http://dx.doi.org/10.3745/KIPSTB.2012.19B.3.189

잠정적 부적합 문서와 어휘 근접도를 반영한 어휘 그래프 기반 질의 확장

조 승 현[†] · 이 경 순[††]

요 약

본 논문에서는 정보검색 성능 향상을 위해 잠정적 적합 문서 및 부적합 문서와 어휘 그래프를 이용한 질의 확장 방법을 제안한다. 언어모델에 의한 초기 검색 결과 상위 문서들은 질의 어휘 조합과 근접도를 기반으로 핵심 질의를 포함하는 문서들로 구성된 핵심 질의 클러스터와 핵심 질의를 포함하지 않는 문서들로 구성된 비핵심 질의 클러스터로 분류된다. 이때, 핵심 질의 클러스터는 잠정적 적합 문서 집합으로, 비핵심 질의 클러스터는 잠정적 부적합 문서 집합으로 본다. 각 클러스터는 어휘들과 질의 어휘와의 가까운 정도에 따라 어휘 그래프로 표현된다. 각 어휘에 대한 중요도는 핵심 질의 클러스터 그래프에서의 어휘 가중치에서 비핵심 질의 클러스터 그래프에서의 어휘의 가중치를 빼서 계산한다. 이는 부적합 문서에서 높은 가중치를 갖는 어휘는 확장 질의에서 제외시키는 역할을 한다. 중요도가 높은 어휘 순으로 확장할 질의를 선택한다. 웹 문서 테스트컬렉션인 TREC WT10g에서의 실험 결과에서 제안 방법이 언어모델(LM)에 비해 평균 정확률의 평균(MAP)에서 9.4% 성능 향상을 보였다.

키워드 : 질의 확장, 어휘 근접도, 잠정적 적합 문서, 잠정적 부적합 문서, 어휘 그래프

Query Expansion Based on Word Graphs Using Pseudo Non-Relevant Documents and Term Proximity

Seung-Hyeon Jo[†] · Kyung-Soon Lee[††]

ABSTRACT

In this paper, we propose a query expansion method based on word graphs using pseudo-relevant and pseudo non-relevant documents to achieve performance improvement in information retrieval. The initially retrieved documents are classified into a core cluster when a document includes core query terms extracted by query term combinations and the degree of query term proximity. Otherwise, documents are classified into a non-core cluster. The documents that belong to a core query cluster can be seen as pseudo-relevant documents, and the documents that belong to a non-core cluster can be seen as pseudo non-relevant documents. Each cluster is represented as a graph which has nodes and edges. Each node represents a term and each edge represents proximity between the term and a query term. The term weight is calculated by subtracting the term weight in the non-core cluster graph from the term weight in the core cluster graph. It means that a term with a high weight in a non-core cluster graph should not be considered as an expanded term. Expansion terms are selected according to the term weights. Experimental results on TREC WT10g test collection show that the proposed method achieves 9.4% improvement over the language model in mean average precision.

Keywords : Query Expansion, Term Proximity, Pseudo Relevant Documents, Pseudo Non-relevant Documents, Word Graph

1. 서 론

정보검색 연구에서 질의 확장은 검색 결과의 정확률과 재현률을 모두 향상시킬 수 있는 방법으로서, 벡터공간모델에서의 Rocchio 알고리즘[1]과 언어모델에서의 적합 모델[2] 등의 연구가 기반이 되었고, 초기검색문서 재샘플링을 통한 적합모델 적용 등 잠정적 적합성 피드백 방법[3] 등 많은 연구가 되어오고 있다. 대부분의 연구에서는 질의 확장을 할 때, 적합 문서만을 이용하여 확장할 어휘를 선택하고 있는데[2, 3], 질의와 연관되어 있지 않은 어휘가 확장 어휘로 선택될 수 있게 된다. 이러한 문제점을 보완하기 위한 방법으로, 잠정적 부적합 문서 집합을 이용하여 부적합 문서에 자주 나타나는 어휘의 중요도를 낮춰준 후 질의 확장을 한 연구가 있다[1, 4].

† 준 회 원 : 전북대학교 컴퓨터공학부 학사과정
†† 정 회 원 : 전북대학교 컴퓨터공학부 영상정보신기술연구센터 부교수(교신저자)
논문접수 : 2012년 1월 11일
수 정 일 : 1차 2012년 4월 4일
심사완료 : 2012년 4월 9일

질의에서 핵심 어휘를 질의로 표현하는 방법에 따라 성능에 변화가 있음을 알고, 질의에서 핵심 개념[5, 6]을 추출하거나 질의에 나타난 어휘들의 조합을 이용하여 모든 부분 질의(sub-query)[7, 8]를 이용해서 질의의 핵심적인 의미는 간직한 채 간결하게 줄임으로써 성능 향상을 꾀하는 연구는 계속 되어왔다. 질의 어휘의 가중치를 정교하게 조정한 질의 확장 연구로는 질의 어휘와의 근접도(proximity)를 이용하여 피드백 문서 안에서 어휘의 위치를 코사인, 가우시안 등 함수의 그래프를 이용하여 적합모델에 적용시켜 성능 향상을 보인 연구[9, 10]가 있다. 또한, 질의 어휘와 가까이 자주 나타난 어휘들을 중요시 하기 위해, 어휘 그래프(word graph)[11, 12]로 표현하여 질의를 확장한 연구가 있다. 그래프 기반 텍스트랭크(TextRank) 알고리즘[13]은 페이지랭크 알고리즘[14]을 텍스트에 적용한 모델로, 한 문서에서 출현한 어휘나 문장 등을 그래프의 노드로 표현하고, 어휘들 사이의 근접도를 에지의 가중치로 표현한 후, 어휘의 중요도가 일정한 값으로 수렴될 때까지 반복적인 연산을 통해 어휘의 가중치 조정하는 방법이다.

본 연구에서는 부적합 문서 집합과 어휘 근접도를 그래프 모델에 적용하여 질의확장을 하기 위해 다음과 같이 접근한다. 첫째, 질의 조합을 기반으로 한 클러스터링에서 질의 어휘 사이의 근접도를 이용하여 핵심 질의를 추출하고, 핵심 질의가 포함된 핵심 질의 클러스터(잠정적 적합 문서 집합)와 비핵심 질의 클러스터(잠정적 부적합 문서 집합)로 나눈다. 둘째, 각 클러스터에 대해 질의 어휘와의 근접도를 어휘 그래프로 표현한다. 셋째, 핵심 질의 클러스터의 어휘 그래프의 가중치를 비핵심 질의 클러스터의 어휘 그래프의 가중치를 이용하여 재조정한다. 마지막으로, 재조정된 그래프에서 중요도가 높은 어휘를 질의 확장 어휘로 선택한다. 제안 방법의 유효성을 검증하기 위해 TREC WT10g 테스트컬렉션에 대해 실험하고, 질의 확장 방법인 잠정적 적합성 피드백 모델에서 우수한 성능을 보인 적합모델(RM; Relevance Model)[9]과 비교 평가한다.

2. 관련 연구

적합 문서를 이용한 질의 확장 연구로, 적합모델(Relevance Model)[15]은 언어모델에 의한 초기 검색 결과에서 상위 문서들은 잠정적으로 질의에 적합한 문서라 가정하고 각 단어에 대해 문서에서 단어가 나타난 확률과 문서의 초기 검색 결과를 곱한 것을 누적한 값으로 각 단어의 확률을 계산하고, 높은 확률을 갖는 단어들로 확장 단어를 선택하는 방법으로, 최근 질의 확장에 효율적인 방법으로 알려져 있다. 부적합 문서를 이용한 질의 확장 연구[4]에서는 SMART 시스템의 초기 검색 결과 상위 20개는 적합 문서로 가정하고, 검색 순위가 501-1000 사이의 문서는 부적합 문서로 가정하여, Rocchio 질의 확장 기법[1]을 이용하여 질의를 확장하였다.

어휘 그래프를 이용한 가중치 계산에 관한 연구로, 텍스트랭크 알고리즘[13]은 텍스트 안에 모든 어휘들 사이의 관계 정보를 이용해 반복적으로 연산한 결과를 고려하여 각 어휘의 중요도를 결정하는 랜덤워크 알고리즘(random-walk algorithm)[16]을 그래프 기반 순위화 알고리즘에 적용한 방법이다. 텍스트랭크를 이용해 키워드를 추출하는 연구에서는 어휘의 공기빈도(co-occurrence)를 가상의 연결고리로 사용하였다. 텍스트랭크 알고리즘은 수식 (1)과 같이 어휘의 중요도가 일정한 값으로 수렴될 때까지 반복적으로 연산을 수행한다.

$$WS(V_i) = (1-d) + d \cdot \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \quad (1)$$

여기서 V_i는 그래프 상의 임의의 어휘이고, V_j는 V_i와 인접한 어휘를 나타낸다. V_k는 V_j와 인접한 어휘다. $In(V_i)$는 V_i와 인접한 어휘의 집합, $Out(V_j)$는 V_j와 인접한 어휘의 집합이다. 초기 단계에서 $WS(V_j)$의 값은 1이다. 본 연구에서는 어휘 근접도를 그래프로 표현할 때, 각 어휘의 초기값을 언어모델에 의한 검색 결과 값으로 지정하였다.

3. 잠정적 부적합 문서 집합의 어휘 그래프를 반영한 질의 확장

3.1 핵심 질의 선택을 통한 잠정적 적합 및 부적합 문서 분류

적합 및 부적합 문서 분류를 위해 기존 연구 [4]에서 초기 검색 결과 1000개에서 상위 문서 20개를 적합 문서로 하고, 하위 문서 500개를 부적합 문서로 이용한 것과는 달리, 본 논문에서는 핵심개념을 포함하는 문서를 적합 문서로, 그렇지 않은 문서를 부적합 문서로 분류하였다.

질의에서 핵심 개념을 추출하는 방법은 [17]의 연구를 적용하였다. 질의의 핵심 개념을 표현하는데 두 개 또는 세 개의 어휘가 있다고 가정하고, 길이가 긴 질의에 대해서 두 개의 어휘로 된 핵심 질의를 추출한다. 원래 질의를 적용해서 검색한 초기 검색 결과에 대해서 핵심 질의를 포함하고 있는 클러스터를 핵심 클러스터로, 그렇지 않은 클러스터를 비핵심 클러스터로 분류하였다.

핵심 및 비핵심 질의 클러스터를 분류하기 위해서 우선, 초기 검색 결과의 문서에 대해서 질의 어휘 조합을 기반으로 문서를 클러스터링 한다. 이때, r개의 어휘로 표현된 질의에서는 최대 2^r-1개의 클러스터가 생길 수 있다. 예를 들어, 세 개의 어휘로 표현된 질의 (q_1, q_2, q_3)에서는 q_1 어휘만을 포함하는 문서들의 q_1 클러스터, q_2 클러스터, q_3 클러스터, q_1과 q_2 어휘를 둘 다 포함하는 q_1&q_2 클러스터, q_1&q_3 클러스터, q_2&q_3 클러스터, q_1과 q_2 그리고 q_3를 모두 포함하는 q_1&q_2&q_3 클러스터로 7개의 문서 클러스터가 생긴다.

질의 개념을 표현하는 핵심 질의를 찾는 방법은 초기 검색 결과 문서 집합에서 임의의 두 질의 어휘가 일정한 거리

(window size)안에 자주 발생하는지, 문서에서의 어휘 빈도수(tf)와 역문서 빈도수(idf)를 반영한 어휘 가중치(tf • idf)가 높은지를 고려하여 수식 (2)과 같이 계산하였다[17]. 각 문서에서 모든 어휘 조합 사이의 공기 빈도를 구하고, 질의 어휘가 두 개 이상 발생한 모든 클러스터의 문서들에서 더한다.

$$CoreQuery(q_i, q_j) = \sum_{q_i, q_j \in D_c \quad D \in S} cooc(q_i, q_j) \cdot (tfidf\ (q_i) + tfidf\ (q_j)) \tag{2}$$

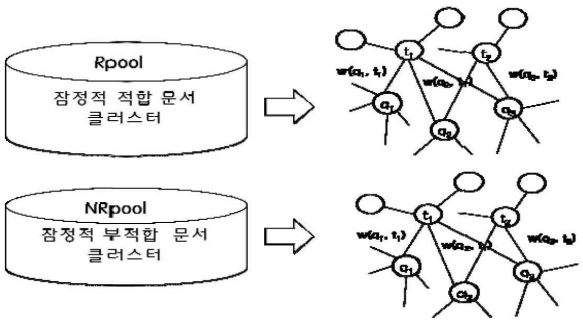
여기서 S는 초기 검색 결과 상위 n개의 문서 집합에서 질의 어휘가 두 개 이상 발생한 클러스터의 문서 집합이다. cooc(q_i, q_j)는 q_i, q_j의 문서에서 일정한 거리 (학습을 통해 가장 성능이 좋은 15로 설정했음) 안에서 발생한 공기 빈도이다. $tfidf\ (q_i)$는 문서 안에서 단어 q_i의 tf • idf값이다. 식을 통해 CoreQuery(q_i, q_j)가 가장 높은 한 쌍의 단어 조합이 핵심 질의로 선택된다.

위의 예에서, 핵심 질의 어휘로 (q_1, q_2)가 선택되었을 때, 문서에 q_1과 q_2를 포함하는 q_1&q_2클러스터와 q_1&q_2&q_3 클러스터 두 개가 핵심 클러스터로 선택된다. 나머지 5개의 클러스터는 비핵심 클러스터로 분류된다.

핵심 질의를 포함한 핵심 클러스터의 문서 집합을 잠정적 적합 문서 집합(pRpool)으로 분류하고, 그렇지 않은 비핵심 질의 클러스터에 속한 문서 집합을 잠정적 부적합 문서 집합(pNRpool)으로 분류한다.

3.2 잠정적 적합 및 부적합 문서 집합에 대한 어휘 그래프 표현

본 논문에서는 잠정적 적합 문서 집합에 나타난 어휘들을 하나의 그래프로 표현하고, 부적합 문서 집합에 나타난 어휘들을 또 다른 그래프로 표현한다. 즉, 두 개의 어휘 그래프를 생성한다.



(그림 1) 잠정적 적합 문서 및 부적합 문서 클러스터의 어휘 그래프 표현

어휘 그래프 생성을 위한 노드 및 에지의 가중치는 다음과 같이 계산한다[18]. 문서에 속하는 각 어휘에 대한 그래프 G = (V, E)로 표현할 수 있다. V는 그래프의 노드로 문서에서 각 어휘를 나타내고, E는 어휘가 질의 어휘와 가까이 발생한 정도(근접도)에 따라 높은 가중치를 부여하기 위해 가중치를 갖는 에지로 표현하였다(그림 1). 노드의 가중치는 수식 (3)로 계산한다[11].

$$f^{r+1}(t) = \alpha \times f^0(t) + (1-\alpha) \times \sum_{q_j \in Near(t)} \frac{w(t, q_j) \times f^r(q_j)}{\sum_{t \in Near(q_j)} w(t, q_j)} \tag{3}$$

여기서 $f^0(t)$는 노드 t의 초기 가중치로서 수식 (4)에서와 같이 적합모델(Relevance Model)을 이용하여 계산한다. Near(t)는 문서 안에서 t와 근접하게 나타난 어휘들의 집합이다. 에지의 가중치를 나타내는 w(t, q_j)는 t와 q_j사이의 근접도이다(수식 (5)에서 계산). 어휘의 가중치가 일정한 값에 수렴할 때까지 반복적으로 계산한다. 수렴하기 위한 임계치 (c = f^{r+1}(t) - f^r(t))는 0.000001로 한다.

$$f^o(t_i) = \sum_{D \in R} P(D) P(t_i \mid D) P(Q \mid D) \tag{4}$$

여기서 R은 질의 Q에 대해 잠정적으로 적합하다고 가정한 문서들의 집합이다. P(D)는 문서가 발생할 확률이므로 모든 값에 균일하게 적용된다. P(t|D)는 문서에서 어휘 t가 발생할 확률을 나타낸다. P(Q|D)는 초기 질의 Q에 대한 문서 D의 언어모델에 의한 확률 값이다.

질의 어휘와 근접한 어휘는 질의와 어떤 연관이 있다고 볼 수 있다. 질의 어휘와의 근접도를 에지의 가중치로 하여 단어 그래프에 적용한다. 수식 (5)는 에지의 가중치를 나타낸다.

$$w(t, q_j) = \sum_{t \in Near(q_j)} prox(t, q_j) \tag{5}$$

여기서 prox(t, q_j)는 문서 안에서 t와 q_j의 근접도이다. w(t, q_j)는 각 문서에서 구한 값을 잠정적 적합 문서 집합 전체에 대해서 더한 값이 된다.

$$prox(t, q_j) = 1 - \frac{dist(t, q_j)}{\delta} \tag{6}$$

여기서 dist(t, q_j)는 t와 q_j 사이의 단어 거리이다. δ는 거리 가중치 적용 파라미터이다. 즉, 질의 어휘 q_j와 어휘 t가 가까이 나타날수록 prox(t, q_j)가 높은 값을 갖게 된다. 또한 두 어휘가 자주 나타날수록 에지의 가중치인 w(t, q_j)가 높은 값을 갖게 된다.

잠정적 적합 문서 집합에 속하는 어휘들을 표현한 어휘 그래프에서 높은 가중치를 갖고 부적합 문서 집합에서 낮은 가중치를 갖는 어휘들은 질의 확장을 위한 어휘로 선택될 가능성이 높다.

3.3 어휘 그래프의 가중치 조정을 통한 질의 확장

잠정적 적합 문서에서 높은 가중치를 갖는 어휘는 질의

어휘와 높은 관련도가 있다고 볼 수 있지만, 부적합 문서에서 높은 가중치를 얻은 어휘는 검색에 도움을 주지 못할 것이다.

본 연구에서는 잠정적 적합 문서 집합(핵심 질의 클러스터에 속하는 문서들)을 어휘 그래프로 표현하여 얻은 어휘들의 가중치와 잠정적 부적합 문서 집합(비핵심 질의 클러스터에 속하는 문서들)을 어휘 그래프로 표현한다. 최종 어휘의 가중치는 수식 (7)를 이용하여 재조정한다. 어떤 어휘가 적합 문서 집합에도 많이 나타나고, 부적합 문서 집합에도 많이 나타나면 확장 어휘로서의 가치가 없다고 본다. 어휘의 가중치를 재조정하는 방법은 수식(7)과 같다.

$$score(t) = \alpha \cdot \frac{1}{|D_{pRpool}|} \cdot \sum_{t \in pRpool} f_R(t) - \beta \cdot \frac{1}{|D_{pNRpool}|} \cdot \sum_{t \in pNRpool} f_{NR}(t) \quad (7)$$

여기서 score(t)는 조정된 어휘의 가중치를 의미하며, $f_R(t)$와 $f_{NR}(t)$은 각각 핵심 질의 클러스터 그래프와 비핵심 질의 클러스터 그래프에서의 어휘의 가중치를 의미한다(수식 3에서의 $f^{r+1}(t)$의 최종값). $|D_{pRpool}|$은 핵심 질의 클러스터에 들어있는 문서의 수, $|D_{pNRpool}|$은 비핵심 질의 클러스터에 들어있는 문서의 수이다. 어휘의 가중치는 핵심 질의 클러스터에서 나온 어휘의 가중치에 α를 곱한 뒤, 비핵심 질의 클러스터에서 나온 어휘의 가중치에 β를 곱하여 뺀 값이 조정된 어휘의 가중치가 된다.

제안된 방법을 통해 조정된 어휘의 가중치가 높은 상위 e개의 단어를 확장 질의 W로 선택한다. 원래 질의와 확장 질의를 포함한 질의 Q′에 대한 문서의 중요도 P(Q′|D)는 다음과 같이 계산한다.

$$P(Q'|D) = \lambda \cdot P(Q|D) + (1-\lambda) \cdot P(W|D) \quad (8)$$

여기서 P(Q|D)는 원래 질의를 적용한 언어모델의 검색 결과 값이고, P(W|D)는 e개의 확장 어휘를 질의로 구성한 후 언어모델에 의한 검색 결과 값이다. 원래 질의와 확장 질의의 결과값을 가중치를 부여해서 결합해서 문서의 최종 결과 값으로 한 후 순위를 매긴다.

4. 실험 및 평가

4.1 실험 집합

제안 방법의 유효성을 검증하기 위해 웹 문서 테스트컬렉션인 TREC WT10g를 사용하여 실험하였다. 파라미터 학습을 위해 질의를 학습 질의와 테스트 질의로 구성하였다. 학습 질의를 이용해서 최적의 파라미터를 결정한 후, 테스트 질의에 적용을 하여 성능을 평가하였다.

TREC WT10g 테스트컬렉션에는 총 100개의 질의(학습 질의 50개, 테스트 질의 50개)로 구성되어 있는데, 본 논문에서 제안하는 방법은 질의가 2개 이하인 경우에는 핵심 질의 조합이 1개밖에 없어서 부적합 문서 클러스터가 생성되지 않는다. 따라서, 질의 어휘가 3개 이상으로 구성된 질의에 클러스터와 비핵심 질의 클러스터로 분류할 수 있으므로, 질의 어휘가 1-2개 인 것과 3개 이상인 것에 대해 따로 성능을 표시한다. 실험집합에 대한 구성은 <표 1>과 같다.

〈표 1〉 TREC WT10g 실험 집합

문서 개수	학습 질의 개수		테스트 질의 개수	
	1~2개 어휘	3개 이상 어휘	1~2개 어휘	3개 이상 어휘
1, 692, 096	29	21	21	29

언어모델(LM)과 적합모델(RM)에 대한 실험 결과는 인드리(Indri-2.8) 시스템[19]을 사용하였다. 각 모델에 대해 학습 질의를 이용하여 파라미터를 학습한 후 테스트 질의에 대해 적용하여 성능을 평가하였다. 파라미터는 수식(3)에서 단어의 초기 가중치(α∈{0.1, 0.15, 0.2, … , 0.9, 0.95}), 수식(7)에서 파라미터 값(α∈{0.1, 0.15, 0.2, … , 0.9, 0.95}, β ∈{0.1, 0.15, 0.2, … , 0.9, 0.95})는 실험에서 가장 좋은 성능을 보인 값으로 선택했다. 피드백 문서의 개수(n∈{5, 10, 25, 50, 75, 100}), 확장 어휘의 개수(e∈{5, 10, 20, 50, 75, 100}), 초기 질의에 대한 가중치(λ∈ {0.1, 0.2, … , 0.9})로 실험하였다.

4.2 비교 실험 결과

제안된 방법과 질의확장의 한 방법인 적합모델을 비교하여 성능을 평가하였다. 언어모델은 원래 질의에 대한 성능을 나타낸다. 성능평가의 척도는 평균 정확률(Average Precision)의 평균인 MAP(Mean Average Precision)이다.

- 언어모델(LM): 질의가 특정 문서에서 발생할 확률을 계산하여 그 확률이 가장 큰 문서를 적합한 문서로 하고 상위에 순위화 한다.
- 적합모델(RM): 언어모델에 의한 초기 검색 결과에서 상위 문서들을 이용해서 확장할 질의를 선택한다.
- pRpool을 이용한 질의확장: 잠정적 적합 문서 집합에 대한 단어 그래프를 이용해 질의와의 어휘 근접도를 계산하고 확장 질의를 선택한다.
- 제안 방법: pRpool과 pNRpool에 해당하는 문서들에 대해 질의 어휘와의 근접도를 각 어휘 그래프로 표현하고, pNRpool에 대한 어휘 그래프의 값을 빼서 확장 질의를 선택한다.

학습 질의에 대하여 실험한 결과, α = 0.95, β = 0.1, 확장 어휘의 개수(e)는 5개와 10개일 때가 가장 좋았으며, 이 파라미터 값을 이용하여 테스트 질의에 대하여 실험을 하였다.

<표 2>에서와 같이 적합모델은 언어모델보다 5.67%의 성능이 향상되었다(즉, 시스템 A의 성능에 비해 시스템 B의

성능 향상률은 (B-A)/A*100으로 계산함). pRpool만을 이용한 방법은 언어모델에 비해 5.62% 향상을 보였다. pRpool에 pNRpool을 반영한 제안 방법이 언어모델보다 9.41%의 성능 향상을 보였다. 이러한 결과를 통해, 잠정적 부적합 문서 집합에 자주 나타나는 어휘의 중요도를 감소시키는 질의 확장 방법이 적절한 방법임을 확인할 수 있다.

〈표 2〉 어휘가 3개 이상으로 구성된 질의에 대한 실험 결과

	LM (언어모델)	RM (적합모델)	잠정적 적합 문서 집합 이용(pRpool)	제안 방법 (pRpool & pNRpool)
MAP	0.2028	0.2143	0.2142	0.2219
향상률	-	+5.67%	+5.62%	+9.41%

〈표 3〉 어휘가 2개 이하로 구성된 질의에 대해 적합 문서 어휘 그래프를 적용한 실험 결과

	LM (언어모델)	RM (적합모델)	pRpool 이용
MAP	0.2411	0.2451	0.2507

2개 이하의 어휘로 구성된 질의에 대한 실험 결과는 <표 3>과 같다. 2개 이하로 구성된 학습 질의를 이용해 결정한 최적의 파라미터를 적용하여 테스트 질의에 대해 성능 평가한 결과이다. 부적합 문서가 없기 때문에 적합 문서 클러스터에 대한 어휘 그래프에서의 질의 확장 결과가 유효함을 알 수 있다.

4.3 결과 분석

제안 방법을 이용해서 성능이 향상된 질의 한 예로, Q537 “sun beds safe”에서 핵심 질의 선택 기법에 의해 선택된 질의는 “beds safe”이다. 초기 질의에 대한 검색 결과에서 잠정적 적합 문서 클러스터에 속하는 문서는 20개이고, 부적합 문서 클러스터에 속한 문서의 개수는 5개이다. 초기언어모델 (LM) 검색 결과 평균정확률은 0.0369, 상위적합 문서를 이용한 적합모델(RM)에 의한 평균정확률은 0.0143, 잠정적 적합 클러스터만을 적용한 방법(pRpool)에 의한 평균정확률은 0.0261이였는데, 제안 방법을 적용한 경우 평균정확률은 0.1978로 부적합 문서를 이용한 질의확장 가중치 조정이 유효함을 보여주었다.

성능 향상이 되지 않은 질의에 대한 오류분석에서는 핵심 질의 선택 방법에서 오류가 발생 했을 때, 적합 문서 클러스터와 부적합 문서 클러스터 구성 자체가 잘못된 질의가 3개 있었다. 이는 핵심 질의를 선택하는 방법을 개선함으로써 성능 향상을 꾀할 수 있을 것이다.

5. 결 론

본 논문에서는 질의 조합에 기반한 질의 클러스터에서 핵심 질의를 추출하였고, 핵심 질의를 포함하는지에 따라 핵심 질의 클러스터와 비핵심 질의 클러스터로 분류하였다. 각 클러스터에 대해서 질의 어휘와의 근접도를 가중치로 반영하여 어휘 그래프로 표현하고, 핵심 질의 클러스터의 어휘 그래프의 값에서 비핵심 질의 클러스터의 어휘 그래프의 값을 빼서 가중치를 재조정해 질의를 확장하는 기법에 대해 제안하였다. 이는 잠정적 적합 문서 집합에서 질의 어휘와 가까이 자주 나타난 어휘는 확장 어휘로서 중요도를 부여하고, 부적합 문서 집합에 질의 어휘와 가까이 많이 나타난 어휘는 확장 어휘로서 선택하지 않도록 하는 방법이다. 실험을 통해, 제안 방법이 언어모델보다 9.4%가 향상됨을 보였다. 이것을 통해 질의를 확장할 때, 부적합 문서를 이용하여 질의와 연관되어 있지 않은 어휘의 가중치를 줄여주면 질의 확장을 할 때 도움을 줄 수 있음을 확인할 수 있었다.

제안 연구의 한계는 핵심 질의 추출에서 오류가 생길 경우 성능 저하에 영향을 미치게 되므로, 향후 연구로는 핵심 질의 추출의 정확도를 높이는 연구와 하나의 어휘로 된 질의에 대해서도 어휘 그래프를 이용한 질의 확장에 관한 연구가 필요하다.

참 고 문 헌

[1] J. J. Rocchio, “Relevance feedback in information retrieval. In The SMART Retrieval System - Experiments in Automatic Document Processing”, Prentice Hall. pp.313-323, 1971.

[2] V. Lavrenko and W.B. Croft, “Relevance-based Language Models”, In Proc. of 24th ACM SIGIR Conference(SIGIR2001). pp.120-127, 2001.

[3] K.-S. Lee, W.B. Croft, and J. Allan, “A Cluster-Based Resampling Method for Pseudo-Relevance Feedback”, In Proc. of 31st ACM SIGIR Conference(SIGIR2008), pp.235-242, 2008.

[4] C. Buckley, M. Mitra, J. Walz, and C. Cardie, “Using Clustering and SuperConcepts within SMART: TREC 6”, In Proc. of the Sixth Text REtrieval Conference(TREC-6), pp.500-240, 1995.

[5] M. Bendersky and W.B. Croft, “Discovering Key Concepts in Verbose Queries”, In Proc 31th ACM SIGIR Conference (SIGIR2008), pp.491-498, 2008.

[6] A. Hulth, “Improved automatic keyword extraction given more linguistic knowledge”, In Proc. Empirical Mothods in Natural Language Processing(EMNLP2003), pp.216-223, 2003.

[7] G. Kumaran and J. Allan, “Effective and Efficient User Interaction for Long Queries”, In Proc 31th ACM SIGIR Conference(SIGIR2008), pp.11-18, 2008.

[8] G. Kumaran, and J. Allan, “A case for shorter queries and helping users create them”, In Proc. HLT-EMNLP Conference. pp.220-227, 2007.

[9] Y. Lv and C.X. Zhai, “Positional Language Model for Information Retrieval”, In Proc. of 32nd ACM SIGIR Conference (SIGIR2009). pp.299-306, 2009.

[10] Y. Lv and C.X. Zhai, “Positional Relevance Model for Pseudo-Relevance Feedback”, In Proc. of 33rd ACM SIGIR Conference (SIGIR2010), pp.579-586, 2010.
[11] Q. Mei, D. Zhang, and C.X. Zhai, “A General Optimization Framework for Smoothing Language Models on Graph Structures”, In Proc. of 31st ACM SIGIR Conference (SIGIR2008), pp.611-618, 2008.
[12] Y. Huang, L. Sun, and J.Y. Nie, “Smoothing Document Language Model with Local Word Graph”, In Proc. of 18th ACM Conference on Information and Knowledge Management (CIKM2009), pp.1943-1946, 2009.
[13] R. Mihalcea, and P. Tarau, “TextRank-Bringing Order into Texts”, In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp.404 - 411, 2004.
[14] L. Page, S. Brin, R. Motowani, and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web”, Unpublished manuscript, Stanford University. 1998.
[15] V. Lavrenko, and W.B. Croft, “Relevance-based language models”, In Proc. of 24th ACM SIGIR Conference (SIGIR2001), pp.120-127, 2001.
[16] S. Hassan, and C. Banea, “Random-Walk Term Weighting for Improved Text Classification”, In Proc. of TextGraphs: 2nd Workshop on Graph Based Methods for Natural Language Processing. pp.53-60, 2006.
[17] 장계훈, 이경순. “핵심 질의 클러스터와 단어 근접도를 이용한 문서 검색 정확률 향상 기법”, 정보처리학회논문지B 제 17권 제 5호, pp.399-404, 2010.
[18] 장계훈, 조승현, 이경순. “단어 근접도를 반영한 단어 그래프 기반 질의 확장”, 제34회 한국정보처리학회 추계학술발표대회, 2010.
[19] T. Strohman, D. Metzler, H. Turtle, and W.B. Croft, “Indri: A language model-based search engine for complex queries”, In Proc. International Conference on Intelligence Analysis. http://www.lemurproject.org. 2005.

조 승 현
e-mail : jackaa@chonbuk.ac.kr
2012년 전북대학교 컴퓨터공학부 학사과정
관심분야 : 정보검색, 정보마이닝

이 경 순
e-mail : selfsolee@jbnu.ac.kr
1994년 계명대학교 컴퓨터공학과(학사)
1997년 한국과학기술원 전자전산학(석사)
2001년 한국과학기술원 전자전산학(박사)
2001년～2003년 일본 국립정보학연구소 (National Institute of Informatics) 연구원
2007년 미국 매사추세츠주립대학 방문교수
2004년～현 재 전북대학교 컴퓨터공학부 영상정보신기술 연구센터 부교수
관심분야 : 정보검색, 정보마이닝, 자연언어처리