질의확장에 의한 단락검색의 성능 향상에 관한 연구

A Study on the Improvement of Retrieval Performance : Query Expansion in Passage-based Retrieval

박지연, 정영미 (연세대학교 문헌정보학과)

Ji-Yeon Park, Young-Mee Chung
Dept. of Library and Information Science, Yonsei University

본 연구에서는 공기기반 질의-용어간 유사도를 이용한 질의확장을 통해 단락검색의 성능을 향상시키는 방안을 제시하고자 하였다. 실험을 통해 전체 문헌집단에 출현한 용어들의 공기정보에 기반한 전역적 질의확장과 이용자의 피드백 없이 초기검색 결과 중 상위 10개 문헌에 출현한 용어들의 공기정보에 기반한 지역적 질의확장의 성능을 비교하고 각각의 성능을 향상시키는 방법을 모색하였다. 마지막으로 문헌집단의 전역 정보와 지역 정보를 함께 이용하는 방안을 제시하고 그 성능을 평가하였다.

1 서론

검색 시스템은 일치하는 색인어를 갖고 있는 문헌을 검색하기 때문에 질의어와 색인어가 정확히 일치하지 않으면 검색이 되지 않는다. 이런 문제점을 보완하기 위한 것이 질의확장이다. 질의확장은 초기 질의와 관련 있는 유용한 용어들은 추가함으로써 이용자의 요구를 잘 표현할 수 있도록 하여 검색효율을 향상시키는 것이다(Buckley, and Salton 1994). 질의확장 방법은 전체 문헌집단의 정보를 이용하는 전역적 질의확장과 초기 검색 후 검색된 문헌 중 일부 정보만을 이용하는 지역적 질의확장을 나뉘어진다.

본 연구에서는 단락검색에서의 질의확장 실험을 통해 전역적 질의확장과 지역적 질의확장의 성능을 비교하고, 각각에서 검색 성능을 향상시킬 수 있는 방안을 연구하고자 한다. 또한 전역 정보와 지역 정보를 함께 이용하여 성능을 더욱 향상시킬 수 있는 방안을 제시하고자 한다.

2 질의확장

2.1 전역적 질의확장

전역 정보를 이용한 질의확장은 시소러스와 같은 개념 사전을 이용한 질의확장과 전체 문헌 내 용어의 공기빈도에 기반한 질의확장이 있다. 용어의 공기빈도에 기반한 방법은 두 용어가 한 문헌 안에서 동시 출현하는 회수가 클수록 밀접히 관련이 있다는 것을 전제하고 있다.

용어의 공기빈도에 기반한 질의확장에서 이

용하는 유사계수로는 자카드 계수(Jaccard coefficient), 다이스 계수(Dice coefficient), 코사인 계수(cosine coefficient), 상호정보량(mutual informaton) 등이 있다.

Qiu와 Frei(Qiu, and Frei 1993)는 질의확장 시 질의를 구성하는 용어들 중 한 용어와 밀접히 관련 있는 용어를 추가하기보다는 전체 질의 개념과 유사한 용어를 추가하는 질의-용어 간 유사도를 이용하여 질의확장의 성능 향상시켰다.

2.2 지역적 질의확장

지역적 질의확장은 이용자 피드백과 시스템 피드백으로 나뉘어진다.

시스템 피드백은 이용자의 개입 없이 초기 검색 결과 상위 순위에 검색된 문헌을 분석하여 질의확장을 하는 것이다. 시스템 피드백에서 확장 용어를 선택하는 방법에는 2가지가 있다. 하나는 초기 검색된 문헌 중 상위 n개를 적합하다는 전제 하에, 상위 n개의 문헌에 출현하는 용어들의 출현빈도를 이용하여 용어를 선택하는 방법으로 지역 피드백(local feedback)이라 한다(Xu, and Croft 1996). 두 번째는 지역 정보인 상위 n개의 문헌에 출현하는 용어들과 질의어와의 공기빈도를 이용하여 확장 용어를 선택하는 것으로 지역적 문맥 분석(local context analysis)이라 한다(Xu, and Croft 1996).

3 질의확장을 통한 단락검색 실험

3.1 실험설계

3.1.1 실험집단 및 질의

본 연구에서 이용한 문헌집단은 한글 테스트 컬렉션인 HANTEC2.0이다. 단락검색에서의 질의확장을 실험하기 위해 200자 이상의 문헌 4,846건을 임의로 선택하여 실험하였다. 질의는 HANTEC2.0의 질의 50개 중 30개를 임의로 선택하여 실험에 사용하였다.

3.1.2 실험 내용 및 방법

단락검색을 위해 4,846건의 실험집단의 문헌을 글자 크기 500자의 길이로 1/2씩 중복되게 단락을 구분하여 총 12,369개의 단락을 생성하여 실험하였다.

본 실험은 3단계로 이루어지는데 그 과정은 다음과 같다.

(1) 전역적 질의확장

전역적 질의확장 실험에서 추가 용어의 선정 기준은 세 가지이다. 로지스틱 함수로 정규화한 상호정보량, 전체 질의 개념과의 연관성, 그리고 전체 문헌집단에서의 역단락빈도를 단계적으로 조합하여 유사도를 측정하며 각각의 공식은 다음과 같다.

N : 질의 Q을 구성하는 질의어 수
PN : 총 단락 수

① 정규화 상호정보량(MI_n)

$$MI_n = \left(\left(\frac{1}{1+e^{-MI}}\right) - 0.5\right)\times 2$$

$$MI(t, Q) = \frac{1}{N}\left(\sum_{q_i \in Q}^{N} \log_2 \frac{PN \times df_{t q_i}}{df_t \times df_{q_i}}\right)$$

② 정규화 상호정보량(MI_n)과 전체 질의 개념과의 연관성(n/N)의 조합

$$SIM_1(t, Q) = MI_n \times \frac{n}{N}$$

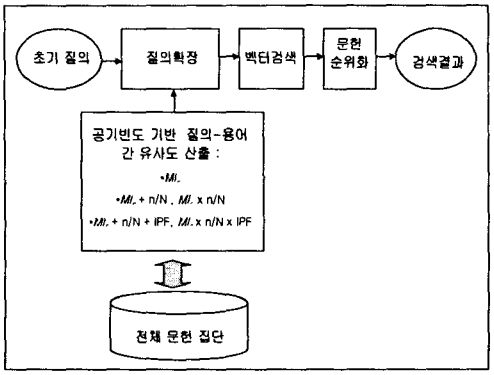
$$SIM_2(t, Q) = MI_n + \frac{n}{N}$$

③ 정규화 상호정보량(MI_n), 전체 질의 개념과의 연관성(n/N), 전체 문헌에서의 역단락빈도(IPF)의 조합

$$SIM_3(t, Q) = MI_n \times \frac{n}{N} \times IPF$$

$$SIM_4(t, Q) = MI_n + \frac{n}{N} + IPF$$

<그림 1>은 전역적 질의확장의 흐름도이다.



<그림 1> 전역적 질의확장 흐름도

(2) 지역적 질의확장

상위 10개 문헌 내 용어들과 질의어와의 공기빈도에 기반한 지역적 문맥 분석을 이용하여 30개의 용어를 추가하였다.

지역적 질의확장 실험은 정규화 상호정보량, 전체 질의 개념과의 연관성, 그리고 상위 10문헌 내 단락빈도를 조합하여 질의-용어간 유사도를 측정하며 각각의 공식은 다음과 같다.

① 정규화 상호정보량(MI_n)

② 정규화 상호정보량(MI_n)과 전체 질의 개념과의 연관성(n/N)의 조합

$$SIM_1(t, Q) = MI_n \times \frac{n}{N}$$

$$SIM_2(t, Q) = MI_n + \frac{n}{N}$$

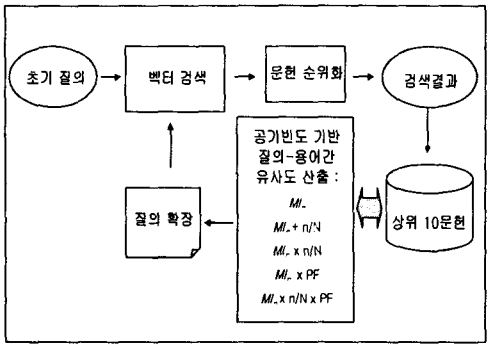
③ 정규화 상호정보량(MI_n)과 상위 10개 문헌 내 단락빈도(PF)의 조합

$$SIM_3(t, Q) = MI_n \times PF$$

④ 정규화 상호정보량(MI_n), 전체 질의 개념과의 연관성(n/N), 상위 10개의 문헌 내 단락빈도(PF)의 조합

$$SIM_4(t, Q) = MI_n \times \frac{n}{N} \times PF$$

지역적 질의확장의 전체적인 흐름은 <그림 2>와 같다.



<그림 2> 지역적 질의확장 흐름도

(3) 전역 정보와 지역 정보를 함께 이용한 질의확장

전역 정보와 지역 정보를 함께 이용하는 방법은 두 가지이다. 첫째는 지역적 질의확장 시 전역 정보인 전체 문헌에서의 역단락빈도를 동시에 이용하는 것으로 공식은 다음과 같다.

$$SIM_5(t, Q) = MI_n \times PF \times IPF$$

PF : 상위 10 문헌에서의 단락빈도(지역 정보)

IPF : 전체 문헌집단에서의 역단락빈도(전역 정보)

둘째는 전역 정보와 지역 정보를 단계적으로 이용하는 것으로, 전역적 질의확장을 통해 1차 검색의 성능을 향상시킨 후 상위 10개 문헌을 대상으로 지역적 질의확장을 실시하는 것

<표 1> 전역적 질의확장 결과

질의	초기 검색	MI_n	$MI_n + \frac{n}{N}$	$MI_n \times \frac{n}{N}$	$MI_n + \frac{n}{N}$ +IPF	$MI_n \times \frac{n}{N} \times$IPF
3-지점 정확률 평균 (증가율)	0.4464	0.4960 (10.53%)	0.5091 (14.03%)	0.4998 (11.96%)	0.4861 (8.88%)	0.4859 (8.84%)
R-정확률 평균 (증가율)	0.4039	0.4528 (12.10%)	0.4704 (17.35%)	0.4641 (15.00%)	0.4590 (13.64%)	0.4614 (14.23%)

<표 2> 지역적 질의확장 결과

질의	초기 검색	MI_n	$MI_n + \frac{n}{N}$	$MI_n \times \frac{n}{N}$	$MI_n \times$PF	$MI_n \times \frac{n}{N} \times$PF
3-지점 정확률 평균 (증가율)	0.4464	0.4830 (8.19%)	0.4877 (9.22%)	0.4885 (9.43%)	0.4921 (10.22%)	0.5042 (13.00%)
R-정확률 평균 (증가율)	0.4039	0.4397 (8.86%)	0.4438 (9.87%)	0.4440 (9.92%)	0.4592 (13.69%)	0.4634 (14.73%)

이다.

3.1.3 실험 결과 평가 방법

본 논문에서는 성능을 평가하기 위해 재현율의 값을 0.25, 0.5, 0.75인 세 지점으로 고정시키고 이에 대한 정확률 값들의 평균을 구하는 방법을 이용하였다. 또한 하위에 검색된 문헌보다는 상위에 검색된 적합 문헌의 비율을 고려한 평가 방법인 R-정확률도 함께 사용하였다.

3.2 실험 결과 분석

3.2.1 전역적 질의확장

정규화 상호정보량, 전체 질의 개념과의 연관성, 역단락빈도를 조합한 유사도 공식에 의해 확장 용어를 선정한 결과 각 방법에 따라 검색 성능이 달라졌다. 용어 선택 방법에 따른 검색효율은 <표 1>과 같다. <표 1>를 보면 정규화 상호정보량과 전체 질의 개념과의 연관성을 조합한 방법이 초기 검색 보다 성능이 최고 17.35% 향상되어 다른 방법보다 성능이 월등히 뛰어난 것을 알 수 있다. 정규화 상호정보량, 전체 질의 개념과의 연관성, 역단락빈도를 조합한 방법은 초기 검색보다는 성능이 향상되었으나, 역단락빈도를 사용하지 않고 정규화 상호정보량과 전체 질의 개념과의 연관성만을 조합한 방법보다는 성능이 저하되었다.

3.2.2 지역적 질의확장

정규화 상호정보량, 전체 질의 개념과의 연관성, 상위 10개 문헌 내 단락빈도를 조합하여 확장 용어를 선정한 결과 검색 성능은 <표 2>와 같다.

정규화 상호정보량과 상위 10개 문헌에서의 단락빈도를 조합한 방법과 정규화 상호정보량과 전체 질의 개념과의 연관성을 조합한 방법 모두 정규화 상호정보량만을 이용한 방법 보다 성능이 향상되었다. 이 결과를 바탕으로, 전체 질의 개념과의 연관성과 상위 10개의 문헌에서의 단락빈도를 조합하여 실험하였다. <표 2>를 보면 이 방법이 초기검색 보다 3-지점 정확률 평균은 13.00%, R-정확률 평균은 14.73% 향상

되어 지역적 질의확장 방법들 중 가장 성능이 우수한 것을 알 수 있다.

3.2.3 전역 정보와 지역 정보를 함께 이용한 질의확장

전역 정보와 지역 정보를 함께 이용한 질의확장은 두 가지 유형의 정보를 동시에 이용하는 방법과 단계적으로 이용하는 방법으로 나누어 실험하였다.

우선 지역 정보인 상위 10개 문헌에서의 정규화 상호정보량(MI_n)과 상위 10개 문헌에서의 단락빈도(PF), 그리고 전역 정보인 전체 문헌에서의 역단락빈도(IPF)를 조합하여 실험한 결과는 <표 3>과 같다.

<표 3> 전역 정보와 지역 정보를 동시에 이용한 질의확장 결과

	초기 검색		$MI_n \times PF \times IPF$	
질의	3-지점 정확률	R-정확률	3-지점 정확률	R-정확률
평균 (증가율)	0.4464	0.4039	0.5320 (19.16%)	0.5002 (23.84%)

초기 검색 성능과 비교해 보면 3-지점 정확률 평균은 19.16%, R-정확률 평균은 23.84% 향상되었는데, 이 결과는 상위 문헌에서는 자주 출현하면서 전체 문헌에는 적게 출현하는 용어들이 적합 문헌을 식별하는 능력이 있음을 보여 주는 것이다.

전역 정보와 지역 정보를 단계적으로 이용한 두 번째 실험에서는 전역적 질의확장에서 가장 성능이 좋았던 방법(MI_n+n/N)으로 질의확장을 한 후, 1) 정규화 상호정보량, 2) 지역적 질의확장에서 가장 성능이 좋았던 방법($MI_n \times n/N \times PF$)을 각각 이용해 2단계 질의확장을 하였는데 그 결과는 <표 4>와 같다. <표 4>를 보면 정규화 상호정보량만 이용한 방법보다는 전체 질의 개념과의 연관성과 상위 10개 문헌 내 단락빈도를 조합한 두 번째 방법이 성능이 우수한 것을 알 수 있다.

3.2.4 질의확장 성능 비교 분석

<표 5>은 전역적 질의확장, 지역적 질의확장 그리고 전역 정보와 지역 정보를 함께 이용한 질의확장의 검색 성능을 비교한 것이다. 각각의 질의확장 실험에서 가장 성능이 우수한 방법을 이용하여 비교했으며, 전역+지역(1)은 전역 정보와 지역 정보를 동시에 이용한 것이

<표 4> 전역 정보와 지역 정보를 단계적으로 이용한 질의확장 결과

질의	초기 검색	MI_n	$MI_n \times \frac{n}{N} \times PF$
3-지점 정확률 평균 (증가율)	0.4464	0.4918 (10.17%)	0.5117 (14.63%)
R-정확률 평균 (증가율)	0.4039	0.4646 (15.03%)	0.4845 (19.96%)

<표 5> 질의확장 방법에 따른 성능 비교

	초기 검색	전역적 질의확장	지역적 질의확장	전역+지역(1)	전역+지역(2)
3-지점 정확률 평균 (증가율)	0.4464	0.5091 (14.03%)	0.5042 (13.00%)	0.5320 (19.16%)	0.5117 (14.63%)
R-정확률 평균 (증가율)	0.4039	0.4704 (17.35%)	0.4634 (14.73%)	0.5002 (23.84%)	0.4845 (19.96%)

고, 전역+지역(2)은 전역적 질의확장 후, 2단계로 지역적 질의확장을 실시한 경우이다. < 표 5>을 보면 전역적 질의확장의 성능이 지역적 질의확장의 성능보다 다소 우수하다. 또한 전역 정보와 지역 정보를 함께 이용한 질의확장을 살펴보면, 어느 한 정보원만 이용한 질의확장 보다 검색효율이 높았다. 특히 전역 정보와 지역 정보를 동시에 이용한 방법은 3-지점 정확률 평균 19.16%, R-정확률 평균 23.84%를 향상시켜 다른 질의확장 방법보다 성능이 월등히 뛰어났다.

4 결론

본 연구에서는 질의확장의 성능을 향상시키기 위해 전체 질의 개념과의 연관성, 전체 문헌집단에서 역단락빈도, 초기 검색의 상위 10개 문헌에서의 단락빈도를 정규화 상호정보량과 함께 조합하여 질의확장 실험을 하였고, 마지막으로 문헌집단의 전역 정보와 지역 정보를 함께 이용하는 방안을 제시하고 그 성능을 평가하였다. 본 연구에서 밝혀진 사실은 다음과 같다.

첫째, 질의-용어간 공기빈도에 기반한 전역적 질의확장과 지역적 질의확장 모두 검색 성능을 향상시켰다.

둘째, 질의확장 시, 전체 질의 개념과의 연관성이 높은 용어들을 우선적으로 추가하여 검색한 결과 검색 성능이 월등히 향상되었다.

셋째, 질의확장 시, 고빈도어가 추가되는 것을 방지하기 위해 역단락빈도를 이용한 결과 검색 성능이 향상되지 않았다.

넷째, 지역적 질의확장에서, 질의어와의 공기빈도가 높고 전체 질의 개념과의 연관성이 크면서 상위 10개의 문헌에 많이 출현하는 용어들을 추가하여 검색 한 방법이 지역적 질의확장 방법들 중에서 성능이 가장 높았다.

다섯째, 전역 정보인 역단락빈도와 지역 정보인 상위 10개 문헌 내 단락빈도를 동시에 이용하여 확장 용어를 선택한 방법이 성능이 가장 뛰어났다.

마지막으로, 전역적 질의확장에서 성능이 가장 좋은 방법, 지역적 질의확장에서 성능이 가장 좋은 방법, 그리고 전역 정보와 지역 정보를 함께 이용한 방법을 비교한 결과 전역 정보와 지역 정보를 함께 이용한 방법이 성능이 뛰어났다.

참고문헌

Buckley, C., and G. Salton. 1994. " The effect of adding relevance information in a relevance feedback environment." *Proceedings of the 17 Annual International ACM SIG Conference on Research a Development in Information Retrie* 292-300.

Qiu, Yonggang, and Hans-Peter Frei. 1993 "Concept based query expansion." *Proceedings of ACM SIG International Conference on Rese and Development in Informati Retrieval,* 160-169.

Xu, Jinxi, and W. Bruce Croft. 1996. "Query expansion using local and global document analysis." *Proceedings of ACM SIG International Conference on Rese and Development in Informati Retrieval,* 4-11.