DOI: 10.3745/KIPSTB.2010.17B.5.399

핵심질의 클러스터와 단어 근접도를 이용한 문서 검색 정확률 향상 기법

장 계 훈[†] · 이 경 순[††]

요 약

본 논문에서는 상위 검색결과 문서의 정확률을 향상시키기 위하여 핵심질의 클러스터와 단어 근접도를 이용해 재순위화하는 방법을 제안한다. 언어모델에 의한 초기 검색결과를 상위 문서에 대해 발생한 질의어휘 조합을 기반으로 문서를 클러스터링한다. 질의어휘 조합 클러스터에 대해 질의어휘 사이의 근접도를 이용하여 핵심질의 클러스터를 선택한다. 질의의 문맥정보를 이용해 핵심질의 클러스터의 문서를 재순위화한다. 뉴스집합인 TREC AP 컬렉션에 대해 언어모델과 제안한 방법의 문서 정확률을 비교한 결과 제안방법이 언어모델에 비해 상위 100개 문서(P@100)에서 11.2% 성능이 향상되었다.

키워드 : 질의 클러스터, 핵심질의, 단어 근접도, 문맥 어휘, 재순위화

A Method for Precision Improvement Based on Core Query Clusters and Term Proximity

Kye-Hun Jang[†] · Kyung-Soon Lee[††]

ABSTRACT

In this paper, we propose a method for precision improvement based on core clusters and term proximity. The method is composed by three steps. The initial retrieval documents are clustered based on query term combination, which occurred in the document. Core clusters are selected by using proximity between query terms. Then, the documents in core clusters are reranked based on context information of query. On TREC AP test collection, experimental results in precision at the top documents(P@100) show that the proposed method improved 11.2% over the language model.

Key Words : Query Term Cluster, Core Query, Term Proximity, Context Term, Reranking

1. 서 론

정보검색에서 검색된 결과를 이용해서 성능을 개선시키는 연구로는 재순위화[1,9]와 상위문서를 피드백하여 질의를 확장하는 잠정적 적합 피드백 기법[2, 3, 4]이 있다. 최근에 정보검색에서 성능을 향상시키기 위한 연구로 길이가 긴 질의에서 핵심개념[5, 6]을 찾거나 질의에서 불필요한 어휘를 제거[3]하여 질의의 핵심적인 의미는 간직한 채 간결하게 줄이는 연구가 성능 향상을 보여오고 있다. 또 어휘 근접도[7, 8]를 이용한 문맥정보를 반영하여 성능을 향상시키는 정보검색 모델이 연구되었다.

사용자가 정보를 검색할 때 사용하는 질의는 사용자의 의도에 따라 하나의 초점에 맞춰져 있다. 즉, 사용자는 한가지 개념에 대한 정보를 찾길 원한다. 질의어휘들의 의미는 서로 연관되어 있지만 각 어휘들의 개념은 다르다. 길이가 긴 질의에는 여러 가지 개념의 단어들이 포함되어 있지만 2~3개의 단어만 사용자가 원하는 핵심개념을 포함하고 있다[7]. 또한 같은 질의어휘를 포함하는 문서들은 유사한 정보를 포함하며[2], 문서 안에서 출현한 단어들은 서로 독립적으로 존재하는 것이 아니라 문서가 쓰여진 의도에 따라 서로 간에 의미적으로 연관되어 있다.

본 논문에서 가정 및 접근방법은 다음과 같다: (i) 같은 질의어휘가 포함된 문서들의 클러스터 즉, 질의어휘 조합 클러스터는 문서들의 행태가 비슷하다는 가정하에, 초기 검색된 결과에 대해 각 문서에서 발생한 질의어휘 조합을 기반으로 클러스터링한다. (ii) 길이가 긴 질의에는 두 개 또

† 정 회 원 : 전북대학교 컴퓨터공학과
†† 정 회 원 : 전북대학교 컴퓨터공학부/영상정보신기술연구센터(교신저자)
논문접수 : 2010년 7월 21일
수 정 일 : 1차 2010년 8월 24일
심사완료 : 2010년 10월 25일

는 세 개의 어휘가 핵심개념을 나타내고 있으며, 두 질의어휘의 근접도가 높으면 두 어휘는 핵심개념을 나타낸다는 가정을 기반으로 하여 질의어휘 사이의 근접도를 이용해 핵심질의를 선택한다. 질의어휘 조합 클러스터에서 핵심질의를 포함하는 클러스터를 핵심질의 클러스터(core query cluster)로 선택하고, 핵심질의를 포함하지 않는 클러스터의 문서는 부적합 문서라고 보고 필터링한다. (iii) 문서에서 핵심질의와 근접해서 나타나는 어휘들인 질의 문맥(context)은 핵심질의와 의미적으로 연관성이 높다는 가정하에, 핵심질의 클러스터 안에 있는 문서들에서 질의의 문맥어휘를 찾아내고 이를 이용하여 핵심질의 클러스터의 문서들을 재순위화한다. 제안된 방법의 유효성을 검증하기 위하여 TREC AP 테스트 컬렉션에 대해 실험한다.

본 논문의 구성은 2장에서 관련 연구를 소개하고, 3장에서는 질의어휘 클러스터링의 유효성과 핵심질의 클러스터 선택 방법에 대해 설명하고, 4장에서는 핵심질의 클러스터 안에 있는 문서들을 재순위화하는 기법을 제안한다. 5장에서 실험에 관한 정보를 6장에서는 실험에 대한 결론 및 향후연구에 대해 논하겠다.

2. 관련 연구

본 논문과 관련된 연구는 질의어휘 클러스터, 핵심질의 선택 및 질의어휘 변이에 따른 성능 변화, 어휘 근접도를 이용한 문맥정보에 관한 연구가 있다.

질의어휘 클러스터와 관련된 연구로 Sakai의 연구[2]에서는 질의어휘 클러스터를 통해 샘플링 문서를 선택하고 잠정적 적합 피드백(Pseudo-relevance feedback)에 사용하여 성능을 향상시키는 알고리즘을 제안했다. 초기 검색결과 상위에 순위화된 문서들은 비슷한 행태를 가지고 있다고 가정하고, 상위에 있는 문서를 피드백에 그대로 사용하면 비슷한 문서만을 가지고 피드백을 하기 때문에 효율적이지 못하다. 따라서 피드백 문서를 선택할 때 어떤 질의어휘 클러스터 안에 있는 문서의 개수가 임의의 개수가 넘으면 더 이상 그 단어조합이 발생한 문서가 나와도 클러스터에 포함시키지 않는 알고리즘을 제안함으로써 다양하고 새로운 문서 집합을 피드백에 사용한다. 본 논문에서는 초기 검색된 결과에서 핵심질의를 선택하기 위해 문서에서 발생한 질의어휘를 기반으로 상위 n개의 문서를 클러스터링하고, 핵심질의를 포함한 클러스터만을 모아놓은 핵심질의 클러스터를 찾는다.

핵심질의 선택에 관한 연구로 Bendersky의 연구[5]는 길이가 긴 질의에서 핵심개념(key concepts)을 선택하는 알고리즘을 제안했다. 길이가 긴 자연어 질의에서 명사만을 추출하고 그 단어의 빈도, 역문서 빈도, 첫 글자가 대문자인지 여부 등을 통해 중요도를 결정하여 그 단어가 핵심질의인지 아닌지를 결정한다. Kumaran의 연구[6]는 길이가 긴 질의에서 부분질의(sub-query)를 찾아내는 알고리즘을 제안했다. 질의에서 발생할 수 있는 모든 부분질의를 고려하여 가장 좋은 부분질의를 찾아낸다.

질의 변이에 따른 성능변화 연구[3]는 질의어휘들 중에서 하나는 불필요한 어휘일 것이라 가정하고 질의어휘 중 하나의 어휘를 제거해 만든 질의 변이(Query variant)를 이용하여 검색결과를 샘플링하고 피드백에 사용하는 방법을 제안하였다. 본 논문에서는 질의어휘가 2개 이상 발생한 모든 문서 안에서 질의어휘 조합 사이의 근접도가 가장 높은 한 쌍의 단어조합을 핵심질의로 선택한다.

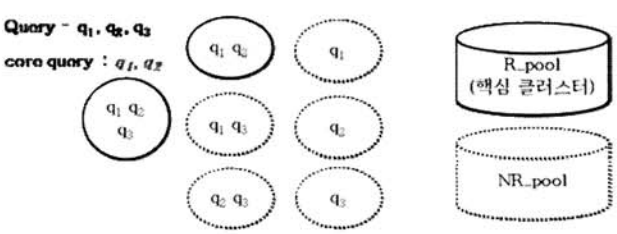
어휘 근접도를 이용한 문맥정보에 관한 연구는 질의어휘 위치기반 언어모델[7]에서 질의어휘들 사이의 거리를 코사인, 가우시안 등의 함수 그래프를 이용해서 표현하고 그 가중치를 이용해 성능을 향상시켰다. 문서에서 하나의 질의어휘 주변에 다른 질의어휘가 발생하면 두 어휘의 그래프가 겹치게 되어 그 문서는 두 어휘의 가중치를 모두 받게 된다. 즉, 두 질의어휘가 가까울수록 더 높은 가중치를 받게 된다. 하지만 질의어휘 사이의 거리만으로 문서의 중요도를 결정하기는 쉽지 않다. 본 논문에서는 문서에서 핵심질의뿐만 아니라 핵심질의의 문맥정보 역시 적용하여 문서의 중요도를 결정한다.

3. 핵심질의 클러스터 선택 기법

핵심질의 클러스터 선택 기법은 본 논문의 첫 번째 단계로 긴 질의에서 핵심질의를 찾고 초기 검색결과에서 일차적으로 부적합한 문서를 필터링하기 위해 사용되는 기법이다. 초기 검색결과에서 상위 n개의 문서를 발생한 질의어휘를 기반으로 클러스터링하고 질의의 위치정보를 이용하여 핵심질의를 찾는다. 핵심질의를 포함하는 클러스터를 적합 문서 후보집합(R_Pool)에 핵심질의를 포함하지 않는 클러스터를 부적합 문서 후보집합(NR_Pool)에 나눈다. 예를 들어, r개의 질의어휘를 가진 질의는 최대 2^r-1개의 클러스터가 생성될 수 있다. (그림 1)에서는 질의어휘가 3개인 질의에서 생성될 수 있는 모든 클러스터를 보여준다. 질의어휘가 3개이면 2^3 - 1, 총 7개의 클러스터가 생성된다. 3개의 질의어휘 q_1, q_2, q_3 중 q_1, q_2를 핵심질의로 선택했을 때 q_1, q_2를 포함하는 두 개의 클러스터가 핵심질의 클러스터로 선택된다. 사람 평가자가 직접 핵심질의를 선택하여 찾은 핵심질의 클러스터와 제안한 방법으로 찾은 핵심질의 클러스터 안에 포함된 적합 문서의 포함률을 비교함으로써 핵심질의 클러스터의 유효성을 검증한다.

3.1 질의어휘 조합을 기반으로 한 클러스터링

질의어휘 클러스터는 같은 질의 조합이 발생한 문서들의



(그림 1). 핵심질의 클러스터 선택 방법

클러스터이다. 같은 질의 조합을 공유하는 문서들은 행태가 비슷하다. 따라서 핵심질의 조합을 포함하고 있는 클러스터의 문서들은 적합한 문서일 가능성이 높다. <표 1>은 실제 질의를 통해 검색된 결과 중 상위 300개의 문서에 대한 질의어휘 클러스터를 보여준다. 3개의 질의어휘를 가진 "fiber optics applications"는 총 2^3-1개, 즉 7개의 클러스터가 발생할 수 있다. 예를 들어, 클러스터 C1은 300개의 문서 중 "fiber"와 "optics", 그리고 "applications" 세 단어를 모두 포함하고 있는 문서들로, 이 클러스터 안에는 19개 문서 중 4개가 적합 문서다. C2는 "fiber"와 "optics"를 포함하는 문서들의 클러스터이고, 158개의 문서 중 36개가 적합 문서다. 실험집합에서 총 7개의 클러스터의 전체 적합 문서의 수는 40개이다.

만약 "fiber optics"를 핵심질의로 선택한다면 "fiber optics"를 포함한 C1, C2 두 개의 클러스터가 적합 문서 후보집합(R_Pool)에 들어가게 된다. 그렇게 되면 R-Pool에는 총 300개 문서 중 177개의 문서가 들어가게 되고 40개의 적합 문서가 모두 들어가게 된다. 초기 검색결과에서 상위 300개 문서의 정확도(Accuracy)는 0.1333(40/300)이 되지만 핵심질의를 통해 찾은 R_Pool의 정확도는 0.2260(40/177)이 된다.

이와 같이 "fiber optics"는 전체 질의에서 핵심개념이라 할 수 있으나 "applications"는 불필요한 단어라고 할 수 있다. 또 "fiber"와 "optics"는 각 어휘만으로는 의미를 전달하기는 어렵다는 것을 알 수 있다. 핵심질의만 잘 찾게 되면 R_Pool에 거의 모든 적합 문서를 포함할 수 있다.

<표 2>는 질의어휘가 3개 이상인 TREC AP 학습 질의 73개(실험집합 정보는 5장에 참조)에 대해 사람이 각 질의에서 직접 핵심질의를 선택한 적합 문서 후보집합(R_Pool)과 부적합 문서 후보집합(NR_Pool)의 포함률과 누락률을 보여준다.

$$\text{• 포함률(recall)} = \frac{\text{R\_Pool의 적합문서수(R\_rel)}}{\text{전체적합문서수(Tot\_rel)}}$$

$$\text{• 누락률(miss alarm)} = \frac{\text{NR\_Pool의 적합문서수(NR\_rel)}}{\text{전체적합문서수(Tot\_rel)}}$$

여기서 Q#n은 n개의 어휘를 가진 질의를 말하며, Tot_rel은 각 질의에 대한 초기 검색결과 상위 300개 안에 모든 적

〈표 1〉 질의 "fiber optics applications"의 검색결과에 대한 질의어휘 클러스터

	질의 조합에 의한 클러스터	검색된 문서 개수	적합 문서 개수
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
합 계		300	40

〈표 2〉 질의어휘가 3개 이상인 학습질의에 대해 사람이 직접 판별한 R_Pool과 NR_Pool의 포함률과 누락률

	질의 개수	총 적합 문서 수 (Tot_rel)	적합 문서 후보집합(R_Pool)			부적합 문서 후보집합(NR_Pool)		
			적합 문서 수 (R_rel)	문서 수 (R_doc)	포함률	적합 문서 수 (NR_rel)	문서 수 (NR_doc)	누락률
Q#3	36	1653	1527	4924	0.9238	126	2876	0.0762
Q#4	27	1073	1002	3040	0.9338	72	3010	0.0671
Q#5	12	777	720	2254	0.9266	57	1346	0.0734
Q#6 이상	8	281	192	1166	0.6833	89	1234	0.3167
총 문서수	73	3784	3441	13384	0.9094	344	8516	0.0909

합 문서의 수, R_rel은 R_Pool에 포함된 적합 문서의 개수, R_doc는 R_Pool에 포함된 전체 문서의 개수, NR_rel은 NR_Pool에 포함된 적합 문서의 개수, NR_doc는 NR_Pool에 포함된 모든 문서의 개수를 의미한다.

보는 바와 같이 질의어휘가 6개 이상인 질의를 제외한 모든 질의에서 90%이상의 포함률을 보이고 있다. 이것을 통해 질의어휘 클러스터가 유효함을 알 수 있다.

3.2 단어 근접도를 이용한 핵심질의 클러스터 선택 방법

핵심질의를 포함하고 있는 모든 클러스터를 핵심질의 클러스터라고 정의한다. 핵심질의 클러스터를 찾기 위해 먼저 질의에서 핵심질의를 찾아야 한다. 임의의 두 질의어휘가 서로 거리가 가깝다거나 일정한 거리(window size)안에 자주 발생하면 두 단어는 서로 의미적인 연관도가 높은 핵심질의라고 할 수 있다. 공기빈도(co-occurrence)란 한 문서에서 두 개의 단어가 일정한 거리 안에서 연속으로 발생한 빈도를 말한다. 공기빈도는 질의어휘가 1개 발생한 클러스터를 제외한 모든 클러스터에서 계산한다. 또한 클러스터 안에 모든 질의어휘 조합을 고려한다. 예를 들어, q_1, q_2, q_3 세 개의 질의어휘를 포함한 클러스터는 (q_1, q_2), (q_1, q_3), (q_2, q_3) 세가지 질의어휘 조합의 공기빈도를 구한다.

각 문서에서 모든 어휘조합 사이의 공기빈도를 구하고 질의어휘가 2개 이상 발생한 모든 클러스터의 문서에서 더한다. 하지만, 문서에서 두 질의어휘 사이의 거리만 가깝다고 해서 핵심질의라고 하긴 어렵다. 문서에서 두 질의어휘의 중요도를 반영하면 거리가 가까운 질의어휘가 문서에서 얼마나 중요한지 나타낼 수 있다. 따라서, 두 단어의 가중치에 따라 근접도가 비례하도록 두 단어의 공기빈도에 문서 안에서 각 단어의 가중치(tfidf(q_i))를 적용한다.

$$CoreQuery(q_i, q_j) = \sum_{q_i, q_j \in D} cooc(q_i, q_j) \cdot (tfidf(q_i) + tfidf(q_j)) \quad (1)$$

여기서 D는 초기검색결과 상위 n개의 문서집합에서 질의어휘가 2개 이상 발생한 클러스터의 문서이다. cooc(q_i, q_j)는 q_i, q_j의 문서에서 일정한 거리(d1;학습을 통해 가장 성능이 좋은 것으로 설정) 안에 발생한 공기빈도이다. tfidf(q_i)는 문서 안에서 단어 q_i의 tf · idf값이다.

식(1)을 통해 CoreQuery(q_i, q_j)가 가장 높은 한 쌍의 어휘 조합이 핵심질의로 선택된다. 핵심질의를 포함한 모든 클러스터를 핵심질의 클러스터라 하고 모든 핵심질의 클러스터는 적합 문서 후보집합(R_Pool)에 들어가게 된다.

4. 핵심질의와 단어 근접도를 이용한 핵심질의 클러스터에서의 재순위화

핵심질의를 포함한다고 해서 모두 적합 문서가 아니다. 핵심질의 클러스터 안에 포함된 문서들은 언어모델(Query-Likelihood Language Model)[10]로 순위화되어 있으며, 부적합 문서도 포함되어 있다. 앞에서 초기 검색결과에서 핵심질의 클러스터를 찾아내면서 부적합 문서를 필터링했다면, 본 장에서는 핵심질의 클러스터 안에 포함된 문서들을 핵심질의와의 문맥정보를 이용해 문서의 중요도를 재조정함으로써 상위에 검색된 문서를 재순위화한다. 문서 안에서 출현한 단어들은 서로 독립적으로 존재하는 것이 아니라 문서가 쓰여진 의도에 따라 서로 간에 의미적으로 연관되어 있다. 또한 문서에서 핵심질의와 일정한 거리를 두고 나타난 단어들은 의미적으로 밀접하게 연관되어 있다. 이를 이용해 핵심질의에 대한 각 단어들의 근접도를 계산하여 질의의 문맥어휘(context term)를 찾아낸다. 핵심질의 클러스터 안에 포함된 문서들에 대해 문맥어휘를 이용해 중요도를 계산하고 재순위화한다.

4.1 단어 근접도를 이용한 문맥어휘 선택 기법

핵심질의 주변에 나타난 단어는 핵심질의와 의미적으로 연관성이 있다. 적합 문서에서 핵심질의 주변에 빈번하게 나타난 단어가 다른 문서에도 많이 나타난다면 그 문서도 적합 문서일 가능성이 높다. 핵심질의와 가까운 거리에서 빈번하게 발생하는 어휘를 질의의 문맥어휘라 한다. 핵심질의 클러스터 안에 있는 문서 중 상위에 순위화된 문서를 적합 문서라 가정하고, 상위에 있는 각 문서들에서 핵심질의 주변에 발생하는 어휘들의 빈도를 계산한다.

$$Context(t) = \sum_{D \in Rdocs} \sum_{t \in D} proxTF(t) \qquad (2)$$

여기서 t는 문서에서 핵심질의와 일정한 거리(d2;학습을 통해 가장 좋은 것으로 설정) 안에 발생한 단어이고, proxTF(t)는 t의 빈도이다. Context(t)는 proxTF(t) 값을 핵심질의 클러스터의 상위문서에서 더한 값이다. D는 핵심질의 클러스터 안에 있는 문서들이고, Rdocs는 문맥어휘를 구하기 위한 핵심질의 클러스터의 상위에 검색된 문서이다. 핵심질의 클러스터의 모든 문서가 적합 문서라고 할 수 없기 때문에 문맥어휘를 구하기 위해 핵심질의 클러스터의 상위 |Rdocs|개로 학습한다. 질의의 문맥어휘는 Context(t)값이 높은 순서대로 e개를 선택한다.

4.2 핵심질의 클러스터 안에 있는 문서의 문맥어휘를 이용한 재순위화

핵심질의 클러스터 안에 포함된 문서는 초기 검색 중요도로 순위화되어 있다. 본 장에서는 핵심질의 클러스터에 포함된 문서들의 초기 검색 중요도를 재조정함으로써 정확률을 향상시킨다. 문맥어휘는 핵심질의 주변에 빈번하게 발생한 어휘로, 질의와 의미적으로 연관성이 있다. 따라서, 문맥어휘는 초기 질의의 확장된 질의라고 볼 수 있다. 핵심질의 클러스터의 문서는 적합모델(Relevance Model)[4]을 이용해 재순위화한다. 적합모델은 초기 질의와 문맥어휘를 통해 확장된 단어를 결합하여 문서의 가중치를 재조정한다.

$$Score(D) = \lambda \cdot P(Q \mid D) + (1-\lambda) \cdot P(Q' \mid D) \qquad (3)$$

여기서 D는 핵심질의 클러스터 안에 있는 문서들이고, λ는 원래 질의 Q에 대한 가중치, P(Q|D)는 언어모델에 의한 초기 검색결과 값이며, Q'은 수식(2)에서 결정한 확장 질의어휘이다. 따라서, P(Q'|D)는 문서 D의 확장된 질의 Q'에 대한 언어모델 검색결과다.

식(3)을 통해 핵심질의 클러스터 안에 포함된 문서들을 재순위화한다. 핵심질의 클러스터에 속하지 않는 문서들은 언어모델에 의한 초기 검색결과 값 순서대로 핵심질의 클러스터에 포함된 문서들 다음에 순위화된다.

5. 실험 및 평가

실험 문서 집합으로 뉴스기사 집합인 TREC AP(88-90)를 사용하였다. 질의 집합은 파라미터 추정을 위해 학습질의 73개를 이용하였고, 테스트 질의 43개에 대해서 평가하였다. 실험 데이터에 대한 정보는 <표 3>에서 보여준다.

색인과 검색은 언어모델(LM)을 기반한 인드리(Indri-2.8)[11] 시스템을 사용하였다.

언어모델(LM)의 수식은 다음과 같다.

$$P(Q \mid D) = \prod_{i=1}^{k} \left(\frac{|D|}{|D|+\mu} \cdot \frac{f_{q_i,D}}{|D|} + \frac{\mu}{|D|+\mu} \cdot \frac{c_{q_i}}{|C|} \right) \qquad (4)$$

여기서 k는 질의어휘의 개수이고, |D|는 문서의 길이, |C|는 전체 컬렉션의 길이, fqi,D는 문서 D에서의 질의 어휘 qi의 빈도수, Cqi는 전체 컬렉션에서의 qi의 빈도수를 나타낸다. μ는 디리슈레 스무딩(Dirichlet smoothing) 파라미터로 μ값은 학습질의에 대한 실험(μ∈{500, 1000, 1500, 2000, … ,5000})에서 MAP(mean average precision)가 가장 높은 값을 보인 2000으로 설정하였다.

질의어휘 클러스터를 위한 상위 n개 문서는 300으로 설정하여 클러스터링하였다. 3.1장에서 공식(1)의 핵심질의를 선택하기 위한 공기빈도 계산에서 단어 사이의 거리는 실험(d1∈{5, 10, 15, 20, 30, 50, 75, 100})을 통해 학습 질의에 대

〈표 3〉 TREC 테스트 컬렉션

컬렉션	문서 수	학습 질의		테스트 질의	
		질의 번호	개수	질의 번호	개수
AP(88-90)	242,918	51-150	100	151-200	50
질의어휘가 3개 이상인 질의의 개수		73		43	

해 가장 좋은 성능을 보인 15로 설정하였다. 4.1장에 식(2)의 문맥어휘를 찾기 위해 사용한 핵심질의 클러스터 문서의 수(|Rdoc|∈{5, 10, 25, 50, 75, 100})는 가장 좋은 성능을 보인 10으로 설정하고, 문맥어휘를 찾기 위한 핵심질의와 단어 사이의 거리(d2∈{5, 10, 15, 25, 50, 75, 100})는 학습을 통해 50으로 설정하였다. 4.1장에 식(2)의 문맥어휘를 통해 확장된 어휘 개수(e∈{5, 10, 25, 30, 35, 40, 45, 50, 75, 100})는 가장 좋은 성능을 보인 45개로 설정했다.

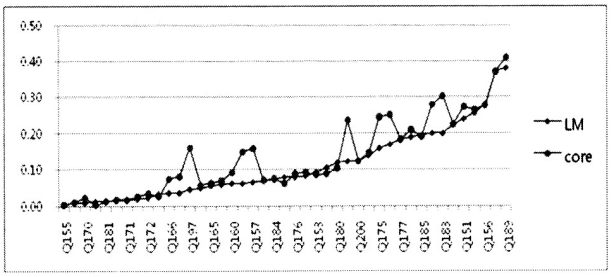
5.1 핵심질의 클러스터 선택 실험 결과

핵심질의 클러스터 선택 실험은 초기 검색결과에서 사람 평가자가 직접 확인하고 선택한 핵심질의 클러스터와 제안한 방법으로 선택한 핵심질의 클러스터의 포함률을 비교하여 평가한다.

전체적으로 사람이 직접 선택한 것과 제안된 방법의 포함률이 비슷한 것을 볼 수 있다. 질의어휘개수가 7개 이상인 질의에서는 클러스터의 개수가 많기 때문에 조금 떨어지는 경향이 있지만 전체적으로 적합 문서 포함률은 83.5%정도로 높은 포함률을 보인다. 질의어휘의 개수가 9개 이상인 질의는 클러스터의 개수가 많아서 핵심질의 클러스터를 선택하기가 어렵기 때문에 결과에서 배제했다.

〈표 4〉 사람이 찾은 핵심질의 클러스터와 제안된 방법을 통해 선택한 핵심질의 클러스터의 결과

	질의 개수	Tot_rel	사람이 찾은 것			제안된 방법 (공기빈도 이용)		
			R_rel	R_doc	포함률	R_rel	R_doc	포함률
Q#3	11	684	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, Q#8	3	89	82	661	0.9213	69	713	0.7753
합 계	40	1393	1212	6383	0.8701	1163	7110	0.8349



(그림 2) 질의어휘가 3개 이상인 각 질의에서 언어모델과 핵심질의 클러스터의 정확률 비교

(그림 2)는 AP문서 집합에 대해 질의어휘가 3개 이상인 각 질의에서 언어모델을 통한 검색결과 상위 300개 문서와 핵심질의 클러스터의 문서(각 질의 마다 평균 219개의 문서를 포함한다.)의 정확률을 보여준다. 그림에서 LM은 언어모델을 core는 핵심질의 클러스터를 나타낸다. 핵심질의 클러스터가 언어모델보다 전체적으로 정확률이 높음을 확인할 수 있다. 결과를 통해 공기빈도를 이용해 선택한 핵심질의 클러스터는 유효함을 알 수 있다.

5.2 핵심질의 클러스터의 재순위화 실험결과

핵심질의 클러스터의 문서들을 문맥 어휘를 이용해 재순위화한 결과와 초기 검색결과 상위에 순위화된 문서들의 정확률을 비교하여 평가한다.

<표 5>는 언어모델의 상위에 검색된 문서와 핵심질의 클러스터의 상위에 검색된 문서의 정확률을 보여준다. P@n은 상위 n개의 문서에서의 정확률을 나타낸다. 언어모델은 인드리 검색엔진을 통한 초기 검색결과다. 실험 결과 제안된 방법이 언어모델보다 상위 100개의 문서(P@100)에서 11.2% 향상되었고, 50개의 문서(P@50)에서 11.1% 성능이 향상되었다.

〈표 5〉 언어모델과 핵심질의 클러스터에서 상위에 검색된 문서의 정확률 비교

	P@100	P@50	P@20
언어모델 (LM)	0.1905	0.2872	0.3023
제안 방법	0.2119(+11.2%)	0.2609 (+11.1%)	0.3000 (+4.5%)

6. 결론 및 향후 연구

본 논문에서는 질의어휘 클러스터에서 단어 근접도를 반영하기 위해 공기빈도를 이용한 핵심질의 클러스터를 찾고 핵심질의 클러스터의 문서들을 핵심질의와 주변단어의 근접도를 통해 문서를 재순위화함으로써 상위문서의 정확률을 향상시키는 기법을 제안했다.

사람이 직접 선택한 질의어휘 클러스터의 유효성 검증을 통해 길이가 긴 질의에서는 핵심질의가 있음을 알 수 있으며 핵심질의를 찾기 위한 방법으로 질의어휘 조합을 기반으로 한 클러스터가 유용하게 쓰일 수 있음을 알 수 있다. 공기빈도를 이용해 찾아낸 핵심질의 클러스터는 적합 문서 포함률(83.5%)이 사람이 직접 선택한 포함률(87%)과 거의 비슷하게 나타났다. 또한 핵심질의와의 단어 근접도를 통해 핵심질의 클러스터 문서에서 적합 문서를 찾아낸 결과는 언어모델의 상위에 검색된 문서와 정확률을 비교해본 결과 상위 100개 문서(P@100)에서 11.2% 향상되었고, 상위 50개 문서(P@50)에서는 11.1% 향상되었다. 실험을 통해 길이가 긴 질의에서 질의어휘 사이의 근접도를 이용해 핵심질의를 찾

을 수 있고, 문맥어휘를 이용해 핵심질의 주변에 있는 단어는 핵심질의와 의미적인 연관성이 있다는 것을 확인할 수 있었다.

향후 연구에서는 핵심질의 클러스터의 상위에 검색된 문서와 유사도를 적용하여 전체 문서의 정확률을 향상시키는 방법에 대해 연구할 것이다.

참 고 문 헌

[1] Balinski, J., Danilowicz, C. 2004. Re-ranking method based on inter-document distances. Information Processing and Management, 41(2005)759-775.

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

[3] Collins-Thompson, K., Callan, J. 2007. Estimation and Use of Uncertainty in Pseudo-relevance Feedback. In Proc. of 30th ACM SIGIR on Research and Development in Information Retrieval. pp.303-310.

[4] Lavrenko, V., Croft, W.B. 2001. Relevance-based language models. In Proc. of 24th ACM SIGIR on Research and Development in Information Retrieval. pp.120-127.

[5] Bendersky, M., Croft, W.B. 2008. Discovering Key Concepts in Verbose Queries. In Proc. of 31st ACM SIGIR on Research and Development in Information Retrieval. pp.491-498.

[6] Kumaran, G., Allan, J. 2008. Effective and Efficient User Interaction for Long Queries. In Proc. of 31st ACM SIGIR on Research and Development in Information Retrieval. pp.11-18.

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

[8] Zhao, J., Yun, Y. 2009. A Proximity Language Model for Information Retrieval. In Proc. of 32nd ACM SIGIR on Research and Development in Information Retrieval. pp.291-298.

[9] Seo, J.W., Jeon, J.W. 2009. High Precision Retrieval Using Relevance-Flow Graph. In Proc. of 32nd ACM SIGIR on Research and Development in Information Retrieval. pp 694-695.

[10] Ponte, J.M., Croft, W.B. 1998. A Language Modeling Approach to Information Retrieval. In Proc. of 21st ACM SIGIR on Research and Development in Information Retrieval. pp.275-281.

[11] Strohman, T., Metzler, D., Turtle, H., and Croft, W.B. 2005. Indri: A language model-based search engine for complex queries. In proc. International Conference on Intelligence Analysis. http://www.lemurproject.org

[12] 신승은, 강유환, 오효정, 장명길, 박상규, 이재성, 서영훈, 2003. 문서필터링을 위한 질의어 확장과 가중치 부여기법. 정보처리학회 논문지. 제10-B권. 제7호. pp.743-750.

장 계 훈

e-mail : ghjang@chonbuk.ac.kr

2009년 전북대학교 컴퓨터공학과 학사

2009~현 재 전북대학교 컴퓨터공학과 석사과정

관심분야 : 정보검색, 정보 마이닝, 자연언어처리

이 경 순

e-mail : selfsolee@chonbuk.ac.kr

1994년 계명대학교 컴퓨터공학과 학사

1997년 한국과학기술원 전자전산학 석사

2001년 한국과학기술원 전자전산학 박사

2001~2003 일본 국립정보학연구소 (National Institute of Informatics) 연구원

2004년~현 재 전북대학교 컴퓨터공학부/영상정보신기술연구센터 부교수

관심분야 : 정보검색, 정보 마이닝, 자연언어처리