검색 문헌의 인용 분석을 통한 질의확장의 성능 평가 연구

An Evaluation of the Performance of Query Expansion Using Citation Information of Retrieved Documents

유소영, 서울여자대학교 도서관, applesnow@swu.ac.kr
정영미, 연세대학교 문헌정보학과, ymchung@yonsei.ac.kr

So Young, Yu, Seoul Women's University Library
Young-Mee Jung, Dept of Library and Information Science, Yonsei University

이 연구에서는 주제검색을 통해 검색된 문헌들의 인용정보를 이용한 질의확장 기법을 제안하였으며 이 제안된 기법의 성능을 일반적 질의확장 기법인 지역적 질의확장 및 전역적 질의확장과 비교 평가하였다. 연구 결과 인용 기반 질의확장 기법이 전역적 및 지역적 질의확장 기법에 비해 우수한 성능을 보임을 확인하였으며, 특히 피인용 표제어를 이용한 질의확장 검색의 효용성을 실험을 통해 밝혀냈다.

1 서론

인터넷의 등장과 최종 이용자의 직접적인 정보검색이 보편화되면서 짧은 질의어로 인해 적합한 문헌이 충분히 검색되어 나오지 못하는 경우가 발생한다.

이에 대한 해결책으로 제시된 것이 질의확장 기법이다(Xu and Croft 1996). 일반적으로 질의확장기법은 전체 실험문헌 집단 내 단어의 동시출현빈도에 근거한 전역적 질의확장(global query expansion)기법과 일차 검색 결과의 적합성 피드백 정보를 이용하는 지역적 질의확장(local query expansion)으로 나눌 수 있다.

전역적 질의확장 기법에는 시소러스를 이용하는 방법과 전역적 문맥 분석을 이용하는 방법이 있다. 일반적으로 시소러스를 이용한 실험의 대부분의 결과는 검색 효율을 증진시키지 못하는 것으로 나타났다.

Qui 와 Frei(1993)는 전역적 문맥분석을 이용할 때 질의가 표현하는 전체 질의 개념과 유사한 용어를 질의어로 추가할 수 있는 질의-용어간 유사도 공식을 고안해 내었다. 이를 이용한 최근 연구들에서는 기존의 질의어-용어간 유사도를 이용한 연구들보다 검색 성능이 향상되는 것으로 나타났다.

지역적 질의확장에서는 적합성 피드백을 기반으로 하는데, 적합성 피드백은 주로 적합문헌을 판정하는 방법에 따라 이용자 피드백과 시스템 피드백으로 나뉜다. 일반적으로 이용자 피드백을 통한 질의확장의 성능이 우수하다고 알려져 있다.

인용정보를 이용한 질의 확장 및 연관 검색 연구에서는 인용 문헌의 인용정보 자체를 이용하는 기법이나, 동시인용과 서지결합의 개념을 도입하여 자동적으로 확장질의를 생성하는 기법, 인용문헌의 본문을 이용하는 등의 기법 등이 제시되었다. 그리고 인용관계 정보를 직접 주제검색과 결합하여 검색 성능을 향상시키고자 한 연구들도 이루어졌다(Ding et al. 2000).

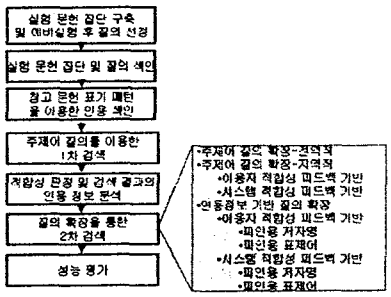
이 연구에서는 주제어를 이용한 일차 검색 결과 문헌들의 인용정보를 분석하여 이를 질의확장에 이용하여 검색 성능을 향상시키고자 하였다. 또 인용정보를 이용한 질의확장 기법이 전역적 및 지역적 질의확장 기법보다 더 좋은 검색 성능을 보이는지 살펴보고자 하였다. 뿐

만 아니라 기존의 인용정보를 이용한 연관검색 연구의 기법들보다 효율적인 질의확장 알고리즘을 제안하고자 하였다.

2 인용정보를 이용한 질의확장실험

2.1 실험 설계

실험의 전체적인 개요는 <그림 1>과 같다.



<그림 1> 실험 개요

이 연구에서는 실험문헌 집단을 자체적으로 구축하였다. 실험문헌 집단은 1996년부터 2003년 사이에 출간된 정보학 분야의 영문 학술회의 논문집과 학술잡지 6종을 대상으로 하였다.

실험문헌 집단은 총 1,396개 문헌으로 이루어졌으며 한 문헌의 평균 길이는 2,708개 색인어이다.

예비실험을 통해 정보학 분야에서 질의어로 많이 사용될 만한 주제어를 이용하여 12개의 영문 질의로 된 질의집단을 구성하였다. 각 질의 및 색인 결과 추출된 질의어는 <표 1>과 같다.

검색 실험을 위하여 각 문헌의 색인어에는 로그TF·IDF의 가중치를, 질의어에는 이진TF의 가중치를 주었다. IDF 공식은 Sparck Jones가 제시한 역문헌빈도(IDF) 공식을 이용하였다.

검색에는 코사인 유사 계수를 이용하였고 검색 후 상위 50개 문헌에 대하여 시스템 및 이용자에 의한 적합성 판정이 이루어졌다.

질의 번호	질의	질의어
1	music retrieval	music, retrieval
2	content-based image retrieval	content, base, image, retrieval
3	web mining and application	web, mine, application
4	information retrieval based on XML	xml, base, information, retrieval
5	user-centered interface design	user, interface, center, design
6	visualization of intellectual structure based on citation analysis	citation, base, analysis, visualization, intellectual, structure
7	the relation among relevance, topicality and preference	relation, relevance, topicality, preference
8	feature selection in document categorization	document, categorization, feature, selection
9	visualization method for document navigation in digital library	digital, library, visualization, document, navigation
10	collection fusion strategy for information retrieval	collection, fusion, strategy, information, retrieval
11	query expansion	query, expansion
12	user's criteria on relevance evaluation	user, criteria, relevance, evaluation

<표 1> 질의집단

이용자에 의한 적합성 판정은 연세대학교 문헌정보학과 대학원생 2명이 미리 정해진 적합성 판정 기준에 따라서 실시하였다.

시스템에 의한 적합성 판정은 문헌 내 용어의 가중치를 TF·IDF, 로그TF·IDF, 이진TF·IDF로 변화시키면서 주제어에 의한 검색실험을 수행한 결과를 조합하여 이루어졌다.

검색성능 평가는 수정정확률과 10위내 순위정확률을 이용하였다.

$$\text{수정정확률} = \frac{\text{검색된 상위 50개 문헌내 적합문헌수}}{50}$$

$$\text{RP@10} = \frac{P@1 + P@2 + \cdots + P@10}{10}$$

(P@i: 순위 i에서의 정확률)

2.2 질의확장 검색 실험

2.2.1 전역적 및 지역적 질의확장

전역적 질의확장 기법으로는 단어동시출현빈도에 기반한 질의확장을 실시하였다.

질의Q와 용어t 간의 유사도는 다음의 공식을 이용하였다.

$$Q=(q_1,q_2,q_3,\ldots,q_n)$$

$$S(Q,t)=\frac{1}{n}\sum_{i=1}^{n}S(q_i,t)$$

질의-용어간 유사도 계산후 유사도값이 큰 상위 50개 용어를 추가 질의어 후보로 선정하였다. 전역적 질의확장 검색 실험에서는 이 추가 질의어 후보를 10개부터 50개까지 10개 단위로 원질의에 추가하여 검색을 수행하였다.

지역적 질의확장 검색 실험에서는 이용자 및 시스템 적합성 피드백을 각각 이용하였다. 적합성 피드백을 바탕으로 다음 공식과 같이 적합문헌에 출현한 용어만을 질의확장에 반영하는 로치오공식을 사용하였다.

$$Q_1=\alpha Q_0+\beta\sum_{i=1}^{n_1}\frac{R_i}{n_1}$$

R_i : 적합문헌 i 의 벡터

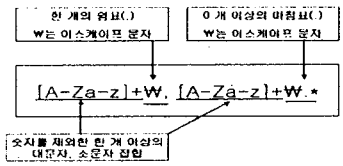
n_1 : 적합문헌 수

재계산된 용어 가중치가 높은 상위 50개 용어를 추가 질의어 후보로 선정하고 이를 10개부터 50개까지 10개 단위로 원질의에 추가하여 검색 실험을 수행하였다.

2.2.2 인용정보 기반 질의확장

인용정보 기반 질의확장 검색 실험에서는 지역적 질의확장 검색 실험과 동일하게 이용자 및 시스템 적합성 피드백 정보를 바탕으로 피인용 저자명 및 피인용 표제어를 추출하여 추가 질의어로 사용하였다. 피인용 저자명과 피인용 표제어는 정규표현식(Friedl 2002)을 이용한 인용정보의 자동생성 방법을 통해 색인하였다.

피인용 저자명 자동 색인을 위해서 사용된 정규표현식은 <그림 2>와 같다.

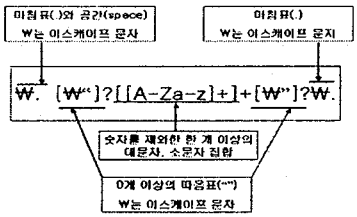


<그림 2> 피인용 저자명 추출을 위한 정규표현식

피인용 저자명 식별 작업 후 결과를 확인하여 발생한 오류는 수작업 확인을 통해 제거하였다.

이 연구에서 사용한 피인용 저자명의 자동 추출 및 색인 방법은 저자의 성은 동일하나 저자의 이름이 다른 두 저자는 구별하지 않았다. 그러나 이 연구에서는 피인용 저자명을 주제어 필드로 이용하였고 질의확장의 목적 자체가 적합문헌의 재현율을 높이고자 하는 것이므로 위와 같은 제한점에도 불구하고 이와 같이 간단한 피인용 저자명의 자동 추출 및 색인 방법으로도 질의확장에 이용할 수 있을 것으로 판단하였다.

피인용 표제어는 <그림 3>의 정규표현식을 이용하여 식별하였다.



<그림 3> 피인용 표제 추출을 위한 정규표현식

피인용 표제를 식별한 후 각 피인용 표제어를 색인하고 각 문헌별 출현 빈도를 계산하였다. 자동 색인 작업에서는 일반적인 주제 색인에서 사용되는 불용어 리스트에 overview, report, research, study, experiment, discovery, approach, method, review, result 등의 용어를

불용어로 추가하였다.

자동 색인한 피인용 저자명은 단순 빈도와 상대빈도를 이용하여 추가 질의어 후보를 선정하였고 피인용 표제어는 단순빈도를 이용하여 추가 질의어 후보를 선정하였다.

그리고 질의확장에 사용할 피인용 표제어는 단순빈도에 따른 상위 10개부터 50개까지 용어를 10개씩 증가시키며 확장하였다. 이는 검색 성능 평가의 비교 대상이 되는 전역적 질의확장 기법과 지역적 질의확장 기법과 동일한 수의 추가 질의어를 이용하기 위함이었다.

3 질의확장 실험 결과 분석 및 평가

3.1 일차 주제어 검색 실험 결과

주제어를 이용하여 일차 검색을 수행한 결과 각 질의당 평균 적합문헌수는 9.5개이고 평균 0.19의 수정정확률을 보였다. 그리고 순위정확률은 평균 0.39이었다. 각 질의별 주제어 검색 결과 상위 50개 문헌 내 적합문헌 수와 수정정확률, 순위정확률은 <표 2>와 같다.

질의 번호	적합문헌수	수정정확률	순위정확률
1	9	0.180	0.857
2	16	0.320	0.266
3	11	0.220	0.479
4	11	0.220	0.484
5	11	0.220	0.400
6	5	0.100	0.000
7	5	0.100	0.010
8	18	0.360	0.615
9	5	0.100	0.010
10	3	0.060	0.486
11	11	0.220	0.385
12	9	0.180	0.645
평균	9.5	0.190	0.386

<표 2>주제어를 이용한 일차 검색 성능

3.2 질의확장 검색 실험 결과

각 질의확장 기법의 평균 수정정확률 및 평균 순위정확률은 <표 3>과 같다. 수정정확률은 모든 질의확장 기법에서 일차 검색보다 향상되었다. 순위정확률도 대부분의 모든 질의확장 기법에서 일차 검색보다 향상되었으나 전역적 질의확장 기법에서는 일차 검색보다 하락하였다. 그리고 지역적 및 인용정보 기반 질의확장 기법에서 이용자 피드백을 이용하는 경우가 시스템 피드백을 이용하는 경우보다 수정정확률과 순위정확률이 높은 것으로 나타났다. 또 모든 질의확장 기법 중에서 인용정보 기반 질의확장 기법이 가장 높은 성능을 보였다.

이용자 피드백을 이용한 경우의 질의확장 기법별 수정정확률을 비교하면 피인용 표제어 > 피인용 저자 상대빈도 > 피인용 저자 단순빈도 > 지역적 > 전역적 질의확장 기법의 순이었다. 그리고 순위정확률은 (피인용 표제어 = 전역적) > 피인용 저자 단순빈도 > 피인용 저자 상대빈도 > 전역적 질의확장 기법의 순이었다.

	일차검색	전역적	지역적		인용정보 기반					
					이용자			시스템		
			이용자	시스템	저자(상대빈도)	저자(단순빈도)	표제	저자(상대빈도)	저자(단순빈도)	표제
수정정확률	0.19	0.21	0.30	0.25	0.37	0.36	0.42	0.26	0.26	0.34
성능향상률(%)	-	7.7	57.4	32.6	93.3	88.6	119.8	8.8	37.4	80.2
순위정확률	0.39	0.37	0.82	0.42	0.70	0.74	0.82	0.46	0.55	0.61
성능향상률(%)	-	-3.5	112.0	9.0	82.0	92.1	112.0	18.0	41.9	58.7

<표 3>질의확장 검색 실험의 성능 비교

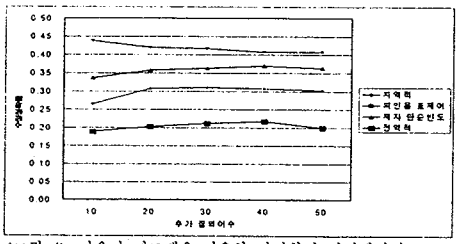
시스템 피드백을 이용한 경우의 질의확장 기법별 수정정확률은 피인용 표제어 > 피인용 저자 단순빈도 > 피인용 저자 상대 빈도 > 지역적 > 전역적 질의확장 기법 순으로 높게 나타났으며, 순위정확률은 피인용 표제어> 피인용 저자 단순빈도 > 피인용 저자 상대빈도 > 전역적 > 지역적 질의확장 기법의 순으로 높게 나타났다.

이용자 피드백 및 시스템 피드백을 이용한 경우 모두 피인용 표제어를 이용한 인용정보

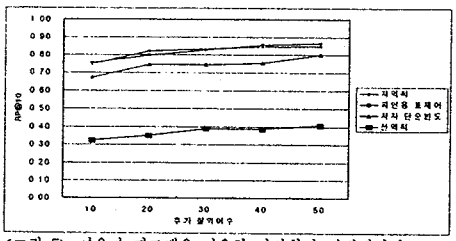
기반 질의확장 기법이 가장 높은 수정정확률과 순위정확률을 보였다. 그러나 인용정보 기반 질의확장 기법 중 피인용 저자의 상대빈도를 이용하는 경우에는 적합성 피드백 방법에 따라 수정정확률과 순위정확률이 차이가 크게 났다.

이용자 피드백을 이용한 경우의 추가되는 질의어 수에 따른 각 질의확장 기법의 수정정확률 변화와 순위정확률 변화는 <그림 4>, <그림 5>와 같다.

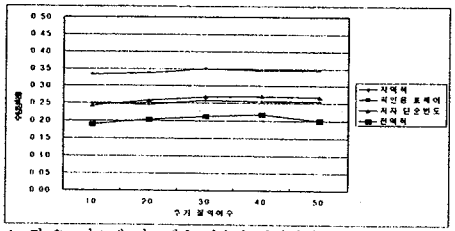
시스템 피드백을 이용한 경우의 추가되는 질의어 수에 따른 각 질의확장 기법의 수정정확률 변화와 순위정확률 변화는 <그림 6>, <그림 7>과 같다.



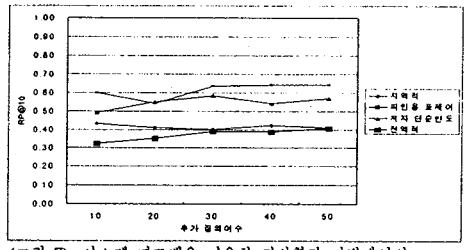
<그림 4> 이용자 피드백을 이용한 질의확장 기법에서의 추가 질의어 수에 따른 수정정확률 변화



<그림 5> 이용자 피드백을 이용한 질의확장 기법에서의 추가 질의어 수에 따른 순위정확률 변화



<그림 6> 시스템 피드백을 이용한 질의확장 기법에서의 추가 질의어 수에 따른 수정정확률 변화



<그림 7> 시스템 피드백을 이용한 질의확장 기법에서의 추가 질의어 수에 따른 순위정확률 변화

또 질의확장 기법들의 성능을 정보 요구가 구체적인 질의집단(질의 번호 7, 9, 10, 12번)과 정보 요구가 보다 덜 구체적인 질의집단(질의 번호 1, 2, 3, 4, 5, 6, 8, 11번)으로 나누어 비교해 보았다.

일반적인 질의와 구체적인 질의집단으로 나누었을 때의 각 기법별 평균 수정정확률 및 순위정확률은 <표 4>와 같다.

일반적인 질의집단										
	일차검색	전역적	지역적		인용정보 기반					
					이용자			시스템		
			이용자	시스템	저자(상대빈도)	저자(단순빈도)	표제	저자(상대빈도)	저자(단순빈도)	표제
수정정확률	0.23	0.24	0.33	0.30	0.36	0.38	0.44	0.25	0.27	0.38
순위정확률	0.44	0.44	0.80	0.50	0.67	0.74	0.86	0.52	0.57	0.70
구체적인 질의집단										
	일차검색	전역적	지역적		인용정보 기반					
					이용자			시스템		
			이용자	시스템	저자(상대빈도)	저자(단순빈도)	표제	저자(상대빈도)	저자(단순빈도)	표제
수정정확률	0.11	0.14	0.23	0.15	0.38	0.32	0.38	0.23	0.25	0.27
순위정확률	0.29	0.24	0.85	0.25	0.77	0.75	0.74	0.43	0.50	0.45

<표 4> 질의집단별 질의확장 검색 실험의 성능 평가

이를 <표 3>의 전체 질의집단의 각 기법별 평균 수정정확률 및 순위정확률과 비교해 보면 일반적인 질의집단의 경우 전체 질의집단에 비해서 수정정확률 및 순위정확률이 높게 나타났다. 반면에 구체적인 질의집단의 경우 전체 질의집단에 비해서 수정정확률 및 순위정확률이 대체적으로 낮게 나타났다.

4 결론

이 연구의 질의확장 검색 실험을 통해 다음과 같은 결과를 얻었다.

첫째, 모든 질의확장 기법은 주제어를 이용한 일차 검색보다 높은 수정정확률 및 10순위내 순위정확률을 보였다. 단 전역적 질의확장 기법을 이용한 경우의 10순위내 순위정확률이 일차 검색의 10순위내 순위정확률보다 다소 하락하였다.

둘째, 전역적 질의확장 기법은 지역적 질의확장 기법 및 인용정보 기반 질의확장 기법에 비해 평균적으로 낮은 성능을 보였다.

셋째, 지역적 질의확장 기법 및 인용정보 기반 질의확장 기법은 이용자 피드백을 이용한 경우가 시스템 피드백을 이용한 경우보다 수정정확률 및 10순위내 순위정확률이 높게 나타났다.

넷째, 인용정보 기반 질의확장 기법이 이용자 피드백을 이용하는 경우와 시스템 피드백을 이용하는 경우 모두에서 지역적 질의확장 기법보다 대체적으로 높은 수정정확률 및 10순위내 순위정확률을 보였다.

다섯째, 인용정보 기반 질의확장 기법은 이용자 피드백을 이용하는 경우에는 추가 질의어로 사용되는 피인용 저자명의 수가 일정 수 이상 많아질수록 수정정확률 및 10순위내 순위정확률이 높아지는 경향을 보였다. 시스템 피드백을 이용하는 경우에는 상대빈도가 0 이상인 저자, 즉 출현한 모든 피인용 저자를 질의확장에 이용하는 경우가 가장 성능이 좋았다. 그리고 피인용 표제어가 피인용 저자명을 이용할 때보다 더 높은 수정정확률 및 10순위내 순위정확률을 나타냈다.

여섯째, 일반적인 질의집단에서는 대체적으로 전체 질의집단에 비해 수정정확률 및 10순위내 순위정확률이 조금 낮게 나타났다.

이상의 실험 결과를 통해 본 연구에서는 인용기반 질의확장 기법이 이용자 피드백을 이용하는 경우와 시스템 피드백을 이용하는 경우 모두 지역적 질의확장 기법보다 높은 성능을 보일 수 있음을 확인하였다.

참고문헌

유소영. 2004. **검색 문헌의 인용 분석을 통한 질의확장의 성능 평가 연구.** 석사학위논문, 연세대학교 대학원, 문헌정보학과.

Ding, Y., Gobinda, C. G., Schubert, F., and Weizhong, Q. 2000. "Bibliometric Information Retrieval System(BIRS): A Web Search Interface Utilizing Bibliometric Research Results", *Journal of the American Society for Information Science*, 51(13): 1190-1204.

Friedl, Jeffrey E. F. 2002. *Mastering Regular Expressions*. 2nd ed. O'Reilly.

Qiu, Y. and Frei, Hans-Peter. 1993. "Concept Based Query Expansion". In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 160-169.

Xu, J., and Croft, W. B. 1996. "Query Expansion Using Local and Global Document Analysis". In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 4-11.