김일성종합대학학보 주체108(2019)년
정보과학 제65권 제2호

질문리력발굴에 의한 질문추천의 한가지 방법

우준민, 김태균

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《과학자, 기술자들은 자기 땅에 발을 붙이고 눈은 세계를 보는 혁신적인 안목을 가지고 두뇌전, 실력전을 벌려 최첨단체신기술과 수단들을 더 많이 연구개발하여야 합니다.》

론문에서는 질문리력발굴에 의한 질문추천의 한가지 방법에 대하여 고찰하였다.

선행방법들[1, 2]은 리력자료에 제출된 질문들의 공간이 성긴것으로 하여 질문들사이의 의미적류사성을 반영할수 없다.

이로부터 다음과 같은 문제점을 설정하였다.

첫째로, 사용자행위자료를 리용하여 질문무리짓기를 하고 질문을 추천하는 방법을 제안한다.

둘째로, 실험을 통하여 제안한 방법의 효과성을 검증한다.

1. 질문추천방법

일반적으로 대용질문을 찾아내는 목적은 검색기구의 문서정렬을 개선하여 검색체계의 성능을 높이기 위해서이다.

결과목록의 순위를 개선하자면 질문과 질문사이의 관계뿐아니라 질문과 련관있는 문서들과의 관계를 밝혀야 한다.

문서가 질문쎄션에서 일정한 정도로 선택되였으면 그 질문과 문서가 일치한다고 약속하자. 그리고 문서모임의 매개가 질문과 일치하면 그 질문이 문서모임과 일치한다고 볼수 있다.

많은 동일한 질문들이 서로 다른 사용자의 요구를 반영할수 있으며 사용자의 의도에 따라 문서들의 특정한 부분모임을 선택할수 있다.

그러므로 질문쎄션에서 선택한 문서모임은 원천질문의 부분주제를 나타낸다. 따라서 질문쎄션기간에 선택된 문서들사이의 관계를 평가하고 질문무리들을 창조하며 매 무리들과 관계되는 질문들을 확정하여 질문을 추천하면 응답결과목록의 순위를 개선할수 있다.

$D(S_q)$를 질문 q의 쎄션 S_q기간에 선택되는 문서들의 모임이라고 하자.

$D(S_q)$가 질문 q의 정보를 표현한다는 가정을 세우면 $D(S_q)$와 일치하는 다른 질문들은 $D(S_q)$의 문서들을 더 잘 정렬할수 있다. 만일 이런 질문들이 존재한다면 원천질문의 매 쎄션에 대하여 우의 과정을 반복하면서 일정한 수의 쎄션들에서 나타나는 추천가능한 질문들을 선택하고 그것들을 q에 흥미를 가지는 사용자들에게 추천한다.

2개의 서로 다른 질문들에 대한 문서모임의 순위화를 비교하기 위하여서는 기준이 있어야 한다. 먼저 질문에서 문서의 순위에 대하여 정의하자.

질문 q에 대한 문서 u의 순위 $r(u, q)$는 검색기구에서 나온 대답목록에서 문서 u의 위치라고 하자.

그러면 질문 q에서 문서모임 U의 순위는 다음과 같이 정의된다.

$$r(U, q) = \max_{u \in U} r(u, q)$$

다시말하여 가장 낮은 순위값을 가지는 문서가 모임의 순위를 결정한다. 직관적으로 1개의 문서모임이 질문 q_a에서 질문 q_b보다 더 높은 순위를 기록하면 q_a가 q_b보다 문서들을 더 잘 정렬한다고 말한다.

U를 q_a와 q_b사이의 일반문서모임이라고 할 때 $r(U, q_a) < r(U, q_b)$이면 질문 q_a는 질문 q_b보다 문서모임 U를 더 잘 정렬한다.

2개 문서를 포함하는 쎄션에 대하여 두 질문의 순위화비교는 그림과 같다.

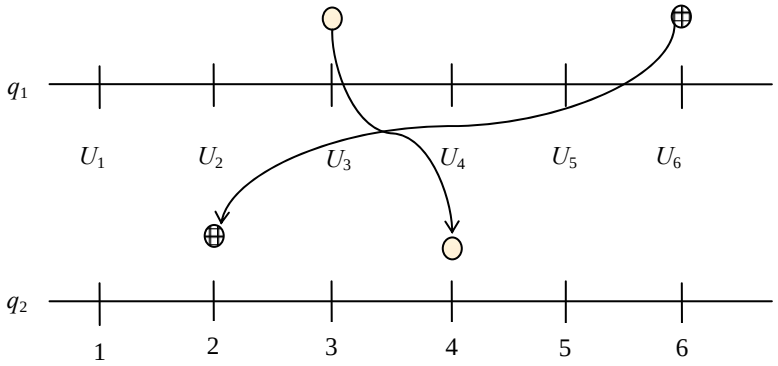

그림. 두 질문의 순위화비교

원천질문 q_1의 쎄션은 3번과 6번에 나타나는 문서선택모임 U_3, U_6을 포함한다. 이 문서모임의 순위는 6이다. 마찬가지로 같은 문서모임에 대하여 질문 q_2는 순위 4를 차지하며 적합한 추천대상으로 된다.

문서모임의 순위화를 비교하는 기준을 정의하였으므로 순위모임을 비교하는 방법으로 질문들을 추천할수 있다.

질문 q가 주어졌을 때

$$SA_q = \{S_q^1, S_q^2, \cdots, S_q^n\}$$

을 질문리력에서 질문 q에 대응하는 질문쎄션모임이라고 하자. 여기서 S_q^i는 질문 q에 대한 질문쎄션$(1 \le i \le n)$이다.

$D(S_q^i)$와 일치하는 질문렬을 $\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_m$이라고 할 때

$$f(\hat{q}_i) > \left|SA_q\right| \times C, \exists j \in \{1, 2, \cdots, n\}, r(q, D(S_i^j)) < r(\hat{q}_i, D(S_q^j))$$

인 $\hat{q}_i$들을 질문 q에 대응질문으로 추천하면 사용자의 의도를 반영하면서 문서순위화를 개선할수 있다. 여기서 $f(\hat{q}_i)$은 질문렬 $\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_m$에서 $\hat{q}_i$의 빈도수를 나타낸다.

이 방법으로 질문을 추천하는데서 문제는 질문리력에 존재하는 모든 질문들가운데서 사용자가 제출한 질문 q의 매 질문쎄션들에서 얻어지는 문서모임 $D(S_q^j)$와 일치하는 질문렬을 찾는것이다. 이 과정은 실시간적으로 진행되여야 하며 질문리력자료가 방대한것으

로 하여 이 방법을 그대로 적용할수 없다. 방도는 질문들을 미리 무리짓기하여 그 무리에 속하는 질문들가운데서 문서모임 $D(S_q^j)$와 일치하는 질문렬을 찾으면 실시간성을 보장할 수 있다. 질문을 무리짓자면 질문들을 어떤 공간으로 넘기고 그 공간의 벡토르들사이의 류사성을 정의하여야 한다.

일반적으로 질문공간은 매우 성글며 사용자가 검색을 진행할 때에는 자기의 의도에 맞는 문서들을 찾을 때까지 련속적으로 질문을 제출하며 그 질문용어들과 그때 얻어지는 응답목록은 사용자의 의도를 반영한것으로 볼수 있다. 그러므로 질문용어를 대응하는 질문쎄션공간에서 선택된 문서들에 있는 용어모임을 리용하여 벡토르로 표현하면 용어공간의 성김성애로를 극복하고 사용자의도를 반영한 질문무리짓기를 진행할수 있다.

질문쎄션은 다음과 같이 쓸수 있다.

QueySession:=(query, (clicked특정자원지적자)*)

질문쎄션은 질문과 그 응답결과에서 선택된 특정자원지적자들의 모임으로 구성된다.

질문 q와 특정자원지적자 u에 대하여 $Pop(q, u)$를 질문 q의 결과에서 u의 일반화값이라고 하고 $Tf(t, u)$를 특정자원지적자 u에서 용어 t의 출현회수라고 하자.

이때 질문 q에 대한 벡토르표현을 용어의 빈도수와 거꿀문서빈도수를 고려한 무게화 방법에 따라 다음과 같이 진행한다.

$$\vec{q}[i]=\sum_{URL_u}\frac{Pop(q,u)\times Tf(t_i,u)}{\max_t Tf(t,u)}$$

여기서 $\vec{q}[i]$는 벡토르의 i번째 요소이며 어휘의 i번째 용어와 련관된다.

직관적으로 질문용어벡토르의 매 구성요소는 질문표시의 용어의 적합성을 표현한다는것을 알수 있다.

이것은 다음과 같이 계산된다.

매 문서에 대하여 용어의 절대빈도수(문서에서 용어의 출현회수)와 문서의 일반화값과의 적을 계산한다. $Pop(q, u)$가 질문 q의 결과목록에서 특정자원지적자 u를 선택한 몫으로 정의되므로 이 값은 [0, 1]구간에 놓인다.

그러므로 매 요소를 표준화하기 위하여 출현회수들을 최대출현회수 $\max Tf(t, u)$로 나눈다. 매 문서용어무게를 일반화값으로 표준화함으로써 선택된 모든 문서들에 대한 총합은 [0, 1]구간에 놓인다. 따라서 질문벡토르표현은 구간 $[0, 1]^n$에서의 n차원문서용어공간을 생성한다. 여기서 n은 어휘의 크기이다.

다음 무리짓기를 위한 류사성기준과 무리짓기알고리듬을 선택한다. 질문리력화일은 용량이 대단히 크므로 무리짓기알고리듬의 고속성을 보장하여야 한다.

그러므로 두 벡토르사이의 류사성척도로는 코시누스류사도를, 무리짓기알고리듬으로는 k-평균법을 리용하였다.

2. 실험결과 및 분석

검색체계의 성능은 주로 적중률(precision)과 완전률(recall)을 리용하여 평가한다.

$$P(\text{적중률})=S_c/S_s$$

$$R(\text{완전률}) = S_c / S_a$$

여기서 S_c는 체계가 정확히 찾은 문서수, S_s는 체계가 찾은 문서의 수, S_a는 찾아야 할 전체 문서수이다.

F-척도는 다음과 같은 식으로 구한다.

$$F = 2 \times \frac{\text{적중률} \times \text{완전률}}{\text{적중률} + \text{완전률}}$$

문서자료기지로 1 242개의 질문과 35개의 특정자원식별자를 포함하는 842개의 질문쎄션을 준비하였다.

그리고 질문련관방법과 질문무리짓기방법을 리용한 질문추천체계의 성능을 이전의 체계와 비교하여 평가하였다.(표 1)

표 1. 질문추천체계의 성능평가

검색방법	적중률/%	완전률/%
종전방법[2]	83.2	78.9
제안방법	87.5	84.7

표 1에서 알수 있는바와 같이 우의 방법으로 검색을 진행하면 적중률은 평균 4%, 완전률은 6%정도 높일수 있다.

다음으로 체계성능평가에서 중요한것은 대답목록에 속하는 문서들에 대한 순서이다. 검색결과가 서로 같다고 하여도 순위화가 사용자의 의도를 반영하여 진행되였으면 더 좋은 검색결과라고 할수 있다.

이것을 평가하기 위하여 매개 질문에 대하여 10개의 특정자원식별자들을 선택하여 론문에서 정의한 순서화기준에 따라 평가하였다.

표 2. 질문추천체계들의 평균질문순위

검색방법	평균질문순위
종전방법[1]	7.34
제안방법	4.21

표 2에 질문모임에 속하는 모든 질문들에 대한 추천체계들의 평균질문순위를 보여주었다.

표 2에서 알수 있는바와 같이 현재의 질문추천체계를 리용하면 종전보다 평균 3만큼 질문순위를 높일수 있다.

맺 는 말

검색질문리력발굴을 통하여 사용자의 의도와 질문들사이의 의미적련관관계를 반영하여 질문을 추천하는 방법을 제기하고 실현하였다.

참 고 문 헌

[1] M. Sahami , T. Heilman; In Proceedings of the 15th International Conference on World Wide Web, 377, 2006.

[2] Y. Dhingra et al.; International Research Journal of Engineering and Technology, 4, 7, 1541, 2017.

주체108(2019)년 2월 5일 원고접수

A Method of Recommending Query Based on Query Log Mining

U Jun Min, Kim Thae Gyun

In this paper we proposed and implemented a method of finding and recommending surrogate query by query log mining in web.

Key words: query log mining, query recomposition