질의의 위치와 문맥을 반영한 클러스터 기반 재순위화

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

Reranking Clusters based on Query Term Position and Context

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

요 약

질의와 질의 주변에 나오는 어휘는 의미적으로 연관되어있다는 가정하에 질의뿐만 아니라 질의 주변에 나오는 문맥 어휘들도 가중치를 높여준다면 검색에 효율을 높일 수 있을 것이다. 본 논문에서는 질의와 질의 주변에 나오는 문맥 어휘들에게 가중치를 주어 질의 어휘의 위치 가중치를 반영한 문서를 표현하고, 위치 가중치가 반영된 문서 벡터들 사이의 유사도를 계산하여 클러스터 기반 재순위화를 하여 성능을 향상시키는 방법을 제안한다. 뉴스 집합인 TREC AP 문서를 이용하여 언어모델, 위치 가중치를 이용한 언어모델, 클러스터 기반 재순위화 모델의 비교실험을 통해 유효성을 검증한다.

1. 서론

사용자가 정보를 검색하기 위해 사용하는 질의가 출현하는 문서에서 질의 주변에 나타나는 문맥 어휘들은 질의와 연관되어 있으며, 질의와 가까이 있을수록 두 어휘 사이는 더 큰 연관성을 가지게 된다. 그렇기 때문에, 질의 주변의 문맥 어휘들은 질의와 함께 사용자가 원하는 정보를 찾는 데에 많은 도움을 주며, 특히, 문맥 어휘가 질의와 가까이 있을수록 더 많은 도움을 준다.

질의 사이의 거리를 그래프로 표현한 연구[1]에서는 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[2]에서는 어휘의 사이의 최소거리, 평균, 어휘 사이의 거리의 합 등을 이용하여 가중치를 결정한 모델링을 통해 성능이 향상됨을 보였다.

클러스터를 이용하여 재순위화하는 기법[3]은 벡터 공간검색 모델에서 성공적인 결과를 보였다. 또한, 클러스터 기반 언어모델 검색 기법[4]은 질의를 생성하는 확률로 클러스터를 순위화한 것으로 질의확률검색 모델에 비해 성능향상을 보였다.

본 연구에서는 (1) 질의 주변의 어휘들은 질의 문맥을 반영하기 때문에, 사용자가 원하는 정보를 찾는 데에 도움을 준다. (2) 질의 주변의 문맥 어휘들이 질의 어휘와 가까이 나타날수록 사용자에게 더 많은 정보를 줄 것이라 가정하고, 문서 유사도 계산에서 질의의 위치와 문맥의 가중치를 반영하기 위하여 질의 어휘와 질의 주변에 나타나는 문맥 어휘들의 근접정도에 따라 그래프를 이용하여 가중치를 부여한 후 문서 벡터표현을 한다. 위치 가중치를 반영한 문서들의 유사도를 계산한 후 클러스터 기반 재순위화를 한다.

2. 질의의 위치와 문맥 가중치를 반영한 문서표현

질의의 위치와 문맥 가중치를 반영한 문서표현 방법은 본 연구의 첫 번째 단계로 질의와 질의 주변 문맥 어휘와의 근접도를 구하기 위해 사용한 기법이다. 문서에는 질의와 질의 주변 문맥 어휘가 있고 둘 사이의 근접도를 구함으로서 질의 주변 문맥 어휘가 질의에 얼마나 영향을 미치는지 검증한다.

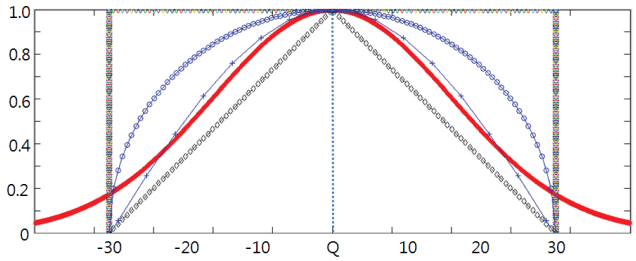
질의와 질의 주변 문맥 어휘와의 근접도를 구할 때, 거리에 따라 어느 정도의 가중치를 주어야 하는지 결정해야 한다. 이 때, 가중치는 그래프를 이용하여 결정할 수 있다.

(그림 1) 에서는 그래프를 이용하여 가중치를 결정하는 방법을 보여준다. 그래프의 가장 높은 부분이 질의이며, 그래프의 옆으로 갈수록 질의 주변 어휘에 적용되는 가중치가 적어진다는 것을 알 수 있다.

그래프에 따라 질의 주변 문맥 어휘에 적용되는 가중치가 다르게 적용된다. 가중치는 가우시안 그래프, 삼각형 그래프, 코사인 그래프, 원 그래프로 표현한다[1].

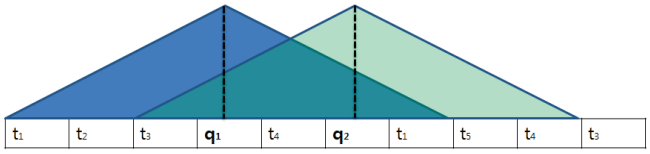
질의와 문맥 어휘 사이의 거리가 일정 거리 이하일 경우($|Q-d| \le \sigma$)에는 가중치가 적용되며, 일정 거

리를 넘어가게 되면 가중치가 적용되지 않는다.



(그림 1) 질의와 질의 주변 문맥 어휘의 가중치 변화 그래프($|Q-d| \le 30$)

(그림 2)는 한 문서 내에서 질의 주변에 나타나는 문맥 어휘들을 삼각형 가중치를 이용하여 가중치를 적용하는 예를 보여준다.



(그림 2) 삼각형 가중치($|Q-d| \le 3$)를 이용하여 주변 문맥 어휘에 가중치를 적용하는 예

t_1	t_2	t_3	q_1	t_4	q_2	t_1	t_5	t_4	t_3
1	1	1	**1**	1	**1**	1	1	1	1

↓

t_1	t_2	t_3	q_1	t_4	q_2	t_1	t_5	t_4	t_3
1	1.333	1.667	**2.333**	2.333	**2.333**	1.667	1.333	1	1

(그림 3) 위치 가중치를 반영했을 때의 어휘와 주변 문맥 어휘의 tf 벡터 변화 예

(그림 3)은 삼각형 가중치를 이용하여 위치 가중치를 반영했을 때의 주변 문맥 어휘의 tf 벡터 변화를 나타낸다. 어휘 t_2 에 적용되는 가중치를 구해보면 질의 q_1 과의 거리는 2 이므로 triangle($|Q-d|$) = 0.333 의 가중치가 적용된다. 또한, 어휘 t_2 와 질의 q_2 와의 거리는 4 이므로 질의 q_2 에 대한 어휘 t_2 의 가중치는 적용되지 않는다. 따라서 어휘 t_2 의 가중치는 1 + 0.333 = 1.333 이 된다.

질의의 위치 정보를 반영한 문서 벡터표현은 다음 세가지 방법으로 표현한다.

- **위치 가중치를 질의 어휘에만 적용하는 방법. pLM(Q) :** 질의의 주변 문맥 어휘도 사용자가 찾는 정보에 대하여 도움을 주긴 하지만 실제 사용자가 찾고자 하는 것은 질의다. 따라서, 질의 어휘의 가중치는 증가시키되, 주변 문맥 어휘는 가중치를 증가시키지 않는다.

- **위치 가중치를 질의 어휘와 주변 문맥 어휘에 동등하게 적용하는 방법. pLM(Q&Context) :** 입력한 질의의 주변 문맥 어휘도 사용자가 원하는 정보를 찾는 데 도움을 준다. 따라서, 질의 어휘와 질의 주변 문맥 어휘에 가중치를 동등하게 준다.

- **위치 가중치를 질의 어휘와 주변 문맥 어휘에 다르게 적용하는 방법. pLM(Q& $\alpha \cdot$ Context) :** 질의의 주변 문맥 어휘가 찾는 정보에 대하여 더 자세한 정보를 줄 수 있지만, 질의는 주변 문맥 어휘보다 더 중요한 정보일 것이다. 따라서, 질의 어휘와 질의 주변 문맥 어휘에 모두 가중치를 주되, 질의 주변 문맥 어휘에 주는 가중치의 증가량은 질의 어휘의 α (0< α <1)배에 해당하도록 한다.

언어 모델(Language Model)[5]은 다음과 같다.

$$P(Q \mid D) = \prod_{i=1}^{m} P(q_i \mid D) \quad (1)$$

여기서 q_i 는 i 번째 질의, m 은 질의 Q 의 어휘 개수이며, D 는 문서 모델을 나타낸다.

문서에 나타나지 않은 질의에 대해 0 이 아닌 값으로 추정하는데 사용하기 위하여 스무딩을 이용하는데, 다음은 디레슈레 스무딩(dirichlet smoothing)을 적용한 식이다.

$$P(w \mid D) = \frac{|D|}{|D|+\mu} \cdot P_{ML}(w \mid D) + \frac{\mu}{|D|+\mu} \cdot P_{ML}(w \mid Coll) \quad (2)$$

$$P_{ML}(w \mid D) = \frac{freq(w, D)}{|D|}, \quad P_{ML}(w \mid Coll) = \frac{freq(w, Coll)}{|Coll|} \quad (3)$$

여기서 $P_{ML}(w|D)$은 문서 D 에서의 어휘 w 의 최대확률추정을 나타내고, $Coll$ 은 전체 문서 집합, μ 는 스무딩 파라미터를 나타낸다. $|D|$는 문서 D 의 길이, $|Coll|$은 전체문서집합의 길이를 나타낸다. $freq(w, D)$는 문서 D 에서의 어휘 w 의 빈도수, $freq(w, Coll)$은 전체 문서집합에서의 어휘 w 의 빈도수를 의미한다.

본 연구에서는 문서에 나타나는 어휘들을 위치 가중치에 따라 변화시킴으로써 언어모델에 변화를 주고 언어모델과 비교하여 유효성을 검증할 수 있다.

3. 위치 정보와 문맥 정보를 반영한 클러스터 기반 재순위화

검색된 N 개의 문서들에 대하여 질의의 위치 가중치를 이용한 문서표현을 한 후, 문서 유사도 계산을 하는데, 본 연구에서는 위치 가중치가 적용된 tf' 를 이용하여 $tf' \cdot idf$ 로 계산한 후 코사인 정규화를 하여 표현한다. 위치 가중치가 반영된 문서들 사이의 유사도를 모두 계산한 후, 각 문서에 대해서 유사도가 높은 k 개의 가장 가까운 문서를 선택해서 그 문서에 대한 클러스터를 형성하는 방법인 최근접이웃(k-Nearest Neighbors; k-NN) 클러스터링 방법을 이용하여

클러스터링한다. 클러스터를 생성한 후, 클러스터 기반 재순위화를 한다.

클러스터 기반 언어모델([3], [4]) 에서 클러스터는 자신의 멤버로 속한 모든 문서를 연결해서 하나의 큰 문서처럼 표현한 후 언어모델에 적용한다.

$$P(Q \mid Clu) = \prod_{i=1}^{m} P(q_i \mid Clu) \quad (4)$$

$$P(w \mid Clu) = \frac{|Clu|}{|Clu|+\mu} \cdot P_{ML}(w \mid Clu) + \frac{\mu}{|Clu|+\mu} \cdot P_{ML}(w \mid Coll) \quad (5)$$

$$P_{ML}(w \mid Clu) = \frac{freq(w, Clu)}{|Clu|},\ P_{ML}(w \mid Coll) = \frac{freq(w, Coll)}{|Coll|} \quad (6)$$

여기서 *Clu* 는 클러스터, *freq(w, Clu)*는 클러스터 *Clu* 에 속하는 문서 *D* 의 *freq(w, D)*를 합한 것을 의미한다.

질의의 위치정보가 반영된 상위 N 개의 문서들에 대하여 *k*-NN 클러스터링 기법으로 생성된 클러스터와 문서에 대한 LM 결과를 결합하여, 문서를 재순위화 한다.

$$P'(Q \mid D) = \lambda \cdot P(Q \mid D) + (1-\lambda) \cdot MAX_{D \in Clu_i} P(Q \mid Clu_i) \quad (7)$$

여기서 문서는 여러 클러스터의 멤버가 될 수 있다. 따라서, 문서 *D* 가 속하는 클러스터 *Clu* 중에서 질의 확률을 최대로 갖는 값으로 선택한다. λ 는 질의의 위치정보가 반영된 문서의 질의확률을 얼마나 적용할 것인지 정한다. λ 가 클수록 질의의 위치정보가 반영된 문서의 질의확률을 더 많이 적용하게 되며, λ 가 작을수록 *k*-NN 기법으로 생성된 클러스터의 질의확률을 더 많이 적용하게 된다.

4. 실험 및 결과

실험 문서집합으로 뉴스기사 집합인 TREC AP(88-89)를 사용하였다. 질의 집합은 학습질의(51-100), 테스트 질의(151-200)를 사용하였다. 색인 및 검색은 인드리(Indri) 검색엔진[6]을 사용하였다.

언어모델에 대하여 질의 어휘의 위치 가중치가 문서 벡터에 반영된 문서표현을 한 후, 질의의 위치 가중치가 문서 벡터에 반영된 문서를 검색 모델을 이용하여 얻은 결과와 언어 모델을 비교한다.

질의와 어휘 사이의 거리는 실험 결과 25 이하일 경우 가중치를 주도록 하였다. 재순위화 하고자 하는 문서는 질의 당 1000 개이며, 스무딩 파라미터는 실험을 통해 μ = 2000 으로 설정하였다.

<표 1> 실험 데이터 집합

컬렉션	문서 수	학습 질의	테스트 질의
AP(88-89)	164,597	51-100(50 개)	151-200(50 개)

<표 2>에서 질의 어휘의 위치 가중치를 문서에 반영했을 경우 언어 모델보다 성능이 향상되었음을 알 수 있었다. 또한, 질의 어휘에만 위치 가중치를 반영하는 것보다 질의 어휘와 주변 어휘 모두에 위치 가중치를 반영했을 경우가 성능이 더 향상되었음을 알 수 있다.

<표 2> 언어모델과 질의 어휘의 위치 가중치를 문서에 반영하여 얻은 결과 비교

	원	코사인	가우시안	삼각형
LM	0.1574			
pLM(Q)	0.1605	0.1612	0.1603	0.1611
pLM(Q&Context)	0.1603	0.1616	0.1609	0.1614
pLM(Q& $\alpha \cdot$ Context)	0.1605	0.1613	0.1609	0.1612

비교실험 방법은 다음과 같다.

1) **언어모델(LM) :** 기준이 되는 검색 모델이다. 앞으로 실험하는 모델은 언어모델보다 높은 성능을 보이는지 비교해야 한다.

2) **위치 가중치를 반영한 언어모델(pLM) :** 본 논문에서 가정한 위치 가중치를 반영한 언어모델로서, 본 연구에서의 핵심 비교 실험 중 하나이다. 삼각형 가중치와 pLM(Q& $\alpha \cdot$ Context)을 이용하였다. α 의 값은 0.5 로 정하였다.

3) **언어모델을 클러스터 기반 재순위화한 모델(CBLM) :** 언어모델과 함께 기준이 되는 결과값이다. 위치 가중치를 반영한 문서를 클러스터 기반 재순위화한 모델이 이 결과보다 높은 성능을 보이는지 비교해야 한다.

4) **위치 가중치를 반영한 문서를 클러스터 기반 재순위화한 모델(pCBLM) :** pLM 을 클러스터 기반 재순위화한 모델로서, 본 연구에서의 핵심 비교 실험 중 하나이다. 이 결과가 높은 성능을 보임으로서 본 논문에서 제안한 방법이 유효한지 검증할 수 있다. pLM 에서 성능이 좋았던 코사인 가중치와 삼각형 가중치에 대하여 실험을 통해 비교한 결과, 성능이 좋았던 삼각형 가중치를 사용하였으며, 실험을 통해 재순위화 결과가 가장 좋았던 pLM(Q& $\alpha \cdot$ Context)를 이용하였다. α 의 값은 0.5 로 정하였다.

비교실험에서 *k*-NN 클러스터링을 위한 k 는 5 로 설정했으며, 재순위화에 적용되는 λ 는 실험을 통해 0.8 로 설정했다.

<표 3> 비교 실험 결과

	LM	pLM	CBLM	pCBLM
MAP	0.3105 (-)	0.3226 (+3.86%)	0.3297 (+6.18%)	0.3502 (+12.79%)
P@5	0.5680 (-)	0.5480 (-4.93%)	0.5760 (+1.41%)	0.5880 (+3.52%)
P@10	0.5080 (-)	0.5200 (+2.36%)	0.5280 (+3.94%)	0.5420 (+6.69%)

<표 3>에서 보는 바와 같이 질의의 위치 가중치를

반영한 언어모델(pLM)이 기존의 언어모델(LM)보다 성능이 향상되었음을 알 수 있다. 클러스터 기반 재순위화(CBLM)는 기존의 언어모델보다 성능이 향상되었음을 알 수 있으며, 기존의 언어모델을 클러스터 기반 재순위화하는 것보다 질의 어휘의 위치 가중치를 반영한 문서를 클러스터 기반 재순위화(pCBLM)했을 경우에 더 높은 성능을 얻을 수 있다는 것을 알 수 있다.

5. 결론

본 연구에서는 문서 내 질의의 주변에서 나오는 문맥 어휘는 사용자가 정보 검색을 하는 데 영향을 준다고 가정하여 질의의 위치 가중치를 이용하여 문서 표현을 하고, 클러스터를 이용하여 재순위화하는 방법을 제안하였다.

언어모델과 질의의 위치정보를 반영한 언어모델의 결과를 비교하였을 때, 질의의 위치정보를 반영한 언어 모델이 기존의 언어모델에 비해 성능이 향상되었다. 또한, 클러스터를 이용한 재순위화에서도 기존의 언어모델을 재순위화한 경우보다 질의의 위치정보를 반영한 언어모델을 재순위화한 경우가 더 성능이 더 많이 향상되었다. 이를 통해, 제안된 알고리즘이 유효함을 알 수 있다.

참고문헌

[1] 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.

[2] Zhao, J.G. and Yun, Y. 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.

[3] Lee, K.S., Park, Y. C. and Choi, K.S., Reranking Model based on Document Clusters, Information Processing & Management, 37(1), pp.1-14, 2001.

[4] Liu, X. Y. and Croft, W. B., Evaluation Text Representations for Retrieval of the Best Group of Documents, In Proc.30th ECIR Conf, pp.454-462, 2008.

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

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

[7] Diaz, F., and Metzler, D. Improving the Estimation of Relevance Models Using Large External Corpora, In Proc. 29th ACM SIGIR Conf on Reserch and Development int Information Retrieval.pp.154-161, 2006.