http://dx.doi.org/10.3745/KIPSTB.2012.19B.1.037

질의 어휘와의 근접도를 반영한 단어 그래프 기반 질의 확장

장 계 훈† · 이 경 순††

요 약

잠정적 적합성 피드백모델은 초기 검색 결과의 상위에 순위화된 문서를 적합 문서라 가정하고, 상위문서에서 빈도가 높은 어휘를 확장 질의로 선택한다. 빈도수를 이용한 질의 확장 방법의 단점은 문서 안에서 포함된 어휘들 사이의 근접도에 상관없이 각 어휘를 독립적으로 생각한다는 것이다. 본 논문에서는 어휘빈도를 이용한 질의 확장을 대체할 수 있는 어휘 근접도를 반영한 단어 그래프 기반 질의 확장을 제안한다. 질의 어휘 주변에 발생한 어휘들을 노드로 표현하고, 어휘들 사이의 근접도를 에지의 가중치로 하여 단어 그래프를 표현한다. 반복된 연산을 통해 확장 질의를 선택함으로써 성능을 향상시키는 기법을 제안한다. 유효성 검증을 위해 웹문서 집합인 TREC WT10g 테스트 컬렉션에 대한 실험에서 언어모델 보다 MAP 평가 기준에서 6.4% 향상됨을 보였다.

키워드 : 어휘 근접도, 단어 그래프, 문맥어휘, 질의 확장, 텍스트랭크 알고리즘

Query Expansion based on Word Graph using Term Proximity

Kye-Hun Jang† · Kyung-Soon Lee††

ABSTRACT

The pseudo relevance feedback suggests that frequent words at the top documents are related to initial query. However, the main drawback associated with the term frequency method is the fact that it relies on feature independence, and disregards any dependencies that may exist between words in the text. In this paper, we propose query expansion based on word graph using term proximity. It supplements term frequency method. On TREC WT10g test collection, experimental results in MAP(Mean Average Precision) show that the proposed method achieved 6.4% improvement over language model.

Keywords : Term Proximity, Word Graph, Context Term, Query Expansion, TextRank

1. 서 론

잠정적 적합 피드백 기법[1,2,3]은 초기 검색 결과의 상위 문서를 잠정적으로 적합한 문서라 가정하고, 상위문서에서 발생한 어휘들 중 질의와 연관되어 있는 어휘를 확장 질의로 선택한다. 전통적인 방법은 어휘빈도를 이용하는 것이다. 상위문서에서 가장 빈도가 높은 어휘를 질의와 연관된 어휘라 가정하고 확장 질의로 선택한다. 어휘빈도는 문서 안에서 자질들의 확률분포를 평가하기 위한 방법으로 오랫동안 사용되어 왔다. 하지만 어휘빈도를 이용한 방법은 문서 안에서 어휘들의 근접도에 상관없이 문서를 어휘들의 집합으로 표현함으로써 어휘들 사이의 의미적인 관계보다는 독립적인 특성들을 이용하여 어휘의 중요도를 결정한다.

문서 안에서 어휘의 위치를 적합모델에 적용시킨 연구[4,5]에서는 코사인, 가우시안, 삼각형 등 함수의 그래프에 따라서 어휘의 위치를 표현하고, 질의 어휘 사이의 거리가 중치를 적용함으로써 확장 질의를 선택한다. 질의 어휘와 가까운 위치에 자주 발생할수록 더 많은 가중치를 받게 되고 가중치가 높은 어휘가 확장 질의로 선택된다.

본 논문은 질의 어휘와 어휘 사이의 관계를 단어 그래프(Word Graph)[6,7,8,9]로 표현하고, 질의 어휘와 근접도[10]를 이용해 질의와 근접한 어휘들을 확장 질의로 선택한다. 단어 그래프를 이용함으로써 각 어휘들의 독립적인 특성이 아니라, 문서 안에서 전체적인 구조와 어휘들 사이의 관계를 반영하여 단어 그래프 안에서 반복적인 연산을 통해 각 어휘들의 가중치를 결정할 수 있다.

질의와 가까이 발생한 어휘는 질의와 의미적으로 연관되어 있다는 가정하에 단어 그래프에서 두 노드(node) 사이

† 준 회 원 : 전북대학교 컴퓨터공학과 석사
†† 정 회 원 : 전북대학교 컴퓨터공학부/영상정보신기술연구센터 부교수
논문접수 : 2011년 5월 18일
수 정 일 : 1차 2011년 7월 7일
심사완료 : 2011년 7월 7일

에지(edge)의 가중치를 어휘와 질의 어휘 사이의 근접도를 적용하여 계산한다. 제안된 방법의 유효성을 검증하기 위해 TREC WT10g 컬렉션에 대해 실험하고, 잠정적 적합 피드백모델에서 우수한 성능을 보인 적합모델(Relevance Model)[1]과 비교하여 성능을 평가한다.

본 논문의 구성은 2장에서 관련연구를 소개하고, 3장에서는 문서 안에서 어휘의 위치에 기반해 단어 그래프를 구성하고 근접도를 적용한 질의 확장 방법에 대해 설명하고, 4장에서 실험에 관한 정보를 5장에서는 실험에 대한 결론에 대해 논하겠다.

2. 관련 연구

2.1 질의 확장

적합모델은(Relevance Model)[1] 초기 검색 결과에서 적합한 문서를 이용해 확장 질의를 선택하고, 초기 질의에 적용함으로써 최근 질의 확장에 효율적인 방법으로 알려져 있다. 잠정적 적합 피드백(pseudo-relevance feedback)은 초기 검색 결과에서 상위문서들은 잠정적으로 질의에 적합한 문서라 가정하고 상위문서에서 빈도가 높은 어휘를 확장 질의로 선택한다. 다음과 같은 식에 의해 확장 어휘가 계산된다.

$$P(w \mid R) = \sum_{D \in R} P(D)P(w \mid D)P(Q \mid D) \qquad (1)$$

여기서 R은 초기 검색 결과 상위문서이고, P(D)는 전체 집합에서 균일하게 적용된다. P(w|D)는 문서에서 어휘가 발생할 확률을 나타내며, P(Q|D)는 언어모델을 통한 문서의 초기가중치를 나타낸다. 결국 P(w|D)와 P(Q|D)의 곱한 것을 피드백 문서 전체에 대해 더하여 값이 높은 순서대로 선택하게 된다. P(w|R)이 가장 높은 e개의 어휘를 질의 확장을 위해 선택한다.

최근에 초기 검색 결과의 상위문서를 그대로 피드백에 사용하지 않고 선택적으로 샘플링하여 피드백하는 연구[3]가 있다. 초기 검색 결과의 상위문서들은 비슷한 행태를 가지고 있다고 가정하면, 상위에 있는 문서를 피드백에 그대로 사용하면 비슷한 문서만을 가지고 피드백 하기 때문에 효율적이지 못하다. 따라서 피드백 문서를 선택할 때 어떤 질의 어휘 클러스터 안에 있는 문서의 개수가 임의의 개수가 넘으면 더 이상 그 어휘 조합이 발생한 문서가 나와도 클러스터에 포함시키지 않는 방법을 제안했다. 다양하고 새로운 문서집합을 피드백에 사용함으로써 성능을 개선시킨다.

2.2 그래프 기반 알고리즘

랜덤워크 알고리즘(random-walk)[11]은 가상의 리더(reader or walker)가 텍스트 안에서 어휘들 위에 임의로 걸어 다니는 동안 목적어휘(target term)를 만날 확률을 계산하여 어휘들의 가중치를 결정하는 알고리즘이다. 랜덤워크 알고리즘은 기존에 어휘의 순서에 상관없이 각 어휘들이 갖는 빈도 정보에 의해서 중요도를 결정하는 것이 아니라, 텍스트 안에 모든 어휘들 사이의 관계정보를 통해 반복적으로 연산한 결과를 고려하여 각 어휘의 중요도가 결정된다. 기본적인 아이디어는 노드 사이의 링크가 다른 노드를 추천하는 하나의 표(vote)로 해석하여 이를 기준으로 중요도를 평가한다. 추천을 많이 받은 노드는 중요도가 높고, 그 노드가 추천한 노드 역시 높은 중요도를 얻게 된다. 문서들 사이의 연결관계를 이용한 페이지랭크 알고리즘과 이를 텍스트 처리에 적용한 텍스트랭크 알고리즘이 랜덤워크 알고리즘을 그래프 기반 순위화 알고리즘에 적용한 방법이다.

페이지랭크(PageRank) 알고리즘[12]은 웹문서 검색의 정확률을 높이기 위해 관련성이 높고 권위있는 페이지를 상위에 순위화 시키는 것에 중점을 두고 있으며, 웹문서 사이의 링크구조를 통해 문서들을 순위화한다. G = (V, E)의 방향성 그래프로 표현할 수 있으며, V(Vertices)는 노드를 나타내고 E(Edges)는 에지를 나타낸다.

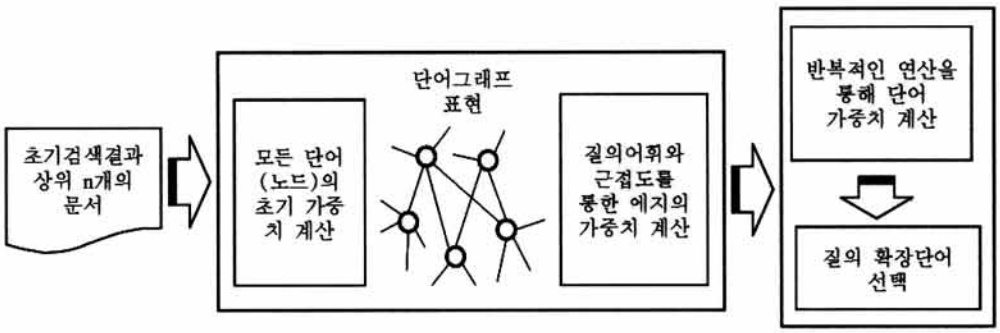
텍스트랭크 알고리즘[9]은 페이지랭크에서 웹페이지 사이의 링크구조를 어휘들 사이의 그래프로 생각하고 중요도를 계산한다. 한 문서에서 출현한 어휘나 문장 등을 노드로 간주하고 방향성이 없는 그래프를 생성한다. 텍스트랭크 알고리즘은 어휘의 중요도가 일정한 값으로 수렴될 때까지 반복적으로 연산을 수행한다. 공식은 식(2)와 같다.

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \qquad (2)$$

여기서 V_i는 그래프 상의 임의의 어휘고 V_j는 어휘 V_i와 인접한 어휘, V_k는 어휘 V_j와 인접한 어휘다. WS(Vi)는 현재 단계에서 어휘 V_i의 중요도이고, $WS(V_j)$는 이전 단계에서 어휘 V_j의 중요도이다. 초기 단계에서 $WS(V_j)$의 값은 1이다. $In(V_i)$는 어휘 V_i와 인접한 어휘의 집합, $Out(V_j)$는 어휘 V_j와 인접한 어휘의 집합이다. 그래프의 방향성이 없으므로 서로 인접한 어휘들은 들어오는 링크와 나가는 링크를 하나씩 갖는다. 또한 그래프에서 에지의 가중치가 없으므로, 가중치 w_{ji}, w_{jk}는 1의 값을 갖는다. d는 일반적으로 0.85로 설정한다.

페이지랭크에서는 웹문서가 가리키는 링크관계를 연결고리로 하여 그래프를 생성했지만 텍스트랭크에서는 한 문서 내에 가시적인 연결고리가 없으므로 어휘나 문장 사이를 연결할 가상의 연결고리가 필요하다. 텍스트랭크를 이용해 키워드를 추출하는 연구에서는 어휘의 공기빈도(co-occurrence)를 가상의 연결고리로 하여 사용하였고 중요문장 추출 연구[9]에서는 문장 사이의 유사도를 연결고리로 하여 사용했다.

본 논문에서는 질의와 의미적으로 연관된 어휘를 확장 질의로 선택하기 위해 질의와의 근접도를 반영한 단어 그래프를 이용한다. 텍스트랭크 알고리즘처럼 모든 어휘들 사이의



(그림 1) 어휘 근접도를 이용한 단어 그래프 기반 질의 확장 시스템 구조

관계를 그래프로 표현하는 것이 아니라 질의와 그 주변에 나타난 어휘들만 그래프로 표현하며, 적합모델(Relevance Model)을 이용해 어휘의 초기가중치를 적용했다. 또한 에지의 가중치는 질의와 그 주변어휘와의 거리를 위치기반 언어모델[4,5]에서 삼각형 그래프에 적용하여 거리가 멀어질수록 적게 적용한 방법을 통해 확장 질의를 선택한다.

3. 질의 어휘와의 근접도를 반영한 단어 그래프 기반 질의 확장

본 논문에서 제안하는 어휘 근접도를 반영한 단어 그래프 기반 질의 확장 기법의 전체적인 시스템 구조는 (그림 1)과 같다. 언어모델(Language Model)[13]에 의한 초기 검색 결과 상위문서에 대해서 적합모델(Relevance Model)을 이용하여 각 문서에서 발생한 모든 어휘들의 초기 가중치를 결정한다. 어휘들을 노드로 하고 어휘들 사이의 근접도를 에지의 가중치로 하여 질의 어휘와 근접한 어휘들을 단어 그래프를 이용하여 표현한다. 반복적인 연산을 통해 어휘의 가중치를 계산하고, 가중치가 높은 어휘를 확장 질의로 선택한다.

3.1 질의 어휘와의 근접도 반영을 위한 단어 그래프 구성

어휘를 확장하기 위해 초기 검색 결과의 상위문서에 있는 모든 어휘들의 가중치를 결정하고, 가중치가 가장 높은 어휘를 확장 어휘로 선택한다. 가중치 결정은 각 어휘와 질의 어휘들 사이의 근접도를 적용한 단어 그래프를 이용한다.

그래프는 G=(V, E)로 표현할 수 있으며, V는 그래프의 노드로써 문서에서 각 어휘를 의미한다. E는 노드 사이의 에지(edge)로써 질의 어휘와의 근접도를 에지의 가중치로 한다. 아래 식(3)을 통해 피드백 문서 안에서 포함된 각 어휘의 가중치를 계산할 수 있다.

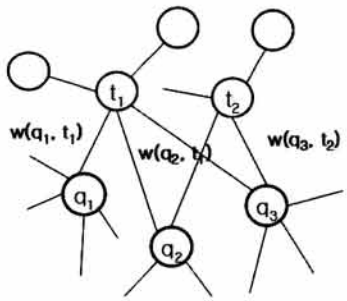
$$f^{r+1}(t_i) = \alpha \times f^0(t_i) + (1-\alpha) \times \sum_{q_j \in Near(t_i)} \frac{w(t_i, q_j) \times f^r(q_j)}{\sum_{t_k \in Near(q_j)} w(q_j, t_k)} \quad (3)$$

여기서 $f^0(t_i)$는 노드 t_i의 초기 가중치이고, 에지의 가중치인 $w(t_i, q_j)$는 t_i와 q_j사이의 근접도이다(식 (5)에서 계산). $Near(t_i)$는 문서 안에서 t_i와 근접하게 나타난 어휘들이다. 어휘의 가중치가 일정한 값에 수렴할 때까지 반복적으로 계산한다. 수렴하기 위한 임계치($c = f^{r+1}(t_i) - f^r(t_i)$)는 0.000001로 한다.

식 (3)에서 $f^0(t_i)$는 어휘의 초기값으로 적합모델(Relevance Model)을 이용하여 계산한다. 적합모델은 언어모델(Language Model)을 기반한 질의 확장 기법으로 질의 Q가 주어졌을 때 어휘 w의 확률을 추정하는 다항분포이다. 식(4)는 적합모델의 식을 보여준다.

$$f^0(t_i) = \sum_{D \in R} P(D)P(t_i \mid D)P(Q \mid D) \quad (4)$$

여기서, R은 질의 Q에 대해 잠정적으로 적합하다고 가정한 문서들의 집합이다. P(D)는 문서가 발생할 확률이므로 모든 값에 균일하게 적용된다. $P(t_i|D)$는 문서에서 어휘가 발생할 확률, P(Q|D)는 초기질의에 대한 문서의 중요도를 의미한다.



(그림 2) 질의 어휘와 근접도를 적용한 단어 그래프

(그림 2)에서 각 노드(node)는 어휘를 의미하고, 식 (3)에서 $w(t_i, q_j)$는 그림에서 에지의 가중치를 의미한다. 식 (3)의 뒷부분에 $w(q_j, t_k)$에서 t_k는 q_j와 근접해있는 모든 어휘들이고 (그림 2)에서 $w(q_1, t_k)$는 q_1과 근접해있는 다섯 개의 어휘 사이에 존재하는 각 에지의 가중치를 의미한다. $w(t_i, q_j)$는 (그림 2)에서 q_1과 t_1사이를 연결한 에지의 가중치를 의

미한다. t_1은 q_1, q_2, q_3모두 근접해 있으므로 모든 질의 어휘들의 가중치를 받는다.

3.2 어휘 근접도를 적용한 에지의 가중치 계산

질의 어휘와 근접한 어휘는 질의와 의미적으로 연관되어 있다. 질의 어휘와의 근접도를 에지의 가중치로 하여 단어 그래프에 적용한다. 식 (5)는 에지의 가중치를 나타낸다.

$$w(t_i, q_j) = \sum_{t_i \in Near(q_j)} prox(t_i, q_j) \quad (5)$$

여기서 prox(t_i, q_j)는 문서 안에서 t_i와 q_j의 근접도이다. w(t_i, q_j)는 각 피드백 문서에서 구한 값을 피드백 문서 전체에서 더한다.

prox(t_i, q_j) 식은 다음과 같다.

$$prox(t_i, q_j) = 1 - \frac{dist(t_i, q_j)}{\delta} \quad (6)$$

여기서 dist(t_i, q_j)는 t_i와 q_j사이의 거리이다. δ는 거리 가중치 적용 파라미터이다.

예를들어, δ가 4라고 한다면, q_1을 중심으로 각 주변 어휘들의 가중치를 (그림 3)과 같이 적용할 수 있다. t_3과 q_1의 거리는 1이기 때문에 dist(t_3, q_1) = 1 이고, prox(t_3, q_1) = 1 - $\frac{1}{4}$= 0.75 이다. 질의 어휘에서 멀어질수록 가중치는 덜 받게 된다. 문서 전체에서 t_3과 q_1사이의 근접도를 더하면 에지의 가중치인 w(t_3, q_1)를 계산할 수 있다.

t1	t2	t3	q1	t4	t5	t6
0.25	0.5	0.75	1	0.75	0.5	0.25

(그림 3) 어휘와 질의 어휘 사이의 근접도 계산 방법

어휘의 위치를 이용한 문맥정보 연구[4,5]에서 질의 어휘 위치기반 언어모델은 질의 어휘들 사이의 거리를 코사인, 가우시안, 삼각형, 사각형 등의 그래프를 이용하여 위치 가중치를 계산했다. 본 논문에서는 그 중 삼각형을 이용한 방법이 학습질의에서 가장 좋은 성능을 보였기 때문에 삼각형을 이용하여 가중치를 계산했다.

제안된 방법을 통해 $f^{r+1}(t_i)$의 가중치가 높은 상위의 r개의 어휘를 확장질의로 선택하여 식(7)의 RM3(Relevance Model number 3)[14] 공식을 통해 문서의 중요도를 결정한다.

$$P(Q'|D) = \lambda \cdot P(Q|D) + (1-\lambda)P(W|D) \quad (7)$$

여기서 P(Q|D)는 언어모델에 의해서 계산된 초기질의에 대한 문서의 중요도이고, W는 r개의 확장된 어휘를 나타내며 P(W|D)는 확장질의를 적용하여 계산된 문서의 중요도를 나타낸다. P(Q'|D)는 초기질의를 통한 문서의 가중치와 확장질의를 통한 문서의 가중치를 결합하여 문서의 가중치를 새롭게 조정한 것이다. λ는 초기질의와 확장질의의 비중을 조절하는 파라미터이다.

4. 실험 및 평가

4.1 실험환경

실험 문서집합은 웹문서 집합인 TREC WT10g를 사용하였다. 학습질의를 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	
		질의번호	개수	질의번호	개수
WT10g	1,692,096	451-500	50	501-550	50

언어모델(LM)과 적합모델(RM)에 대한 실험결과는 인드리(Indri-2.8)시스템[15]을 사용하였다. 언어모델의 수식은 다음과 같다.

$$P(Q|D) = \prod_{i=1}^{k} \left(\frac{|D|}{|D|+\mu} \cdot \frac{f_{q_i,D}}{|D|} + \frac{\mu}{|D|+\mu} \cdot \frac{c_{q_i}}{|C|} \right) \quad (8)$$

여기서 k는 질의 어휘의 개수이고, |D|는 문서의 길이, |C|는 전체 컬렉션의 길이, $f_{q_i,D}$는 문서 D에서의 질의 어휘 q_i의 빈도수, C_{qi}는 전체 컬렉션에서의 q_i의 빈도수를 나타낸다. μ는 디리슈레 스무딩(Dirichlet smoothing) 파라미터로 μ 값은 학습질의에 대한 실험 ($\mu \in$ {500, 1000, 1500, 2000, … ,5000})에서 MAP가 가장 높은 값을 보인 2000으로 설정하였다.

3.1절의 식(3) 에서 어휘의 초기가중치 ($\alpha \in$ {0.1, 0.2, … , 0.9}), 3.2절의 식(6)에서 거리 가중치를 적용하기 위한 파라미터 ($\delta \in$ {5, 10, 25, 50, 74, 100})는 학습집합에서 가장 좋은 성능을 보인 값으로 선택했다. 피드백 문서의 개수 (n∈ {5, 10, 25, 50, 75, 100}), 확장 어휘의 개수 (r∈{5, 10, 25, 50, 75, 100}), 식 (7)에서 초기질의에 대한 가중치($\lambda \in$ {0.1, 0.2, … , 0.9})로 실험했다.

4.2 실험결과

단어 그래프를 이용한 질의 확장 실험결과 적합모델(RM)과 텍스트랭크를 이용한 질의 확장 방법과 제안된 기법으로 질의 확장한 결과를 비교하여 평가한다. 평가의 척도는 MAP(Mean Average Precision)이다.

- 언어모델(LM) : 질의가 특정 문서에서 발생할 확률을 계산하여 그 확률이 가장 큰 문서를 적합한 문서로 하고 상위에 순위화 된다.
- 적합모델(RM) : 언어모델에 의한 초기검색결과에서 상위에 순위화된 문서를 피드백하고, 어휘의 빈도를 반영해 확장질의를 선택한다.
- 텍스트랭크(TextRank) : 단어 그래프를 이용해 전체 어휘들을 대상으로 근접도를 계산하고 확장질의를 선택한다. 또한 어휘사이의 근접도는 사각형 그래프를 적용하여 계산한다.
- 제안방법 : 단어 그래프를 이용해 질의와 가까이 위치한 어휘들을 상대로 근접도를 계산하고 확장질의를 선택한다. 어휘 사이의 근접도는 삼각형 그래프를 적용하여 계산한다.

<표 2>는 제안된 방법과 초기 검색 결과와 적합모델을 통해 질의 확장한 결과에 대한 비교실험결과를 보여준다.

〈표 2〉 WT10g 컬렉션에 대해 제안된 방법과 적합모델 및 텍스트랭크와 비교실험결과

	언어모델(LM)	적합모델(RM)	텍스트랭크	제안방법
MAP	0.2125 (-)	0.2171 (+2.2%)	0.2217 (+4.3%)	0.2261 (+6.4%)

실험결과 제안된 방법이 실험집합에 대해서 언어모델보다 6.4%의 성능향상을 보였으며, 이는적합모델과 텍스트랭크를 이용해 확장질의를 선택한 것 보다 성능이 향상됨을 보였고, 어휘의 빈도만을 적용한 적합모델(RM)과 모든 어휘들 사이의 근접도를 적용한 텍스트랭크 보다 질의와의 근접도를 적용한 제안 방법이 질의 확장에 효율적임을 알 수 있다.

5. 결 론

본 논문에서는 적합성 피드백에서 질의 확장을 위해 각 어휘들의 독립적인 특성만을 고려한 어휘빈도를 이용한 기존의 방법을 문서 안에서 어휘들 사이의 관계와 문맥적인 특성을 이용하여 단어 그래프를 구성하고 어휘 근접도를 이용해 각 어휘들의 가중치를 결정하여 질의 확장질의를 선택하는 기법으로 대체 함으로써 적합모델(RM)과 텍스트랭크(TextRank) 보다 MAP가 향상되었다. 이 실험을 통해 문서 안에서 발생한 어휘들은 가까운 거리에 발생한 어휘와 서로 의미적으로 연관이 있음을 확인할 수 있었고, 이것을 통해서 질의 주변에 가까이 빈번하게 발생한 어휘들을 문맥어휘로 선택함으로써 기존의 어휘빈도를 이용한 방법을 대체할 수 있음을 보였다. 또한, 기존의 그래프 기반 순위화 알고리즘을 문서 안에 포함된 어휘들의 그래프에 적용함으로써 질의 확장 어휘를 선택하는데 유효함을 확인했다.

참 고 문 헌

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Blanco, R., Lioma, C. 2007. Random Walk Term Weighting for Information. In Proc. of 30th ACM SIGIR on Research and Development in Information Retrieval.

[7] 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.

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

[9] Mihalcea, R., Tarau, P., 2004. TextRank-Bringing Order into Texts. In Proc. of the Conference on Empirical Methods in Natural Language Processing(EMNLP 2004).

[10] 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.

[11] S. Hassan and C. Banea, 2006. Random-Walk Term Weighting for Improved Text Classification. In Proc. of TextGraphs: 2nd Workshop on Graph Based Methods for Natural Language Processing. ACL. pp.53-60.

[12] Page, L., Brin, S., Motowani, R. and Winograd, T. 1998. The PageRank Citation Ranking: Bringing Order to the Web, Unpublished manuscript, Stanford University.

[13] 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.

[14] Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Smucker, M.D., Wade, C. 2004. UMASS at TREC 2004-novelty and hard. In proc. Of the Thirteenth Text Retrieval Conference(TREC-13). pp.715-725.

[15] 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

장 계 훈

e-mail : ghjang@chonbuk.ac.kr
2009년 전북대학교 컴퓨터공학과(학사)
2011년 전북대학교 컴퓨터공학과(석사)
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)
2001년~2003년 일본 국립정보학연구소 (National Institute of Informatics) 연구원
2007년 미국 매사추세츠주립대학 방문교수
2004년~현 재 전북대학교 컴퓨터공학부/영상정보신기술연구센터 부교수
관심분야 : 정보검색, 정보 마이닝, 자연언어처리