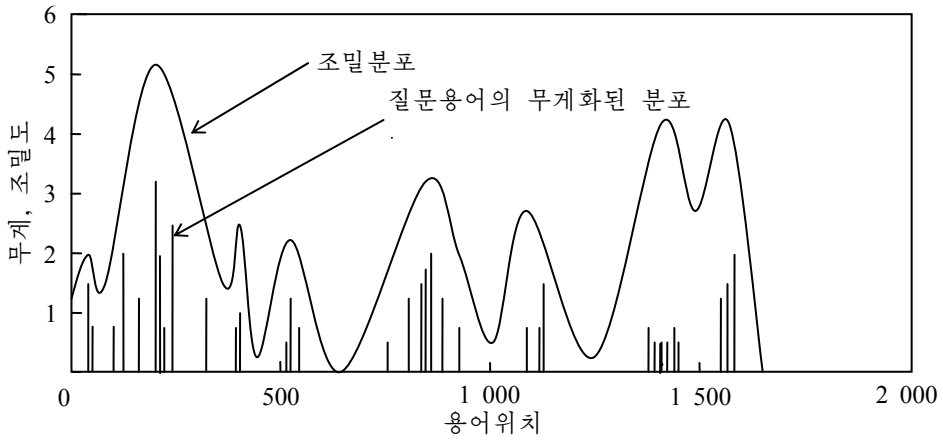

그림. 조밀분포의 실례

그림에서 수평축은 문서에 있는 용어들의 위치이다. 그리고 문서에서 질문용어의 분포는 그림에서 막대도표로 표시되며 그 막대도표의 높이는 용어의 무게와 같다. 그림에서 조밀분포는 창문함수를 리용하여 막대도표를 원활화하는것으로 얻어진다.

$a_j(l)(1 \le l \le L_j)$가 문서 d_j에서 l번째 위치에 있는 용어라고 하자. 여기서 L_j는 단어의 개수로 계산된 문서 d_j의 길이이다.

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

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

문서 d_j에 대하여 $b_j(l)$을 원활화하면 조밀분포 $dd_j(l)$을 얻을수 있다.

$$dd_j(l) = \sum_{x=-W/2}^{W/2} f(x) b_j(l-x)$$

여기서 $f(x)$는 창문크기가 W인 창문함수인데 그 창문함수는 다음과 같이 정의된다.

$$f(x) = \begin{cases} \frac{1}{2}\left(1 + \cos 2\pi \frac{x}{W}\right), & |x| \le W/2\text{인 경우} \\ 0, & \text{기타 경우} \end{cases}$$

질문 q에 대한 문서 d_j의 득점은 조밀분포의 최대값으로 얻어진다.

$$score(d_j, q) = \max_l \{dd_j(l)\}$$

3. 실험 및 결과분석

론문에서는 전통적인 문서검색방법들과 패쎄지에 의한 문서검색의 성능을 평가하기 위하여 문서의 크기와 길이가 서로 다른 IT부문, 의학부문, 생물부문, 물리부문의 도서들을 가지고 평가를 진행하였다.(표 1)

표 1. 실험에 참가한 시험모임의 통계

분 류	IT부문	의학부문	생물부문	물리부문
크기/MB	1.1	1.6	235	209
문서의 수/개	1 033	1 398	27 922	19 789
용어의 수/개	8 870	5 276	45 717	50 866
평균문서길이/개	143	153	1 745	2 320
질문의 수/개	30	225	34	112
평균질문길이/개	7.4	8.5	4.5	4.3

표 1에서 보는바와 같이 IT부문의 도서와 의학부문의 도서는 생물부문의 도서와 물리부문의 도서보다 크기와 문서의 개수, 평균문서길이를 훨씬 작게 하였으며 질문의 크기는 IT부문과 의학부문보다 생물부문과 물리부문을 작게 하였다.

선행한 문서검색방법(VSM, PS, LSI)들과 패쎄지에 의한 문서검색방법이 문서의 길이와 질문의 길이에 어떤 영향을 주는가를 고찰하였다.(표 2)

표 2에서 보는바와 같이 문서의 길이가 작은 IT부문과 의학부문에 대해서 LSI와

PF는 VSM과 DD보다 더 좋은 결과를 산생하였으나 긴 문서집합인 생물부문과 물리부문 집합에서는 질문의 크기가 작을 때 DD가 가장 좋은 결과를 산생하였다.

표 2. 모든 적합문서에 대한 평균적중률

모 형	IT부문	의학부문	생물부문	물리부문
VSM	0.512	0.375	0.124	0.071
PF	0.622	**0.427**	0.173	0.087
LSI	**0.704**	0.413	0.076	0.043
DD	0.491	0.357	**0.175**	**0.198**

맺 는 말

론문에서는 문서의 길이가 큰데 비해 사용자의 질문이 짧은 경우 가장 효과적인 문서검색방법이 어느 방법인가를 실험을 통하여 검증하였다.

검증결과 조밀분포(DD)에 의한 패쎄지검색방법이 가장 효과적이라는것이 확증되였다. 결국 문서의 크기에 비해 사용자의 질문이 짧게 제기되는 질문응답체계에서의 문서검색은 조밀분포에 의한 패쎄지검색방법이 가장 효과적이라는것을 확증하였다. 앞으로 창문의 크기 W를 자동적으로 결정하는 문제와 이 방법을 웨브문서에 어떻게 적용하겠는가 하는 문제가 더 연구되여야 한다.

참 고 문 헌

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

[2] Mio Kobayashi et al.; Proceedings of the 8th International Joint Conference on Natural Language Processing, 967, 2017.

[3] R. Berant et al.; Proceedings of Empirical Methods in Natural Language Processing, 1533, 2013.

주체109(2020)년 8월 5일 원고접수

Experimental Evaluation of Conventional Document Retrieval and Passage-based Document Retrieval Method

Ri Chong Han

In this paper, we experimentally show that the passage-based retrieval is also advantageous for dealing with short queries on condition that documents are long. We employ a passage-based method based on density distributions of query terms in documents, and compare it with three conventional methods-the vector space model, pseudo-feedback and latent semantic indexing.

Keywords: document retrieval, passage retrieval, density distribution