핵심 질의 어휘와 근접도를 이용한 핵심 문서 선택 기법

장계훈, 김설영, 이경순
전북대학교 컴퓨터공학과
e-mail : {ghjang, xying, selfsolee}@chonbuk.ac.kr

Core Document Selecting Method Using Core Query Term and Proximity

Gye-Hun Jang, Seol-Young Kim, Kyung-Soon Lee
Dept. of Computer Engineering, Chon-buk National University

요 약

길이가 긴 질의에는 검색에 불필요한 단어가 포함되어 있어서 사용자가 원하지 않는 문서가 검색결과에 포함된다. 질의에서 불필요한 단어를 제거하고 핵심 단어만 선택한다면 검색에 효율을 높일 수 있다. 본 논문에서는 질의 조합을 기반한 클러스터를 이용해 핵심 단어를 찾고 핵심 단어와의 근접도를 이용해 상위 문서의 정확율을 향상시키는 기법을 제안한다. 실험은 뉴스 집합인 TREC AP 문서를 검색한 결과를 제안한 알고리즘으로 재순위화하여 초기 검색 결과의 상위 문서의 정확율을 비교함으로써 제안된 알고리즘의 유효성을 검증한다.

1. 서론

웹에서 사용자가 원하는 정보를 검색하기 위해 사용하는 질의 어휘는 대부분 의미가 모호하다. 특히 길이가 긴 질의는 불필요한 단어가 포함되어서 사용자가 원하는 정보와 더 멀어진다. 사용자들은 긴 질의를 사용함으로써 더 정확한 의미를 전달했다고 믿지만 2~3 개의 어휘들을 제외한 대부분의 어휘들은 관련 없는 어휘들이다. 핵심 어휘인 2~3 개의 단어만 원하는 정보를 찾는데 도움을 준다.

Collins-Thompson 의 연구[1]에서는 질의 어휘들 중에서 하나는 불필요한 어휘일 것이라 가정하고 질의 어휘 중 하나의 어휘를 제거해서 만든 질의 변이(query variant)에 의한 검색결과를 샘플링하여 피드백하는 방법을 제안하였다. 질의에서 핵심 개념[2, 3]을 찾거나 질의에서 발생하는 모든 부분질의(sub-query)[4, 5]를 이용해서 질의의 핵심적인 의미는 간직한채 간결하게 줄이려는 연구는 계속 되어왔다.

본 연구에서는 다음을 가정한다. (ⅰ) 길이가 긴 질의에는 2~3 개의 어휘가 핵심 개념을 나타낸다. (ⅱ) 길이가 긴 질의 어휘 중에서 두 단어가 일정한 거리 안에 발생 빈도가 높으면 두 단어는 핵심 질의 어휘이다. (ⅲ) 핵심 질의 주변에 발생한 단어는 핵심 질의와 의미적으로 연관성이 있다.

상위 문서의 정확율 향상을 위한 첫 번째 단계는 질의 조합을 기반한 클러스터에서 질의 어휘 사이의 근접도를 이용해 핵심 클러스터를 선택하여 초기 검색 결과에서 부적합 문서를 필터링한다.

두 번째 단계는 핵심 클러스터 안에 문서들을 핵심 질의와 주변단어와의 관계를 통해 선택한 문맥어휘(Context term)를 이용해 문서의 가중치를 결정함으로써 적합문서를 상위에 순위화 하여 정확율을 높일 수 있다.

실험은 뉴스기사 집합인 TREC AP 문서를 인드리(indri) 검색엔진[6]을 통해 초기 검색 결과를 얻고 실험을 통해 재순위화된 결과와 비교하여 그 유효성을 검증한다.

2. 관련연구

Sakai 의 연구[7]에서는 질의 어휘 클러스터를 통해 문서를 선택하여 잠정적 적합 피드백에 사용한다. 초기 검색 결과 상위에 순위화된 문서들은 비슷한 행태를 가지고 있다고 가정하고 질의 어휘 기반 클러스터에서 선택적으로 클러스터를 선택함으로써 효율을 높이는 알고리즘을 제안 하였다.

본 연구에서는 상위 n 개의 문서를 질의 어휘 기반으로 클러스터링하여 그 중에 핵심 클러스터만을 선택하여 정확율을 향상시킨다.

Bendersky 의 연구[2]에서는 길이가 긴 질의(verbose queries)에 대해서 질의-의존, 코퍼스-의존, 그리고 코퍼스-독립적인 자질을 이용해서 핵심 개념(key concepts)을 선택하는 알고리즘을 제안 했다. Kumaran 의 연구[4] 에서는 역시 길이가 긴 질의에서 부분질의(sub-query)를 찾아내는 알고리즘을 제안했다. 사용자와의 상호작용(user interaction)을 최소화하면서 질의에서 발생할 수 있는 모든 부분질의를 고려하여 50%

이상의 평균정확률을 향상시켰다.

본 연구에서는 문서 안에서 발생한 모든 질의 어휘 조합 사이의 공기빈도를 구하고 가장 빈도가 높은 한 쌍의 단어 조합을 핵심 질의 어휘로 선택한다.

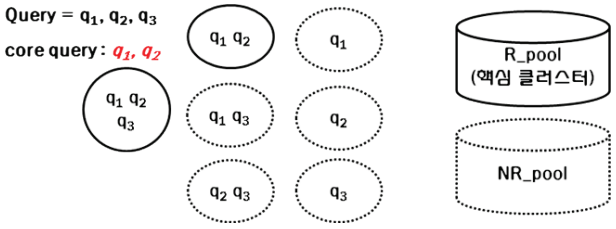
Yuanhua 의 연구[8]에서는 질의 사이의 거리를 그래프로 표현한 언어모델 검색 방법을 제안하였다. 질의 사이의 어휘 근접도가 클수록 사용자가 원하는 정보에 가까울 것이라 가정하고 그래프로 표현된 가중치를 이용한 모델링을 통해 성능이 향상됨을 보였다. 질의 사이의 근접도를 이용한 연구[9]에서는 어휘 사이의 최소거리, 평균, 어휘 사이 거리의 합 등을 이용하여 가중치를 결정한 모델링을 통해 성능이 향상됨을 보였다.

본 연구에서는 문서에서 핵심 질의 주변에 발생한 어휘들의 빈도를 기반으로 문서의 가중치를 결정한다.

3. 핵심 클러스터 선택 기법

핵심 클러스터는 긴 질의에서 핵심 질의를 찾고 초기 검색 결과에서 부적합한 문서를 필터링 하기 위해 사용된다. 초기 검색 결과는 상위 n 개의 문서를 사용한다. (실험에서는 300 개를 사용하였다.)

길이가 긴 질의에는 2~3 개의 핵심 어휘가 있다. 핵심 어휘를 찾기 위해 초기 검색결과의 문서에서 발생한 질의 어휘 조합을 기반으로 문서를 클러스터링한다. r 개의 질의 어휘를 가진 질의는 최대 2^r-1 개의 클러스터가 발생할 수 있다.



(그림 1) 핵심 클러스터 선택 방법

(그림 1)에서는 질의 어휘가 3 개인 질의에서 발생할 수 있는 모든 클러스터를 보여준다. 3 개의 질의 어휘 q_1, q_2, q_3 중 q_1, q_2를 핵심 질의로 선택했을 때 q_1, q_2를 포함하는 두 개의 클러스터가 핵심 클러스터로 선택된다. 핵심 클러스터는 적합문서 집합(R_Pool)에 나머지 클러스터는 부적합 문서 집합(NR_Pool)에 들어간다.

<표 1>은 실제 질의에서 300 개의 문서에 대한 질의 어휘 클러스터를 보여준다. “fiber optics applications” 3 개의 질의 어휘를 가진 질의에는 총 2^3-1 개 즉, 7 개의 클러스터가 발생할 수 있다. 이 실험 집합에서 총 7 개의 클러스터의 총 적합문서의 수는 40 개다. 만약 “fiber optics”를 핵심 질의 어휘로 선택 한다면 “fiber optics”를 포함한 C1, C2 두 개의 클러스터가 R_Pool 에 들어가게 된다. 그렇게 되면 R_Pool 에는 총 177 개의 문서가 들어가게 되고 40 개의 적합문서가 모두 들어가게 된다. 초기 검색 결과에서 상위 300 개의 문서의 정확율은 0.1333(40/300)이 되지만 핵심 질의 어휘를 통해 찾은 R_Pool 의 정확율은 0.2260(40/177)이 된다.

<표 1> 질의 "fiber optics applications"의 검색결과 300 개 문서에 대한 질의 어휘 클러스터

	질의 조합 클러스터	검색된 문서 수	적합 문서 수
C1	fiber optics applications	19	4
C2	fiber optics	158	36
C3	optics applications	26	0
C4	fiber applications	21	0
C5	optics	39	0
C6	fiber	37	0
C7	applications	0	0

이와 같이 “fiber optics”는 질의에서 핵심 개념이라 할 수 있으나 “applications”는 불필요한 단어라고 할 수 있다. 또 “fiber”와 “optics”는 각각의 어휘만으로는 의미를 전달하기는 어렵다는 것을 알 수 있다.

<표 2>는 학습 질의 전체에 대해 사람이 각각의 질의에서 직접 핵심 질의를 선택한 R_Pool 과 NR_Pool 의 포함율과 누락율을 보여준다.

- 포함율(recall) $= \frac{R\_Pool\text{의적합문서수(R\_Pool)}}{\text{전체적합문서수(Tot\_rel)}}$
- 누락율(miss alam) $= \frac{NR\_Pool\text{의적합문서수}(NR\_rel)}{\text{전체적합문서수}(Tot\_rel)}$
- Tot_rel 은 모든 적합문서의 수, R_rel 은 R_Pool 에서 적합문서의 개수, R_doc 는 R_Pool 에 있는 모든 문서의 개수, NR_rel 은 NR_Pool 에서 적합문서의 개수, NR_doc 는 NR_Pool 에 있는 모든 문서의 개수, Q#n 은 n 개의 질의 어휘를 가진 질의를 말한다.

<표 2>에서와 같이 질의 어휘가 6 개 이상인 질의를 제외하고 모든 질의에서 90%이상의 포함율을 보

<표 2> 학습 질의 전체에 대해 사람이 직접 판별한 R_Pool 과 NR_Pool 의 포함율과 누락율

	총 적합 문서 수 (Tot_rel)	R_Pool			NR_Pool		
		적합문서 수 (R_rel)	문서 수 (R_doc)	포함율 (recall)	적합문서 수 (NR_rel)	문서 수 (NR_doc)	누락율 (miss_alarm)
Q#3	1653	1527	4924	0.9238	126	2876	0.0762
Q#4	1073	1002	5040	0.9338	72	3060	0.0671
Q#5	777	720	2254	0.9266	57	1346	0.0734
Q#6 이상	281	192	1166	0.6833	89	1234	0.3167
총 문서 수	3784	3441	13384	0.9094	344	8516	0.0909

이고 있다. 이는 질의어휘기반 클러스터를 이용해 핵심 클러스터를 찾는 방법이 유효함을 알 수 있다.

핵심 질의 어휘를 포함하고 있는 모든 클러스터를 핵심 클러스터라고 정의한다. 핵심 클러스터를 찾기 위해 먼저 질의에서 핵심 질의 어휘를 찾아야 한다. 사용자가 검색을 위해 사용한 질의 어휘들 사이에는 의미적인 관계를 갖고 있다. 임의의 두 질의 어휘가 일정한 거리(window size)안에 자주 발생하면 두 단어는 서로 의미적인 연관도가 높고 질의 안에서 핵심 질의 어휘라고 생각할 수 있다.

공기 빈도란 한 문서에서 두 개의 단어가 일정한 거리 안에서 연속으로 발생한 빈도수를 말한다. 공기 빈도는 질의 어휘가 2 개 이상인 모든 클러스터에서 계산한다. 또한 클러스터 안에 모든 질의 어휘 조합을 고려한다. 예를 들어, q_1, q_2, q_3 세 개의 질의 어휘를 포함한 클러스터는 (q_1, q_2), (q_1, q_3), (q_2, q_3) 세 가지 질의 어휘 조합의 공기빈도를 구한다.

각각의 문서에서 모든 어휘 조합 사이에 공기 빈도를 구하고 각각의 문서에서 구한 어휘 조합의 공기빈도를 모든 문서에서 더한다. 또 문서에서 질의 어휘의 발생 빈도가 높으면 핵심 질의일 확률이 높다. 하지만 전체 문서 집합에서 빈도가 높다면 단어의 중요도가 떨어지게 되므로 공기빈도에 문서에서 단어의 빈도와 전체 문서 집합에서 단어의 빈도를 적용시켜서 각각 단어 조합들의 가중치를 구한다.

$$CoreQuery(q_i, q_j) = \sum_{D \in S} cooc(q_i, q_j) \cdot \left(\frac{tf(q_i)}{cf(q_i)} + \frac{tf(q_j)}{cf(q_j)}\right) \quad (1)$$

S 는 초기 검색 결과 상위 n 개의 문서집합에서 질의 어휘가 2 개 이상 발생한 클러스터의 문서이다. $Cooc(q_i, q_j)$는 문서에서 발생한 q_i, q_j 의 공기빈도이다. $tf(q_i)$는 문서에서 단어의 빈도수 이고, $cf(q_i)$는 전체 문서 집합에서 단어의 빈도수이다.

가중치 $CoreQuery(q_i, q_j)$가 가장 높은 한 쌍의 단어 조합이 핵심 질의로 선택된다. 핵심 질의를 포함한 모든 클러스터를 핵심 클러스터라 하고 모든 핵심 클러스터는 R_Pool 에 들어 간다.

4. 핵심 질의와 단어 근접도를 이용해 핵심 클러스터에서 적합문서 찾기

핵심 질의를 포함한다고 해서 모두 적합문서가 아니다. 즉, 핵심 클러스터 안에는 부적합 문서도 포함되어 있다. 핵심 클러스터 안에 문서들은 Query likelihood[10] 순서로 되어 있다. 초기 검색 결과에서 핵심 클러스터를 찾아 내면서 비적합문서를 필터링 하였다면 다음은 핵심 클러스터 안에서 핵심 질의와의 단어 근접도를 이용해 적합문서를 찾아냄으로써 상위 문서의 정확율을 높일 수 있다. 먼저 핵심 질의 주변에 나타난 단어들의 빈도를 이용해 핵심 질의 어휘에 대한 단어 근접도를 계산한다. 단어 근접도를 이용하여 핵심 클러스터 안에 문서들의 가중치를 계산하고 가중치가 높은 순서대로 순위화 한다.

핵심 질의 주변에 나타난 단어는 핵심 질의와 의미적으로 연관성이 있다. 적합문서에서 핵심 질의 주변에 빈번하게 나타난 단어가 다른 문서에도 많이 나타난다면 그 문서도 적합문서일 확률이 높다. 핵심 클러스터안에 문서 중 상위에 순위화된 문서를 적합 문서라 가정하고, 상위에 있는 문서들에서 핵심 질의 주변에 발생하는 문맥어휘의 근접도를 계산한다.

$$Context(t) = \sum_{d \in Rdocs} \sum_{t \in d} proxTF(t) \quad (2)$$

proxTF(t)는 핵심 질의 주변에 발생한 단어의 빈도수이고, Rdocs 는 단어 근접도를 구하기 위한 핵심 클러스터의 상위 문서의 개수이다. Rdocs 는 학습을 통해 실험한 결과 핵심 클러스터의 상위 10 개의 문서를 사용했고, 문맥어휘의 빈도는 핵심 질의 앞뒤로 각 50 개씩 사용하였다.

(2)번 공식을 통해 어떤 단어가 핵심 질의 주변에 몇 번 발생했는지 알 수 있다. 빈도가 높은 단어는 핵심 질의와 의미적으로 관계가 있다고 할 수 있다.

핵심 클러스터 안에 각각의 문서들을 문맥어휘를 이용해 가중치를 계산하고 가중치가 높은 순서대로 순위화한다. 문서에서 핵심 질의와 근접도가 높은 단어가 많이 발생했다면 그리고 그 단어가 문서에서 중요도가 높은 단어라면 그 문서는 적합 문서일 확률이 높다. 가중치 계산은 문서에서 문맥 어휘가 몇 번 발생했는지 그리고 문맥어휘가 문서에서 중요도가 얼마인지 또 문서의 초기 검색의 중요도는 얼마인지를 반영해서 계산한다. 문맥어휘 중 빈도가 높은 상위 e 개 단어를 사용한다. (실험에서 e=45). 문서에서 단어의 중요도는 tf·idf 를 사용했다. 그리고 초기 검색 결과의 중요도는 문서의 Query Likelihood 를 적용했다.

$$wgt(D) = \sum_{t_i \in d} \frac{1}{|Rdocs|} \cdot Context(t_i) \cdot tfidf(t_i) \cdot P(Q|D) \quad (3)$$

|Rdocs|는 문맥어휘를 구하는데 사용한 문서의 개수, $Context(t_i)$는 문맥어휘의 빈도수, P(Q|D)는 문서의 Query Likelihood 이다.

(3)번 공식을 핵심 클러스터 안에 문서들의 가중치를 결정하고 그에 따른 순위를 결정할 수 있다. 상위 문서의 정확율을 통해 유효성을 검증할 수 있으며 문맥어휘와 핵심 질의 사이에 근접도가 적합문서를 찾는데 유용함을 알 수 있다.

5. 실험 및 평가

실험 문서집합으로 뉴스기사 집합인 TREC AP(88-90)를 사용하였다. 질의 집합은 학습질의 (51-150), 테스트 질의 (151-200)를 사용하였다. 표 3 에서는 실험 데이터에 대한 정보를 보여준다. 초기 검색 결과를 위한 스무딩(smoothing) 파라미터는 학습을 통해 $\mu = 2000$ 으로 설정하였다.

<표 3> 실험데이터 집합

컬렉션	문서수	학습 질의		테스트 질의	
		질의 번호	개수	질의 번호	개수
AP(88-90)	242,918	51-150	100	151-200	50

<표 4> 사람이 찾은 핵심 클러스터와 공기빈도를 통해 선택한 핵심 클러스터의 결과

	질의 수	Tot_rel	사람이 찾은 것			공기빈도를 이용한 것		
			R_rel	R_doc	포함율	R_rel	R_doc	포함율
Q#3	11	685	615	2565	0.8978	602	2565	0.8788
Q#4	14	281	243	2021	0.8648	233	2180	0.8292
Q#5	7	238	209	757	0.8782	199	1315	0.8361
Q#6	4	100	63	379	0.6300	60	337	0.6000
Q#7 이상	4	89	82	661	0.9213	69	713	0.7753
합 계	40	1393	1212	6383	0.8701	1163	7110	0.8349

초기 검색 결과에서 사람 평가자가 직접 확인하고 선택한 핵심 클러스터와 제안한 알고리즘으로 선택한 핵심 클러스터의 포함율을 비교하여 평가한다.

<표 4>를 보면 전체적으로 사람이 직접 선택한 것과 제안된 알고리즘의 포함율이 비슷한 것을 볼 수 있다. 전체적인 적합문서 포함율을 83.5% 정도로 높은 포함율을 보인다.

핵심 클러스터의 문서들을 문맥어휘를 이용해 순위화 한 결과와 초기 검색 결과(Language Model) 상위에 순위화된 문서들의 정확율을 비교하여 평가한다.

<표 5> 언어모델과 핵심 클러스터에서 상위 문서의 정확율 비교

	P@100	P@50	P@5	P@1
언어모델 (LM)	0.1905	0.2349	0.2930	0.3023
핵심 클러스터	0.2022 (+6.14%)	0.2456 (+4.56%)	0.3209 (+9.52%)	0.3256 (+7.71%)

<표 5>는 언어모델의 상위 문서와 핵심 클러스터의 상위 문서들의 정확율을 보여준다. P@n 은 상위 n 개의 문서에서의 정확율을 나타낸다. 언어모델은 인드리 검색엔진을 통한 초기 검색결과이다. 핵심 클러스터가 언어모델보다 상위의 문서에서 더 높은 정확율을 보임을 확인할 수 있다. 특히 P@5 에서는 10%가까운 향상을 보였다.

6. 결론 및 향후 연구

본 연구에서는 질의 어휘 클러스터에서 공기빈도를 이용해 핵심 클러스터를 찾고 핵심 클러스터의 문서들을 핵심 질의와 주변 단어의 근접도를 통해 적합문서를 찾아냄으로써 상위문서의 정확율을 향상 시키는 알고리즘을 제안했다.

사람이 직접 선택한 질의 어휘 클러스터의 유효성 검증을 통해 길이가 긴 질의에서는 핵심 질의가 있음을 알 수 있으며 공기빈도를 이용해 찾아낸 핵심 클러스터는 적합문서 포함율이 83.5%로 사람이 직접 선택한 정확율과 거의 비슷한 정확율을 보였다. 또한 핵심 질의와의 단어 근접도를 통해 핵심 클러스터 문서에서 적합문서를 찾아낸 결과는 언어모델의 상위 문서와 정확율을 비교해본 결과 P@5 에서는 9.52%, P@1 에서는 7.71% 성능이 향상 됐음을 확인했다.

적합문서와 유사도가 높은 문서는 적합문서일 확률이 높다. 본 연구에서는 핵심 클러스터의 상위 문서의 정확도를 향상시켰다. 향후 연구는 핵심 클러스터의 상위 문서와 유사도를 적용하여 전체 문서의 정확율을 향상 시키는 알고리즘에 대해 연구 할 것이다.

참고문헌

[1] Collins-Thompson, K., and Callan, J., Estimation and use of uncertainty in pseudo-relevance feedback. In Proc. 30th ACM SIGIR conference, pp.303-310. 2007

[2] Bendersky, M., Coft, W.B., Discovering Key Concepts in Verbose Queries. In Proc 31th ACM SIGIR Conf on Research and Development in Information Retrieval, pp.491-498. 2008.

[3] A Hulth,. Inproved automatic keyword extraction given more linguistic knowledge. In Proc. Empirical Mothods in Natural Language Processing Conf. pp.216-223. 2003.

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

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

[6] Indri. http://www.lemurproject.org

[7] Sakai, T., Manabe, T., and Koyoma, M., Flexible pseudo-relevance feedback via selective sampling. ACM Transaction on Asian Language Information Processing(TALIP), 4(2), pp.111-135. 2005.

[8] Lv, Y., Zhai, C.X., Positional Language Models for Information Retrieval. In Proc. 32th ACM SIGIR Conf on Research and Development in Information Retrieval. pp.299-306, 2009.

[9] Zhao, J.G. and Yun.G., A Proximity Language Model for Information Retrieval. In Proc. 32th ACM SIGIR Conf on Research and Development in Information Retrieval. pp.291-298, 2009.

[10] Ponter, J.M., Croft, W.B., A language modeling approach to information retrieval. In Proc. 21th ACM SIGIR conference, pp.275-281. 1998.