단어 근접도를 반영한 단어 그래프 기반 질의 확장

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

Query Expansion based on Word Graph using Term Proximity

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

요 약

질의 확장은 초기 검색결과에서 질의와 연관된 단어를 선택하여 질의를 확장함으로써 검색 성능을 향상시키는 기법이다. 페이지 랭크(PageRank) 알고리즘은 웹문서 사이의 링크구조를 이용하여 문서들의 상대적인 중요성을 측정하기 위해 제안되었다. 본 논문에서는 문서들 사이의 관계가 아니라 문서 안에서 단어 그래프(Word Graph)를 통해 단어들 사이의 상대적인 중요성을 계산하였다. 질의와 가까이 위치한 단어들 사이의 관계를 단어 그래프에 적용하여 중요도를 계산하고 확장단어를 선택한다. 본 논문의 유효성을 검증하기 위해 웹문서 집합인 TREC WT10g 에 대해 실험하였고, 적합모델(Relevance Model)보다 MAP(Mean Average Precision)가 4.1% 향상되었다.

1. 서론

정보의 양이 늘어날수록 하나의 질의에 의해 다양한 주제의 문서가 검색되어 사용자가 필요로 하지 않는 정보까지 검색될 수 있다. 이것은 질의 어휘가 표현하는 의미가 모호하기 때문에 발생하는 것으로 사용자가 원하는 정보에 맞는 정확한 질의를 선택하는 것은 쉬운 일이 아니다. 질의 확장[1]은 이런 질의 어휘 선택의 어려움을 해결할 수 있는 방법 중 하나이다.

질의를 확장하는 연구로는 피드백 문서 안에서 단어의 위치를 적합모델에 적용시킨 연구[2, 3]가 있다. 단어의 위치를 코사인, 가우시안 등 함수의 그래프에 따라서 질의 어휘 사이의 거리 가중치를 적용하여 확장단어를 선택한다. 질의와 가까운 위치에 자주 발생할수록 더 많은 가중치를 받게 되고 가중치가 높은 단어가 확장단어로 선택된다.

본 논문은 질의 어휘와 단어 사이의 거리를 단어 그래프(Word Graph)[4, 5]에 적용하여 질의를 확장한다. 기본적으로 그래프 기반 랭킹 알고리즘은 노드 사이의 링크가 다른 노드를 추천하는 하나의 표로 해석하여 이를 기준으로 중요도를 평가한다. 페이지 랭크(PageRank)[6]는 정확률을 높이기 위해 관련성이 높고 권위 있는 페이지를 상위에 랭크 시키는 것에 중점을 두고 있으며, 웹문서 사이의 링크구조를 통해 문서들을 순위화한다. 구글 검색기에서는 페이지 랭크를 사용하여 많은 사람들이 추천하는 권위 있는 문서를 상위 결과로 제시하여 좋은 성능을 보였다.

페이지 랭크 알고리즘의 개념을 텍스트 문서에 대해 적용한 텍스트 랭크(TextRank) 알고리즘[7]은 페이지 랭크에서 웹페이지 사이의 링크구조를 단어들 사이의 그래프로 생각하고 중요도를 계산한다. 랜덤 워크 알고리즘(Random-Walk Algorithms)[8] 역시 단어 그래프를 기반으로 문서 안에 포함된 단어와 단어 사이의 관계를 표현한다. 두 단어의 공기빈도를 에지(edge)의 가중치로 하여 그래프를 표현한다.

단어 그래프는 단어 사이의 연결 관계를 이용하여 가중치를 결정하기 때문에 단어의 빈도에 의존하는 기존의 질의 확장을 대체할 수 있는 방법 중 하나이다.

본 논문의 가정 및 접근방법은 다음과 같다: (ⅰ) 질의와 근접하게 발생한 단어는 질의와 의미적으로 연관되어 있다는 가정하에 단어 그래프에서 두 노드 사이의 에지(edge)가중치를 이용하여 단어와 질의 어휘 사이의 근접도를 적용한다. (ⅱ) 주변에 비슷한 단어가 나타나는 두 단어는 서로 의미적으로 연관되어 있다는 가정하에 전체 피드백 문서에서 각 단어들의 주변 문맥 단어를 찾아서 단어들 사이의 유사도를 구하고 그 값을 노드 사이의 가중치로 하여 단어 그래프에 적용한다.

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

2. 단어 그래프를 이용한 단어의 가중치 결정

단어를 확장하기 위해 초기 검색결과의 상위문서에 있는 모든 단어들의 가중치를 결정하고, 가중치가 가장 높은 단어를 확장단어로 선택한다. 가중치 결정은 각 단어와 질의 어휘들 사이의 근접도를 적용한 단어 그래프를 이용하며, G=(V, E)로 표현할 수 있다. V 는 그래프의 노드로써 문서에서 각 단어를 의미하며, E 는 노드 사이의 에지(edge)로써 질의 어휘와의 근접도를 에지의 가중치로 한다. 아래 식 (1)을 통해 피드백 문서 안에 포함된 각 단어의 가중치를 계산할 수 있다.

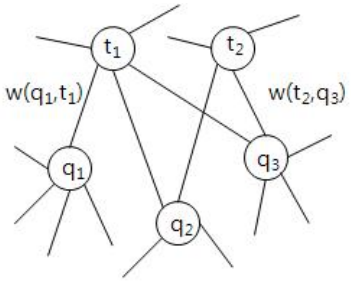
$$f^{r+1}(t_i) = \alpha \times f^0(t_i) + (1-\alpha) \times \sum_{j=1}^{k} \sum_{q_j \in Near(t_i)} \frac{w(t_i, q_j) \times f^r(q_j)}{\sum_{t_k \in Near(q_j)} w(q_j, t_k)} \quad (1)$$

여기서 $f^0(t_i)$는 t_i 의 초기 가중치이고, k 는 질의 어휘의 수, $w(t_i, q_j)$는 t_i 와 q_j 사이의 근접도, $Near(t_i)$는 문서 안에서 t_i 와 근접하게 나타난 단어들이다. 단어의 가중치가 일정한 값에 수렴할 때까지 반복적으로 계산한다. 수렴하기 위한 임계치($c = f^{r+1}(t_i) - f^r(t_i)$)는 10^{-6} 으로 하였다.

식 (1)에서 $f^0(t_i)$는 적합모델(Relevance Model)을 이용하여 계산한다. 적합모델은 언어모델(Language Model)[10]을 기반한 질의 확장 기법으로 질의 Q 가 주어졌을 때 어휘 w 의 확률을 추정하는 다항분포이다.

$$f^o(t_i) = \sum_{D \in R} P(D)P(t_i \mid D)P(Q \mid D) \quad (2)$$

R 은 질의 Q 에 대해 잠정적으로 적합하다고 가정한 문서들의 집합이다. P(D)는 문서가 발생할 확률이므로 모든 값에 균일하게 적용된다. $P(t_i|D)$는 문서에서 단어 t_i 가 발생할 확률, P(Q|D)는 초기 질의에 대한 문서의 중요도를 의미한다.



(그림 1) 질의 어휘와 근접도를 적용한 단어 그래프

그림 1 에서 각 노드(node)는 단어를 의미하며, q_1, q_2, q_3 은 질의 어휘를 나타낸다. 에지(edge)는 주변 단어들과의 근접도를 의미한다. 식 (1)에서 $w(t_i, q_j)$는 그림에서 에지의 가중치를 의미하며 한 노드에 연결되어 있는 모든 에지들의 가중치를 더하면 1 이 된다. 식 (1)의 뒷부분에 $w(q_j, t_k)$는 q_j 와 근접해있는 모든 단어들로 그림에서 q_1 과 근접해있는 다섯 개의 단어와 에지를 의미하며 $w(t_1, q_1)$는 그림에서 q_1 과 t_1 사이를 연결한 에지를 의미한다. t_1 은 q_1, q_2, q_3 모두 근접해 있으므로 모든 질의 어휘들의 가중치를 받는다.

본 논문에서는 $w(t_i, q_j)$의 값을 두 가지 방법으로 계산한다.

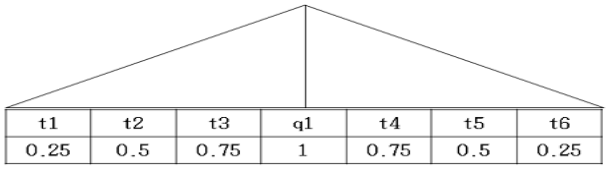
1) 질의 근접도를 반영

질의와 근접한 단어는 질의와 의미적으로 연관되어 있다. 질의와의 거리를 에지의 가중치로 하여 단어 그래프에 적용한다.

$$w(t_i, q_j) = \sum_{D \in fbDocs, t_i \in Near(q_j)} prox(t_i, q_j) \quad (3)$$

여기서 fbDocs 는 단어를 확장하기 위해 피드백에 사용할 초기 검색결과의 상위 문서이고, $Prox(t_i, q_j)$는 문서 안에서 t_i 와 q_j 의 근접도이다.

그림 2 에서 가중치 적용 거리 파라미터 δ 가 4 라고 한다면, q_1 을 중심으로 각 주변 단어들의 가중치를 그림과 같이 적용할 수 있다. q_1 은 4/4 만큼의 가중치를 받고, q_1 옆에 있는 t_3 과 t_4 는 3/4 만큼, 그 옆에 t_2 와 t_5 는 2/4 만큼, t_1 과 t_6 은 1/4 만큼 가중치를 받는다. 질의 어휘에서 멀어질수록 가중치는 덜 받게 된다.



(그림 2) 단어와 질의 어휘 사이의 근접도 계산 방법

2) 문맥어휘의 유사도와 질의 근접도를 반영

어떤 두 개의 단어가 비슷한 문맥어휘를 공유하고 있으면 두 단어는 의미적으로 연관되어 있다. 두 단어의 문맥어휘에 대한 유사도를 계산함으로써 비슷한 문맥어휘를 공유하는지 계산한다.

$$w(t_i, q_j) = \sum_{D \in fbDocs} (contextSim(t_i, q_j) + prox(t_i, q_j)) \quad (4)$$

여기서 $contextSim(t_i, q_j)$은 두 단어의 문맥어휘 사이의 유사도이다.

단어 주변에 빈번하게 나타나는 단어들을 그 단어의 문맥어휘라 한다. 식 (5)와 같이 문맥어휘는 피드백 문서 전체에서 구한다.

$$Context(t_i) = \sum_{D \in fbDocs} cooc(t_i, t_j) \quad (5)$$

여기서 $cooc(t_i, t_j)$는 t_i 와 일정한 거리 안에 발생한 단어 t_j 의 빈도이다.

그림 2 에서 q_1 의 주변에는 $t_1, t_2, \ldots, t_5, t_6$ 이 있으며, 이 단어들은 q_1 의 문맥어휘이다. q_1 과 가까울수록 문맥 가중치를 많이 받게 된다. 이렇게 모든 피드백 문

서에서 q_l 주변에 나온 단어들의 문맥 가중치를 합하면 q_l 의 문맥어휘가 하나의 문서처럼 의미가 연관된 단어들의 집합이 만들어진다. 모든 단어에 대해서 문맥어휘를 구한 후, 각 단어들의 문맥어휘를 하나의 문서로 생각하고, 두 단어들 사이의 코사인 유사도를 구한다. contextSim(t_i, q_j)는 전체 피드백 문서에서 구한 두 단어의 문맥어휘 사이의 유사도이기 때문에 하나의 문서 안에서는 두 단어가 어느 위치에 발생해도 값은 같다. 식 (3)에서 prox(t_i, q_j)는 하나의 문서 안에서 두 단어의 거리이기 때문에 이 두 값을 더해주면 전체 피드백 문서에서 가중치와 한 문서에서 가중치를 모두 적용할 수 있다.

제안된 두 가지 방법을 통해 $f^{r+1}(t_i)$의 가중치가 높은 상위의 fbTerms 개의 단어를 확장단어로 선택하여 식 (6)의 RM3(Relevance Model number 3) 공식을 통해 문서의 중요도를 결정한다.

$$P'(Q \mid D) = \lambda \cdot P(Q \mid D) + (1-\lambda)P(w \mid D) \qquad (6)$$

여기서 P'(Q|D)는 확장단어를 적용한 문서의 중요도이고, P(Q|D)는 원래 질의를 적용한 문서의 가중치, P(w|D)는 확장단어를 적용한 문서의 가중치이다.

3. 실험 및 평가

실험 문서 집합은 웹문서 집합인 TREC WT10g 를 사용하였다. 학습 질의를 통해 파라미터를 추정하고 테스트 질의에 대해 성능을 평가했다. 테스트 컬렉션에 대한 정보는 표 1 에서 보여준다.

<표 1> 실험 데이터 집합

컬렉션	문서 수	학습 질의		테스트 질의	
		질의 번호	개수	질의 번호	개수
WT10g	1,692,096	451-500	50	501-550	50

언어모델과(LM)과 적합모델(RM)에 대한 실험 결과는 인드리(Indri-2.8) 시스템[11]을 사용하였다. 언어모델의 수식은 다음과 같다.

$$P(Q \mid D) = \prod_{i=1}^{k} \left(\frac{|D|}{|D|+\mu} \cdot \frac{f_{q_i,D}}{|D|} + \frac{\mu}{|D|+\mu} \cdot \frac{c_{q_i}}{|C|} \right) \qquad (7)$$

여기서 k 는 질의 어휘의 개수이고, |D|는 문서의 길이, |C|는 전체 컬렉션의 길이, $f_{qi,D}$ 는 문서 D 에서의 질의 어휘 q_i 의 빈도수, C_{qi} 는 전체 컬렉션에서의 q_i 의 빈도수를 나타낸다. μ는 디리슈레 스무딩(Dirichlet smoothing) 파라미터로 μ값은 학습질의에 대한 실험 (μ∈{500, 1000, 1500, 2000, … ,5000})에서 MAP 가 가장 높은 값을 보인 2000 으로 설정하였다.

피드백 문서의 개수(fbDocs∈{5, 10, 25, 50, 75, 100}), 확장 어휘의 개수(fbTerms∈{5, 10, 25, 50, 75, 100}), 원래 질의에 대한 가중치(λ∈ {0.1, 0.2, … , 0.9})로 실험하였다. 식 (1)에서 단어의 초기 가중치 (α∈{0.1, 0.2, … , 0.9}), 식 (3)에서 거리 가중치를 적용하기 위한 파라미터(δ∈{5, 10, 25, 50, 74, 100})는 훈련집합에서 가장 좋은 성능을 보인 값으로 선택했다.

제안된 방법과 적합모델을 비교하여 성능을 평가한다. 평가의 척도는 MAP 이다. 표 2 에서 LM 은 단어를 확장하지 않은 언어모델을 나타내고, RM 은 언어모델을 기반으로 단어를 확장한 적합모델을 나타낸다.

<표 2> 비교 실험 결과

컬렉션	LM	RM	질의 근접도 적용	문맥어휘의 유사도 적용
WT10g	0.2125	0.2171	0.2261 (+4.1%)	0.2256 (+3.9%)

표 2 에서와 같이 실험 결과는 질의 근접도를 적용한 방법에서 4.1%의 성능이 향상되었고 문맥어휘의 유사도를 적용한 방법에서 3.9%의 성능이 향상되었다.

4. 결론

본 논문에서는 질의 어휘와의 근접도를 적용한 단어 그래프를 이용하여 질의 확장단어를 선택하는 기법에 대해 제안하였다. 질의와 근접하게 나타난 단어들이 질의와 의미적으로 연관되어 있음을 확인할 수 있었고, 질의와 근접도를 적용한 단어 그래프를 이용해 적합모델보다 MAP 가 4.1% 향상됨을 보였다. 이것을 통해 웹문서 사이의 링크관계를 통해 상대적인 가중치를 결정하는 페이지 랭크 알고리즘을 문서 안에 포함된 단어들의 그래프에 적용할 수 있음을 확인했고, 특히 질의 확장단어를 선택하는데 유효함을 확인했다.

참고 문헌

[1] Kalmanovich, I.G., Kurland, O. 2009. Cluster-Based Query Expansion. In Proc. of 32nd ACM SIGIR on Research and Development in Information Retrieval. pp.646-647.

[2] 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

[3] 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.

[4] 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.

[5] 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.

[6] Page, L., Brin, S., Motowani, R. and Winograd, T. 1998. The PageRank Citation Ranking: Bringing Order to the Web, Unpublished manuscript, Stanford University.
[7] Mihalcea, R., Tarau, P., 2004. TextRank-bringing order into texts. In Proc. of the Conference on Empirical Methods in Natural Language Processing(EMNLP 2004)
[8] Blanco, R., Lioma, C. 2007. Random Walk Term Weighting for Information. In Proc. of 30th ACM SIGIR on Research and Development in Information Retrieval.
[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] 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