핵심질의 클러스터를 이용한 어휘 그래프 기반 질의 확장

조승현, 장계훈, 이경순
전북대학교 컴퓨터공학과
e-mail : {jackaa, ghjang, selfsolee}@jbnu.ac.kr

Query Expansion Based on Word Graph Using Core Query Clusters

Seung-Hyeon Jo, Gye-Hun Jang, Kyung-Soon Lee
Dept. of Computer Engineering, Chonbuk National University

요 약

본 논문에서는 질의 조합을 기반한 핵심질의 클러스터와 비핵심질의 클러스터를 각각 어휘 그래프로 표현한다. 이 때, 핵심질의 클러스터는 잠정적 적합 문서 집합으로, 비핵심질의 클러스터는 잠정적 부적합 문서 집합으로 본다. 핵심질의 클러스터의 어휘 그래프에서 비핵심질의 클러스터의 어휘 그래프를 빼서 확장어휘를 선택한다. 본 논문의 유효성을 검증하기 위해 웹문서 테스트컬렉션인 TREC WT10g 에 대해 실험하였고, 언어모델보다 평균정확률의 평균(MAP)이 9.4% 향상되었다.

1. 서론

정보검색 분야에서 질의 확장은 검색결과의 정확률 및 재현률을 향상시키는 방법으로 많은 연구가 되어오고 있다

질의 확장을 할 때 적합 문서만을 이용할 경우, 질의와 연관되어 있지 않은 어휘가 확장 어휘로 선택될 수 있게 된다. 따라서 부적합 문서를 이용하여 질의와 연관되어 있지 않은 어휘의 가중치를 줄여주면 질의 확장을 할 때 도움을 줄 수 있다.

질의에서 핵심 개념[1, 2]을 찾거나 질의에서 발생하는 모든 부분질의(sub-query)[3, 4]를 이용해서 질의의 핵심적인 의미는 간직한 채 간결하게 줄이려는 연구는 계속 되어왔다.

질의를 확장하는 연구로는 피드백 문서 안에서 어휘의 위치를 코사인, 가우시안 등 함수의 그래프를 이용하여 적합모델에 적용시킨 연구[5, 6]가 있다. 또한, 질의 어휘와 어휘 사이의 거리를 어휘 그래프(Word Graph)[7, 8]에 적용하여 질의를 확장한 연구가 있다.

본 연구에서의 접근방법은 다음과 같다. (ⅰ) 질의 조합을 기반한 클러스터에서 질의 어휘 사이의 근접도를 이용하여 핵심질의 클러스터(잠정적 적합 문서 집합)와 비핵심질의 클러스터(잠정적 비적합 문서 집합)로 나눈다 (ⅱ) 핵심질의 클러스터와 비핵심질의 클러스터에서 어휘 그래프를 이용하여 질의와 어휘 사이의 근접도를 적용한다. (ⅲ) 비핵심질의 클러스터의 어휘 그래프의 결과를 이용하여 핵심질의 클러스터의 어휘 그래프의 결과를 재조정한 후 질의 확장을 한다.

제안된 방법의 유효성을 검증하기 위해 TREC WT10g 테스트컬렉션에 대해 실험하고, 잠정적 적합성 피드백 모델에서 우수한 성능을 보인 적합모델(RM; Relevance Model)[9]과 비교함으로써 성능을 평가한다.

2. 잠정적 적합 및 부적합 문서클러스터를 이용한 어휘 확장

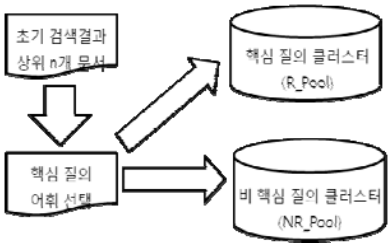
2.1 핵심질의 및 비핵심질의 클러스터 선택

클러스터는 긴 질의에서 핵심질의를 찾고 언어모델(LM; Language Model)에 의한 초기 검색 결과에서 부적합한 문서를 필터링하기 위해 사용된다. 초기 검색 결과는 상위 n 개의 문서를 사용한다.

길이가 긴 질의에는 2 개 또는 3 개의 핵심 어휘가 있다고 가정하고, 핵심 어휘를 찾기 위해 초기 검색 결과의 문서에서 발생한 질의 어휘 조합을 기반으로 문서를 클러스터링한다. r 개의 질의 어휘를 가진 질의는 최대 $2^r - 1$ 개의 클러스터가 발생할 수 있다.

클러스터 중 핵심질의 어휘를 포함하고 있는 모든 클러스터를 핵심 클러스터라고 정의하며, 이 클러스터들은 Rpool(핵심질의 클러스터)로 들어가게 되며, 핵심 클러스터가 아닌 경우에는 NRpool(비핵심질의 클러스터)로 들어가게 된다(그림 1). 예를 들어, 3 개의 질의 어휘 q_1, q_2, q_3 중 q_1, q_2 를 핵심질의로 선택했을 때 q_1, q_2 를 포함하는 두 개의 클러스터가 핵심 클러스터로 선택된다.

핵심 클러스터를 찾기 위해 먼저 질의에서 핵심질의 어휘를 찾아야 한다.



(그림 1) 핵심질의 클러스터 및 비핵심질의 클러스터 선택 방법

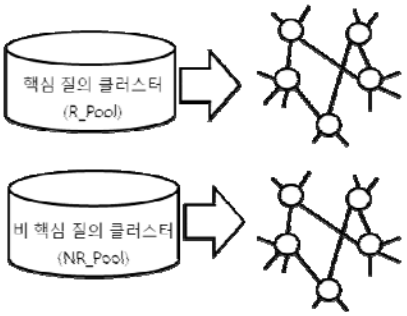
임의의 두 질의 어휘가 일정한 거리(window size)안에 자주 발생하면 두 어휘는 질의 안에서 핵심이 되는 질의 어휘라고 생각할 수 있다. 핵심이 되는 질의 어휘는 수식(1)과 같이 공기 빈도를 이용하여 계산하였다[10].

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

가중치가 가장 높은 한 쌍의 어휘 조합이 핵심질의로 선택되며, 핵심질의를 포함한 모든 클러스터를 핵심 클러스터(잠정적 적합 문서 집합)라 하고 모든 핵심 클러스터는 Rpool 에, 비핵심질의 클러스터(잠정적 부적합 문서 집합)들은 NRpool 에 들어가게 된다.

2.2 잠정적 적합 및 부적합 문서집합에 대한 어휘 그래프 표현

본 연구에서는 핵심질의 클러스터에 속하는 문서들을 하나의 그래프로 표현을 하고, 비핵심질의 클러스터에 속하는 문서들을 또 다른 그래프로 표현을 한다. 즉, 두 개의 어휘 그래프를 생성한다(그림 2).



(그림 2) 핵심질의 클러스터와 비핵심질의 클러스터에서 어휘그래프를 표현 방법

문서에 속하는 각 어휘에 대한 그래프는 G=(V, E)로 표현할 수 있다. V 는 그래프의 노드로써 문서에서 각 어휘를 나타내고, E 는 어휘 사이의 근접도를 에지로 가중치로 표현한다. 이 때, 노드의 가중치는 수식(2)로 계산한다.[11].

$$f^{r+1}(t_i) = \alpha \times f^0(t_i) + (1-\alpha) \times \sum_{j=1}^{K} \sum_{q_j \in Near(t_i)} \frac{w(t_i, q_j) \times f^r(q_j)}{\sum_{t_k \in Near(q_j)} w(q_j, t_k)} \quad (2)$$

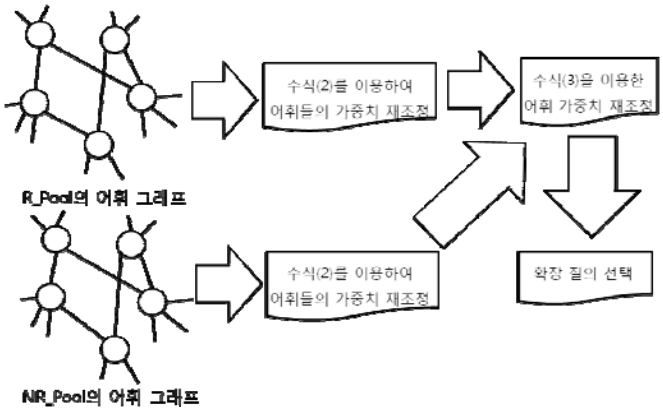
수식(2)에서 $f^0(t_i)$는 t_i 의 초기 가중치, k 는 질의 어휘의 수, $w(t_i, q_j)$는 t_i와 q_j사이의 근접도, $Near(t_i)$는 문서 안에서 t_i 와 근접하게 나타난 단어들이며, $f^0(t_i)$는 적합모델을 이용하여 계산한다.

본 논문에서는 핵심질의 클러스터(잠정적 적합 문서 집합)과 비핵심질의 클러스터(잠정적 부적합 문서 집합)에 대하여 어휘 그래프로 표현한다.

2.3 어휘 그래프를 이용한 질의 확장

적합 문서에서 높은 가중치를 얻은 어휘는 검색에 도움을 주지만, 비적합 문서에서 높은 가중치를 얻은 어휘는 검색에 도움을 주지 못할 것이다.

본 연구에서는 핵심질의 클러스터에 속하는 문서들을 어휘 그래프로 표현하여 얻은 어휘들의 가중치와 비핵심질의 클러스터에 속하는 문서들을 어휘 그래프로 표현하고 수식(2)를 이용하여 가중치를 재조정한다.



(그림 3) 핵심질의 클러스터와 비핵심질의 클러스터의 어휘그래프를 이용한 질의 확장 방법

어휘의 가중치를 재조정하는 방법은 수식(3)과 같다.

$$score(t) = \alpha \cdot \frac{1}{|D_{Rpool}|} \cdot \sum_{t \in Rpool} wgt(t) - \beta \cdot \frac{1}{|D_{NRpool}|} \cdot \sum_{t \in NRpool} wgt(t) \quad (3)$$

수식(3)에서 score(t)는 조정된 어휘의 가중치를 의미하며, wgt(t)는 어휘 그래프를 이용하여 구한 어휘의 가중치를 의미한다. $|D_{Rpool}|$은 핵심질의 클러스터에 들어있는 문서의 수, $|D_{NRpool}|$은 비핵심질의 클러스터에 들어있는 문서의 수이다. 어휘의 가중치는 핵심질의 클러스터에서 나온 어휘의 가중치에 α 를 곱한 뒤, 비핵심질의 클러스터에서 나온 어휘의 가중치에 β 를 곱하여 뺀 값이 조정된 어휘의 가중치가 된다.

제안된 방법을 통해 조정된 어휘의 가중치가 높은 상위 e 개의 단어를 확장 단어 w 로 선택하여 수식(4)를 통해 문서의 중요도를 결정한다.

$$P(Q'|D) = \lambda \cdot P(Q|D) + (1-\lambda) \cdot P(w|D) \quad (4)$$

여기서 P(Q'|D)는 확장 단어를 적용한 언어모델이며, P(Q|D)는 원래 질의를 적용한 언어모델, P(w|D)는 확장 단어를 적용한 언어모델이다.

3. 실험 및 평가

실험 문서 집합은 웹문서 테스트컬렉션인 TREC

WT10g 를 사용하였다. 실험을 통해 파라미터를 추정하고 성능을 평가한다. 실험 데이터 집합에 대한 정보는 <표 1>에서 보여준다.

본 연구에서는 전체 질의 개수 100 개(학습 질의 50 개, 테스트 질의 50 개) 중에서 질의가 3 개 이상인 경우에만 핵심질의 클러스터와 비핵심질의 클러스터로 분류하였다. 질의가 2 개 이하인 경우에는 핵심질의 조합이 1 개밖에 없어서 핵심질의를 구할 필요가 없기 때문이다.

<표 1> 실험 데이터 집합

문서 수	질의 개수	
	학습질의	테스트질의
1, 692, 096	21	29

언어모델(LM)과 적합모델(RM)에 대한 실험결과는 인드리(Indri-2.8)시스템[12]을 사용하였다.

수식(2)에서 단어의 초기 가중치($\alpha \in \{0.1, 0.15, 0.2, \dots, 0.9, 0.95\}$), 수식(3)에서 파라미터 값($\alpha \in \{0.1, 0.15, 0.2, \dots, 0.9, 0.95\}$, $\beta \in \{0.1, 0.15, 0.2, \dots, 0.9, 0.95\}$)는 실험에서 가장 좋은 성능을 보인 값으로 선택했다. 피드백 문서의 개수($n \in \{5, 10, 25, 50, 75, 100\}$), 확장 어휘의 개수($e \in \{5, 10, 20, 50, 75, 100\}$), 초기 질의에 대한 가중치($\lambda \in \{0.1, 0.2, \dots, 0.8, 0.9\}$)로 실험하였다.

제안된 방법과 적합모델을 비교하여 성능을 평가한다. 평가의 척도는 평균정확률의 평균인 MAP(Mean Average Precision)이다. <표 2>에서 LM 은 언어모델을 나타내고, RM 은 언어모델을 기반으로 질의를 확장한 적합모델을 나타낸다.

<표 2> 비교 실험 결과

LM (언어모델)	RM (적합모델)	제안 방법
0.2028	0.2143	0.2219
(-)	(+5.67%)	(+9.41%)

학습 질의에 대하여 실험한 결과, $\alpha = 0.95$, $\beta = 0.1$, 확장 어휘의 개수(e)는 5 개와 10 개일 때가 가장 좋았으며, 이 파라미터 값을 이용하여 테스트질의에 대하여 실험을 하였다.

그 결과, <표 2>에서와 같이 적합모델이 언어모델보다 5.67%의 성능이 향상되었으며, 제안 방법이 언어모델보다 9.41%의 성능이 향상되었음을 알 수 있었다.

4. 결론

본 논문에서는 핵심질의 클러스터와 비핵심질의 클러스터를 어휘 그래프로 표현하고, 핵심질의 클러스터의 어휘 그래프의 값에 비핵심질의 클러스터의 어휘 그래프의 값을 빼서 가중치를 재조정해 질의를 확장하는 기법에 대해 제안하였다. 실험을 통해, 제안 방법이 언어모델보다 9.41%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.

참고 문헌

[1] 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.

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

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

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

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

[6] Lv, Y., Zhai, C.X. 2010. Positional Relevance Model for Pseudo-Relevance Feedback. In Proc. of 33rd ACM SIGIR on Research and Development in Information Retrieval.

[7] Mei, Q., Zhang, D., Zhai, C.X., 2008. A General Oprimization FrameWork for Smoothing Language Models on Graph Structures. In Proc. of 31st ACM SIGIR on Research and Development in Information Retrieval.

[8] Huang, Y., Sun, L., Nie, J.Y., 2009. Smoothing Document Language Model with Local Word Graph. In Proc. of 18th ACM Conference on Information and Knowledge Management.

[9] 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.

[10] 장계훈, 김설영, 이경순. 2010. 핵심질의 어휘와 근접도를 이용한 핵심 문서 선택 기법. 제 33 회 한국정보처리학회 춘계학술발표대회.

[11] 장계훈, 조승현, 이경순. 2010. 단어 근접도를 반영한 단어 그래프 기반 질의 확장. 제 34 회 한국정보처리학회 추계학술발표대회.

[12] 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