온라인 이용자 피드백을 사용한 정보필터링 시스템의 수정질의 최적화에 관한 연구*

A Study on Query Refinement by Online Relevance Feedback in an Information Filtering System

최 광(Kwang Choi)**, 정영미(Young Mee Chung)***

초 록

이 연구의 목적은 대량의 최신정보를 제공하는 정보필터링 시스템에서 이용자 피드백에 의해 수정질의를 자동생성하여 재검색을 수행함으로써 검색 성능을 최적화할 수 있는 방안을 찾는 데 있다. 이용자가 입력한 초기질의를 사용하여 정보필터링 시스템이 검색한 문헌에 대해 이용자가 적합성 여부를 온라인으로 입력하도록 하고, 이 피드백 결과를 토대로 '중복제거법'과 '저빈도제거법' 두 가지 방법에 의해 각각 17개의 수정질의를 생성하여 재검색한 결과를 초기 검색결과와 비교 분석하였다. 수정질의는 각각의 방법마다 17개 패턴의 불논리 질의형태를 미리 만든 다음 초기질의에 디스크립터와 분류기호를 결합하여 생성하였으며, 재검색 결과에 대한 적합성 평가를 통해 최적의 수정질의식을 도출하였다.

ABSTRACT

In this study an information filtering system was implemented and a series of relevance feedback experiments were conducted using the system. For the relevance feedback, the original queries were searched against the database and the results were reviewed by the researchers. Based on users' online relevance judgements a pair of 17 refined queries were generated using two methods called "co-occurrence exclusion method" and "lower frequencies exclusion method." In order to generate them, the original queries, the descriptors and category codes appeared in either relevant or irrelevant document sets were applied as elements. Users' relevance judgments on the search results of the refined queries were compared and analyzed against those of the original queries.

키워드: 질의수정, 질의확장, 정보필터링, 이용자 피드백, 적합성 피드백, 불논리 검색, query refinement, query expansion, information filtering, user feedback, relevance feedback, Boolean retrieval

---

* 이 논문은 연세대학교 문헌정보학과 박사학위논문을 요약한 것임.

** (주)오롬정보 교육정보화사업팀 이사(choik@orom.com)

*** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr)

■ 논문 접수일 : 2002. 11. 11

■ 게재 확정일 : 2002. 11. 26

1 서 론

인터넷을 비롯한 정보기술의 발전은 정보시스템 상에 유통되는 정보의 양적 증가와 함께 웹, 전자우편, 게시판 등 정보 유형의 다양화 및 정보 발생주기의 일상화를 초래하였고, 이에 따라 새로운 형태의 정보제공 환경의 구축이 필요하게 되었다.

실제로 특정 주제분야의 연구원들이 활용하는 최신 정보원으로는 도서관이 전통적으로 제공하는 단행본, 학술지 등의 신간 입수정보 외에 대량의 신간 학술지 목차정보, 특정 데이터베이스 갱신판의 최신 정보, 전자저널 제공 사이트의 신규정보, 전문 웹 사이트의 수록정보 등 매우 다양한 정보원이 존재한다.

새로 생산된 다양한 정보원을 대상으로 한 필요한 정보의 검색과 불필요한 정보의 제거(즉 필터링)은 동전의 양면과 같은 것으로서, 이제는 궁극적으로 이용자가 원하는 정보만을 제공하기 위하여 어떤 정보원을 어떤 방법으로 검색할 것인가가 매우 중요한 문제로 대두되고 있다.

최신의 정보원을 중심으로 이용자에게 적합한 정보를 제공하려는 노력은 SDI(Selective Dissemination of Information) 서비스 기술을 통하여 추구되어 왔으며, 최근 생산되는 웹 정보원에 대해서는 웹 로봇(Web robot) 또는 웹 에이전트(Web agent)를 통한 정보 수집 및 제공에 초점을 맞추고 있다. 특히 스팸메일의 제거와 필요 정보의 획득이라는 양면적인 문제를 안고 있는 전자우편의 경우 불필요한 정보의 제거가 강조되고 있다.

최근 웹을 통한 대량의 정보 제공 및 입수를 위해 다양한 웹 검색엔진들이 개발되어 사용되고 있으나 정보검색의 효과 측면에서는 여러 문제점을 안고 있다. 특히 특정 검색도구를 사용하여 검색한 정보의 양이 많아지게 됨에 따라 이용자가 일일이 검색결과를 확인하는 것이 거의 불가능해졌고, 또한 직접 검색을 통해서만 필요한 정보를 얻을 수 있다는 점은 새로운 정보시스템에 대한 요구를 증대시키고 있다.

정보필터링 시스템(information filtering system)은 다양한 정보원으로부터 이용자의 요구에 적합한 최신 정보를 제공하기 위한 검색시스템의 일종으로, 전자우편이나 웹과 같은 동적인 정보원을 대상으로 이용자의 정적인 정보요구를 비교하여 적합한 정보를 제공하도록 고안된 시스템이다.

SDI에서 시작된 정보필터링은 1958년 Luhn(1958)의 연구에서 시작되었으며, 비순위 불논리 검색모형에 기반을 두고 있다(Callen 1992). Denning (1982)은 정보 제공이 정보 이용자에게 도달한 정보의 통제와 필터링과 같

은 정보의 입수에 초점을 두어야 한다고 지적하면서, 입수된 정보의 주제, 키워드, 또는 본문을 검색함으로써 원하지 않는 정보를 삭제할 수 있어야 한다고 주장하였다. 정보필터링 시스템에서는 이용자의 프로파일 작성이 필수적이므로 이용자 모형화 및 이용자 행태에 관한 연구들이 다수 수행되었다(Malone, Grant, and Turbak 1986; Stadnyk and Kass 1992; Fisher and Stevens 1991; Stevens 1992; Ram 1992; Kay and Kummerfeld 1996; Rodriguez-Mula, Garcia-Monila, and Paepcke 1998; Fidel and Crandall 1998).

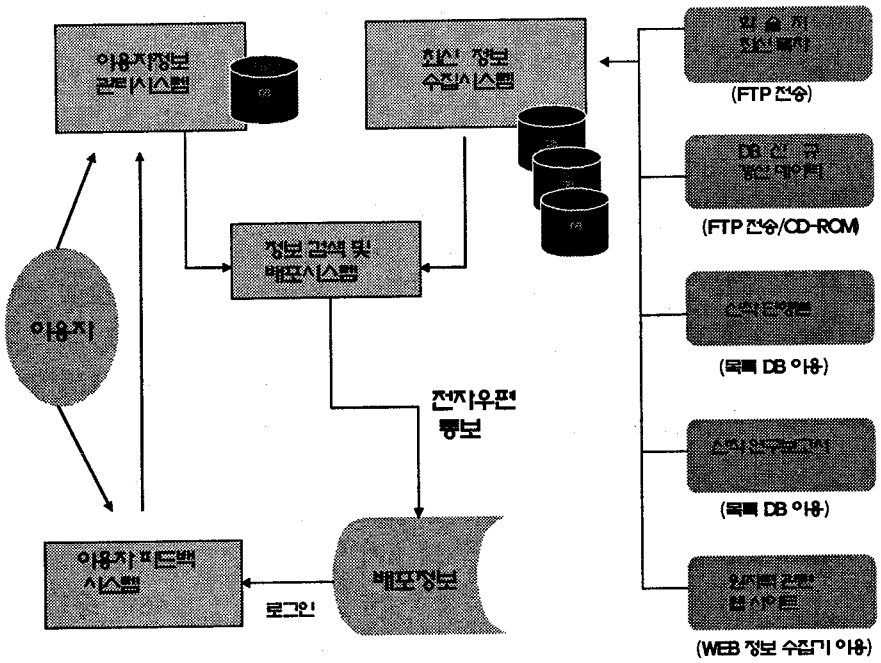
그러나 정보필터링 시스템에서 최신의 정보원들을 대상으로 이용자의 정보요구에 부합하는 정보만을 제공할 수 있는 방법은 이용자의 정보선호도를 그 바탕으로 해야 한다. 따라서 이용자에게 적합한 정보원은 이용자의 피드백 또는 이용자의 정보이용 기록에서 그 실마리를 찾는 것이 중요한 접근점이 된다.

이러한 점에서 이용자가 작성한 질의를 기본으로 일차적인 검색을 수행한 후 이 결과들에 대하여 이용자가 적합, 부적합 여부를 판정하고, 그 판정 결과를 토대로 새로운 질의를 생성함으로써, 검색성능을 향상시킬 수 있는 정보필터링 방안에 대한 연구가 필요하다. 지금가지 정보검색 과정에서 이용자 피드백을 도입하여 검색 성능을 향상시키려는 연구는 피드백 검색 또는 질의확장 등의 주제 아래 계속 수행되어 왔으나(Ide 1971; Rocchio 1971; Salton 1981; Dillon and Desper 1980; Qiu and Frei 1993; Xu and Croft 1996; Greenberg 2001; Mandala, Takenobu, and Tanaka 2000; 박지연 2001), 정보필터링 시스템 환경에서의 연구는 찾아보기 힘들다.

본 연구의 목적은 연구소 환경에서 활용할 수 있는 다양한 정보원을 대상으로 하여 정보필터링 시스템을 구축하고, 이 정보필터링 시스템의 정보검색 성능을 향상시키기 위한 방안을 찾아내는 것이다. 이를 위해 일차 검색 후 이용자의 적합성 피드백을 이용하여 초기 질의를 수정한 다음 수정질의를 사용하여 재검색을 수행하는 실험을 수행하고, 검색 실험 결과의 분석을 통해 검색 성능을 최적화할 수 있는 수정질의식의 패턴을 찾아냄으로써 정보필터링 시스템의 성능을 향상시키고자 한다.

2 이용자 피드백을 사용한 정보필터링 시스템 구현

이 연구에서는 실제 정보환경에서 온라인 피드백에 의한 질의수정 실험이 가능하도록 하기 위하여 이용자의 정보요구를 검색질의로 저장하고, 검색결과



〈그림 1〉 정보필터링 시스템의 구성도

에 대한 이용자의 평가를 반영하는 정보필터링 시스템을 구현하였다.

이 정보필터링 시스템은 〈그림 1〉과 같이 이용자정보 관리시스템, 최신정보 수집시스템, 정보검색 및 배포시스템, 이용자 피드백 시스템 등의 하부시스템으로 구성된다.

이용자정보 관리시스템에서는 이용자 번호, 성명, 소속, 이메일 등의 개인정보와 이용자의 정보요구를 표현한 불논리 질의식 및 제공받기를 원하는 정보원을 입력한다. 또한 웹 정보원의 경우 이용자가 제공받기를 원하는 사이트의 URL과 질의를 입력한다.

최신정보 수집시스템은 정보필터링 시스템에서 제공할 정보원들로부터 최신의 정보원만 선별적으로 수집하여 데이터베이스로 생성하는 하부시스템이다. 신착 학술지 목차는 SWETS의 외국학술지 목차를 FTP를 통하여 파일로 전송받아 시스템에 저장하고, 데이터베이스 신규 갱신분은 시디롬으로 배포한 INIS 데이터베이스의 갱신분을 대상으로 저장한다. 신착 단행본과 신착 연구보고서는 목록 데이터베이스에 입력된 데이터 중에서 최근에 입력되어 정보여과 서비스를 제공하지 않은 정보를 대상으로 하며, 웹자원의 경우 이용자가 입력

한 URL을 가지고 웹크롤러(WebCrawler)를 이용하여 해당 사이트의 웹문서들을 수집한 다음 저장한다.

정보검색 및 배포시스템은 이용자정보 관리시스템에 저장된 불논리 질의를 검색엔진에 적합한 탐색문으로 변환하고, 이 탐색문을 가지고 최신정보 수집시스템에 저장된 최신정보를 이용자정보 관리시스템에 기록된 요청주기별로 검색하여 배포한다. 정보검색시 이용자 프로파일 내에 있는 질의는 Verity에 적합한 탐색문으로 변환되고, 이 탐색문을 Verity 검색엔진을 이용하여 각각의 정보 유형별 데이터베이스에 대하여 검색한 다음 그 결과를 통합하여 저장한다.

검색 후 이 시스템은 검색결과의 배포시에 이용자들에게 특정 프로파일에 해당하는 결과가 처리되어 배포되었음을 전자우편으로 통보하게 된다. 이용자는 통보된 전자우편을 클릭하여 시스템으로 로그인한 다음 자신에게 배포된 결과를 확인할 수 있으며, 개인별로 검색된 각각의 결과에 대한 저장기능을 가진다.

이용자 피드백 시스템은 이용자가 검색된 결과의 서지와 초록을 보고 판정한 적합, 부적합 여부와 질의에서 제외되기를 바라는 키워드를 입력받아 저장하고, 이것을 바탕으로 적합문헌 집단과 부적합문헌 집단의 차이점을 찾아내어 새로 수정된 질의를 만들어 내는 시스템이다.

이용자가 하나의 프로파일에 대하여 검색된 각 검색결과에 대하여 적합문헌 또는 부적합문헌으로 판정을 완료하면 피드백시스템은 우선 적합문헌의 분류기호 집단과 부적합문헌의 분류기호집단을 추출하여 출현빈도순으로 정렬하여 저장한다. 다음으로 적합문헌 집단과 부적합문헌 집단의 디스크립터들을 출현빈도순으로 정렬하여 저장한다. 또한 이용자가 질의에서 제외되기를 바라는 문헌의 키워드를 정렬하여 저장한다.

3 온라인 이용자 피드백을 사용한 정보필터링 실험

3.1 실험의 개요

정보필터링 실험에서는 이용자가 초기에 작성한 정보요구 질의를 정보필터링 시스템을 통하여 검색하여 제공한 다음, 이용자가 그 결과를 보고 시스템에 적합, 부적합의 판정을 피드백하면 그 결과를 토대로 두 가지 방법으로 수정질의를 생성하여 재검색한 다음 초기질의의 판정결과와 비교 분석하였다.

이 실험을 위하여 원자력분야 연구에 종사하는 박사급 전문가 10인을 선정하여 실험목적을 설명하고, 최근의 관심분야에 대한 INIS 디스크립터를 제시받아 불논리 질의를 생성하였다. 〈표 1〉은 각각의 전문가로부터 받은 초기질

〈표 1〉 실험대상 초기질의

전문가	초기질의
1	Alloy AND Deformation
2	Corrosion AND (Pitting OR Condenser)
3	Corrosion AND Zirconium AND Alloy
4	Hydrogen AND Zirconium
5	Hydrogen and Nickel and Alloy
6	Plasma AND Diagnostics AND TOKAMAK
7	Plasma AND Heating AND TOKAMAK
8	Steam Generator AND Tube
9	Stress AND Corrosion AND Cracking
10	Tritium AND Application

의이다.

이 실험에 사용한 문헌집단은 원자력 연구분야의 핵심 데이터베이스인 INIS 데이터베이스에 추가되는 최신정보들을 대상으로 하였다. INIS 데이터베이스는 년간 52회 갱신되며, 갱신되는 데이터는 FTP 전송 또는 CD-ROM을 통해서 제공된다. 이 실험에서는 2001년 갱신분 중 Vol. 32, No. 39 - 46까지 8회분을 사용하였으며, 실험문헌의 총 건수는 11,947건이었다.

이 연구에서는 이용자의 피드백 결과 부적합문헌에서만 사용된 분류기호를 사용한 수정질의나 적합문헌 집단과 부적합 문헌 집단에 동시에 출현한 디스크립터를 제거하고 남은 디스크립터를 이용한 수정질의, 또는 이 두 가지를 함께 사용한 수정질의가 초기질의보다 검색결과가 우수할 것이라는 전제하에 수정질의를 생성, 검색하는 정보필터링 실험을 수행하였다.

수정질의 생성은 1차와 2차로 나누어 두 가지 방법을 사용하였는데, 1차에는 적합문헌 집단과 부적합문헌 집단 양측에 발생하는 분류기호와 디스크립터를 동시에 삭제하고 남은 분류기호와 디스크립터를 사용하여 수정질의를 만드는 '중복제거법'을 사용하였다. 2차에는 적합문헌 집단과 부적합문헌 집단에 동시에 출현한 디스크립터나 분류기호를 적합문헌 집단에서의 출현빈도와 부적합문헌 집단에서의 출현빈도를 비교하여 낮은 쪽을 제거하는 '저빈도제거법'을 사용하였다.

수정질의는 17개 패턴의 불논리 질의 형태로 미리 만든 다음 초기의 질의와 적합문헌 집단 및 부적합문헌 집단의 디스크립터와 분류기호를 활용하여 수정질의를 생성하고, 재검색을 수행하였다.

실험 과정은 다음과 같다.

(1) 이용자의 질의들을 이용자정보 관리시스템에 입력한다.
(2) 최신정보 수집시스템을 통하여 신규 입수된 INIS 데이터를 반입한다.
(3) 정보검색 및 배포시스템을 통하여 각각의 질의를 검색하고 그 결과를 배포한다.
(4) 각 이용자가 배포된 검색결과에 대하여 이용자 피드백 시스템을 통하여 적합, 부적합 판정을 내린다. 이용자가 전혀 관계없다고 판단하는 키워드는 별도로 입력할 수 있도록 하였다.
(5) 이용자정보 관리시스템에서 피드백된 결과에서 적합문헌 집단과 부적합문헌 집단의 디스크립터와 분류기호를 추출하여 따로 저장한다.
(6) 두 집단간에 동시에 발생한 디스크립터와 분류기호를 중복제거법으로 제거한 후 1차로 17개의 수정된 질의를 생성하여 다시 검색을 실시한다.
(7) 두 집단간에 동시에 발생한 디스크립터와 분류기호를 저빈도제거법으로 제거한 후 2차로 17개의 수정된 질의를 생성하여 다시 검색을 실시한다.
(8) 1차 수정질의 17개의 검색결과와 2차 수정질의 17개의 검색결과를 초기 검색결과에 대한 판정결과와 비교한다.

3.2 실험용 수정질의의 생성

이 실험은 최초의 검색결과에 대하여 이용자가 적합, 부적합 판정을 내리면 그 결과를 분석하여 새로운 질의를 만들도록 이루어져 있다. 우선 적합문헌 집단과 부적합문헌 집단의 디스크립터와 분류기호를 별도로 저장하였다. 그 다음 초기질의에 들어 있는 디스크립터를 두 문헌집단의 디스크립터에서 우선 삭제하였다. 다음으로 빈도를 기준으로 중복제거법과 저빈도제거법 두 가지를 사용하여 초기의 질의어와 함께 수정질의에서 사용할 분류기호 및 디스크립터 집단을 생성하였다.

중복제거법을 이용한 수정질의에 사용한 여섯 가지 추가 질의어 유형은 다음과 같으며, 해당되는 불논리식을 괄호안에 제시하였다.

A : 부적합문헌 집단의 분류기호에서 적합문헌 집단의 분류기호를 제거한 후의 부적합문헌 분류기호(NOT A)
B : 적합문헌 집단의 최고빈도 디스크립터(AND B)
C : 적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터를 제거하고 남은 적합문헌 집단 최고빈도 디스크립터(AND C)
D : 부적합문헌 집단의 최고빈도 디스크립터(NOT D)
E : 적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터를 제거

〈표 2〉 17가지의 수정질의식

불논리 연산자 사용	질의식 패턴	비 고
AND 연산자만 사용	OQ AND B	
	OQ AND C	
	OQ AND (B OR C)	OR 사용
AND, NOT 연산자 사용	OQ AND C NOT E	
	OQ AND B NOT A	
	OQ AND C NOT A	
	OQ AND C NOT E NOT A	
	OQ AND (B OR C) NOT A	OR 사용
NOT 연산자만 사용	OQ NOT A	
	OQ NOT D	
	OQ NOT E	
	OQ NOT (D OR E)	OR 사용
NOT, NOT 사용	OQ NOT D NOT A	
	OQ NOT E NOT A	
	OQ NOT (D OR E) NOT A	OR 사용
	OQ NOT F	
	OQ NOT F NOT A	

하고 남은 부적합문헌 집단 최고빈도 디스크립터(NOT E)

F : 이용자가 직접 입력한 부적합 디스크립터 중 최고빈도 디스크립터(NOT F)

저빈도제거법을 이용한 수정질의에 사용한 여섯 가지 추가 질의어 유형은 다음과 같으며, 해당되는 불논리식을 괄호안에 제시하였다.

A : 부적합문헌 집단 분류기호에서 적합문헌 집단 분류기호를 제거한 후의 부적합문헌 분류기호(NOT A)

B : 적합문헌 집단의 최고빈도 디스크립터(AND B)

C′ : 적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터들 중 부적합문헌 집단 디스크립터보다 낮은 빈도의 적합문헌 집단 디스크립터를 제거하고 남은 적합문헌 집단 최고빈도 디스크립터(AND C′)

D : 부적합문헌 집단 최고빈도 디스크립터(NOT D)

E′ : 적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터들 중 적합문헌 집단 디스크립터보다 낮

〈표 3〉 초기질의 10개에 대한 검색결과

초기질의	검색된 문헌수	적합 판정 문헌수	부적합 판정 문헌수	정확률 (%)	재현율 (%)	F 척도
OQ 1	58	12	46	20.7	100	45.9
OQ 2	40	13	27	32.5	100	61.0
OQ 3	44	7	37	15.9	100	38.1
OQ 4	68	25	43	36.8	100	65.4
OQ 5	39	24	15	61.5	100	83.8
OQ 6	39	9	30	23.1	100	49.4
OQ 7	25	12	13	48.0	100	75.0
OQ 8	35	11	24	31.4	100	59.8
OQ 9	29	24	5	82.8	100	94.0
OQ 10	37	11	26	29.7	100	57.9
합 계	414	148	266	35.7	100	64.3
평 균	41.4	14.8	26.6	35.7	100	64.3

은 빈도의 부적합 문헌집단 디스크립터를 제거하고 남은 부적합문헌 집단 최고빈도 디스크립터(NOT E′)

F : 이용자가 직접 입력한 부적합 디스크립터 중 최고빈도 디스크립터(NOT F)

앞의 각 방법에 대하여 여섯 가지 질의어 유형을 초기질의(이하 OQ)에 결합시켜 〈표 2〉와 같이 17가지의 새로운 수정질의식을 형성하였다.

3.3 실험결과 평가척도

이 실험에 사용된 전체 질의의 수는 초기질의 10개에 대하여 중복제거법 및 저빈도제거법 각각 17개의 수정질의를 생성하여 총 360개이다. 10개의 초기질의에 대하여 중복제거법에 의한 OQ와 수정질의 17개 및 저빈도제거법에 의한 OQ와 수정질의 17개에 대하여 각각의 검색 문헌 수, 적합 문헌 수, 부적합 문헌 수와 함께 검색 효율척도인 정확률, 재현율, 배제율 및 F 척도를 구하였다.

F 척도는 검색효율을 나타내는 복합척도 중 하나이며, 다음과 같은 식으로 표현한다.

$$F_{\beta}(r, p) = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

이 실험에서 Van Rijsbergen이 원

〈표 4〉 OQ와 중복제거법에 의한 수정질의의 연산자별 검색결과

질의 형태		검색 문헌수	적합 문헌수	부적합 문헌수	정확률 (%)	재현율 (%)	배제율 (%)	*F* 척도
OQ		414	148	266	35.7	100.0		64.3
AND 연산자만 사용	OQ AND B	344	131	213	38.1	88.5	19.9	62.9
	OQ AND C	94	75	19	79.8	50.7	92.9	57.1
	OQ AND (B OR C)	354	140	214	39.5	94.6	19.5	66.2
AND, NOT 연산자 사용	OQ AND B NOT A	325	131	194	40.3	88.5	27.1	64.7
	OQ AND C NOT A	93	75	18	80.6	50.7	93.2	57.2
	OQ AND C NOT E	81	70	11	86.4	47.3	95.9	54.9
	OQ AND C NOT E NOT A	80	70	10	87.5	47.3	96.2	55.1
	OQ AND (B OR C) NOT A	335	140	195	41.8	94.6	26.7	68.1
NOT 연산자만 사용	OQ NOT A	371	148	223	39.9	100.0	16.2	68.3
	OQ NOT D	36	10	26	27.8	6.8	90.2	8.8
	OQ NOT E	307	140	167	45.6	94.6	37.2	71.1
	OQ NOT (D OR E)	27	9	18	33.3	6.1	93.2	8.1
NOT, NOT 연산자 사용	OQ NOT D NOT A	26	10	16	38.5	6.8	94.0	9.1
	OQ NOT E NOT A	272	140	132	51.5	94.6	50.4	75.2
	OQ NOT (D OR E) NOT A	17	9	8	52.9	6.1	97.0	8.4
	OQ NOT F	296	128	168	43.2	86.5	36.8	66.1
	OQ NOT F NOT A	270	128	142	47.4	86.5	46.6	69.0
수정질의 평균		196	92	104	51.4	61.8	60.8	51.2

래 제시한 *E4* 척도 대신 텍스트범주화 성능평가에 주로 사용하는 *F* 척도 (Yang 1999)를 사용한 이유는 정확률과 재현율의 가중치에 차이를 줄 수 있고, 큰 값이 높은 검색효율을 나타내도록 하기 위함이다. 이 실험에서는 OQ의 판정결과 중 적합문헌의 재현과 부적합문헌의 제거에 중점을 두었기 때문에, 재현율이 정확률보다 중요하다고 판단해서 $\beta=1.5$로 정하였다.

3.4 실험결과 및 분석

3.4.1 초기질의에 대한 검색결과

〈표 5〉 중복제거법에 의한 수정질의의 재현율에 따른 구분

질의 형태		검색 문헌수	적합 문헌수	부적합 문헌수	재현율 (향상률)	정확률 (향상률)	배제율	F 척도 (향상률)
OQ6		414	148	266	100.0	35.7		64.3
재현율 80% 이상	OQ NOT A	371	148	223	100.0 (0)	39.9 (11.8)	16.2	68.3 (6.2)
	OQ NOT E NOT A	272	140	132	94.6 (-5.4)	51.5 (44.3)	50.4	75.2 (17.0)
	OQ NOT E	307	140	167	94.6 (-5.4)	45.6 (27.7)	37.2	71.1 (10.6)
	OQ AND (B OR C) NOT A	335	140	195	94.6 (-5.4)	41.8 (17.1)	26.7	68.1 (5.9)
	OQ AND (B OR C)	354	140	214	94.6 (-5.4)	39.5 (10.6)	19.5	66.2 (3.0)
	OQ AND B NOT A	325	131	194	88.5 (-11.5)	40.3 (12.9)	27.1	64.7 (0.6)
	OQ AND B	344	131	213	88.5 (-11.5)	38.1 (6.7)	19.9	62.9 (-2.2)
	OQ NOT F NOT A	270	128	142	86.5 (-13.5)	47.4 (32.8)	46.6	69.0 (7.3)
	OQ NOT F	296	128	168	86.5 (-13.5)	43.2 (21.0)	36.8	66.1 (2.8)
재현율 80% 미만	OQ AND C NOT A	93	75	18	50.7 (-49.3)	80.6 (125.8)	93.2	57.2 (-11.0)
	OQ AND C	94	75	19	50.7 (-49.3)	79.8 (123.5)	92.9	57.1 (-11.2)
	OQ AND C NOT E NOT A	80	70	10	47.3 (-52.7)	87.5 (145.1)	96.2	55.1 (-14.3)
	OQ AND C NOT E	81	70	11	47.3 (-52.7)	86.4 (142.0)	95.9	54.9 (-14.6)
	OQ NOT D NOT A	26	10	16	6.8 (-93.2)	38.5 (7.8)	94.0	9.1 (-85.8)
	OQ NOT D	36	10	26	6.8 (-93.2)	27.8 (-22.1)	90.2	8.8 (-86.3)
	OQ NOT (D OR E) NOT A	17	9	8	6.1 (-93.9)	52.9 (48.2)	97.0	8.4 (-86.9)
	OQ NOT (D OR E)	27	9	18	6.1 (-93.9)	33.3 (-6.7)	93.2	8.1 (-87.4)

이 실험에서 가장 기초가 되는 데이터는 초기질의 10개에 대한 이용자의 적합성 판정 결과와 평균재현율, 평균 정확률 및 평균 F 척도 값으로서 〈표 3〉과 같다.

10개의 질의 모두가 재현율이 100인 이유는 초기질의에 대한 검색결과에 대하여 이용자가 적합하다고 판정한 문헌의 수를 전체 적합문헌의 수로 하여 수정질의에 대한 재현율을 산출하였기 때문이다.

3.4.2 수정질의에 대한 비교 분석

3.4.2.1 중복제거법에 의한 수정질의 전체 비교 분석

OQ와 중복제거법에 의한 수정질의 17개 전체의 연산자별 검색결과는 〈표 4〉와 같다. 수정질의의 검색결과에 대하여 재현율 측면에서 80% 이상을 충족시키는 수정질의와 그 이하의 재현율을 갖는 수정질의가 확연히 구분되었다. 〈표 5〉는 재현율을 기준으로 수정질의를 구분하여 나타낸 것이다.

중복제거법을 이용한 수정질의의 검색결과에서 나타난 특징은 다음과 같다.

(1) 재현율을 기준으로 17개의 수정질의가 재현율 80% 이상인 수정질의 집단과, 47%~51%인 수정질의 집단, 10% 미만의 수정질의 집단으로 구분되었다.

(2) 배제율을 기준으로 90%가 넘는 수정질의 집단은 재현율이 51%미만의 수정질의 집단과 일치하며, 배제율이 90% 미만인 수정질의 집단은 재현율이 80% 이상인 집단과 일치하였다.

(3) 재현율이 80% 이상인 수정질의 집단(배제율이 51% 미만인 수정질의 집단)은 요소 B를 AND로 결합한 경우, 요소 A를 NOT으로 결합한 경우, 요소 E를 NOT으로 결합한 경우, 요소 F를 NOT으로 결합한 경우이다. 이 수정질의 집단은 OQ에 비하여 정확률이 모두 높게 나타났다.

(4) 재현율이 47%~51%인 수정질의 집단(배제율이 90% 이상인 수정질의 집단중 일부)은 요소 C를 AND로 결합한 경우이다. 이 집단은 집단은 OQ에 비하여 정확률이 모두 2배 이상 향상된 결과를 보인다.

(5) 재현율이 10%미만인 수정질의 집단(배제율이 90% 이상인 수정질의 집단중 일부)은 요소 D를 NOT으로 결합한 경우이다.

(6) 재현율 100%인 경우는 요소 A를 NOT으로 결합한 수정질의 한 가지이다.

3.4.2.2 저빈도제거법에 의한 수정질의 전체 비교분석

OQ와 저빈도제거법에 의한 수정질의 17가지 전체의 검색결과는 〈표 6〉과 같다. 저빈도제거법을 이용한 수정질의

〈표 6〉 저빈도제거법에 의한 수정질의의 연산자별 검색결과 비교

질의 형태		검색 문헌수	적합 문헌수	부적합 문헌수	정확률 (%)	재현율 (%)	배제율 (%)	F 척도
	OQ	414	148	266	35.7	100.0		64.3
AND 연산자만 사용	OQ AND B	344	131	213	38.1	88.5	19.9	62.9
	OQ AND C′	308	127	181	41.2	85.8	32.0	64.4
	OQ AND (B OR C′)	345	131	214	38.0	88.5	19.5	62.8
AND, NOT 연산자 사용	OQ AND B NOT A	325	131	194	40.3	88.5	27.1	64.7
	OQ AND C′ NOT E′	65	38	27	58.5	25.7	89.8	31.0
	OQ AND C′ NOT A	290	127	163	43.8	85.8	38.7	66.3
	OQ AND C′ NOT E′ NOT A	61	38	23	62.3	25.7	91.4	31.3
	OQ AND (B OR C′) NOT A	326	131	195	40.2	88.5	26.7	64.6
NOT 연산자만 사용	OQ NOT A	371	148	223	39.9	100.0	16.2	68.3
	OQ NOT D	36	10	26	27.8	6.8	90.2	8.8
	OQ NOT E′	93	51	42	54.8	34.5	84.2	38.9
	OQ NOT (D OR E′)	27	9	18	33.3	6.1	93.2	8.1
NOT, NOT 연산자 사용	OQ NOT D NOT A	26	10	16	38.5	6.8	94.0	9.1
	OQ NOT E′ NOT A	80	51	29	63.8	34.5	89.1	40.1
	OQ NOT (D OR E′) NOT A	18	9	9	50.0	6.1	96.6	8.3
	OQ NOT F	296	128	168	43.2	86.5	36.8	66.1
	OQ NOT F NOT A	270	128	142	47.4	86.5	46.6	69.0
수정질의 평균		잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식

의 검색결과는 재현율이 80% 이상인 수정질의와 40% 미만인 수정질의로 뚜렷히 구분되었다. 재현율이 80%이상인 수정질의들은 모두 다 배제율이 40% 미만이었으며, 재현율이 40%미만인 수정질의들은 모두 배제율이 80% 이상을 나타내었다. 〈표 7〉은 저빈도제거법을 이용한 수정질의 17개의 검색결과를 재현율 80%를 기준으로 구분하여 재배열한 것이다.

저빈도제거법을 이용한 수정질의의 검색결과에서 나타난 특징은 다음과 같다.

(1) 재현율을 기준으로 17개의 수정질의가 재현율 80% 이상인 수정질의

〈표 7〉 저빈도제거법에 의한 수정질의의 재현율에 따른 구분

질의 형태		검색 문헌수	적합 문헌수	부적합 문헌수	재현율 (향상률)	정확률 (향상률)	배제율	F 척도 (향상률)
OQ		414	148	266	100.0	35.7		64.3
재현율 80% 이상	OQ NOT A	371	148	223	100.0 (0)	39.9 (11.8)	16.2	68.3 (6.2)
	OQ AND B NOT A	325	131	194	88.5 (-11.5)	40.3 (12.9)	27.1	64.7 (0.6)
	OQ AND (B OR C′) NOT A	326	131	195	88.5 (-11.5)	40.2 (12.6)	26.7	64.6 (0.5)
	OQ AND B	344	131	213	88.5 (-11.5)	38.1 (6.7)	19.9	62.9 (-2.2)
	OQ AND (B OR C′)	345	131	214	88.5 (-11.5)	38.0 (6.4)	19.5	62.8 (-2.3)
	OQ NOT F NOT A	270	128	142	86.5 (-13.5)	47.4 (32.8)	46.6	69.0 (7.3)
	OQ NOT F	296	128	168	86.5 (-13.5)	43.2 (21.0)	36.8	66.1 (2.8)
	OQ AND C′ NOT A	290	127	163	85.8 (-14.2)	43.8 (22.7)	38.7	66.3 (3.1)
	OQ AND C′	308	127	181	85.8 (-14.2)	41.2 (15.4)	32.0	64.4 (0.2)
재현율 80% 미만	OQ NOT E′ NOT A	80	51	29	34.5 (-65.5)	63.8 (78.7)	89.1	40.1 (-37.6)
	OQ NOT E′	93	51	42	34.5 (-65.5)	54.8 (53.5)	84.2	38.9 (-39.5)
	OQ AND C′ NOT E′ NOT A	61	38	23	25.7 (-74.3)	62.3 (74.5)	91.4	31.3 (-51.3)
	OQ AND C′ NOT E′	65	38	27	25.7 (-74.3)	58.5 (63.9)	89.8	31.0(-51 .8)
	OQ NOT D NOT A	26	10	16	6.8 (-93.2)	38.5 (7.8)	94.0	9.1 (-85.8)
	OQ NOT D	36	10	26	6.8 (-93.2)	27.8 (-22.1)	90.2	8.8 (-86.3)
	OQ NOT (D OR E′) NOT A	18	9	9	6.1 (-93.9)	50.0 (40.1)	96.6	8.3 (-87.1)
	OQ NOT (D OR E′)	27	9	18	6.1 (-93.9)	33.3 (-6.7)	93.2	8.1 (-87.4)

집단과, 25%~35%인 수정질의 집단, 10% 미만의 수정질의 집단으로 구분되었다.

(2) 배제율을 기준으로 80%가 넘는 수정질의 집단은 재현율이 35% 미만의 수정질의 집단과 일치하며, 배제율이 39% 미만인 수정질의 집단은 재현율이 80% 이상인 집단과 일치하였다.

(3) 재현율이 80% 이상인 수정질의 집단(배제율이 39% 미만인 수정질의 집단)은 요소 C′을 AND로 결합한 경우이다. 이 수정질의 집단은 OQ에 비하여 정확률이 모두 높게 나타났다.

(4) 재현율이 35% 미만인 수정질의 집단은 요소 E′을 NOT으로 결합한 수정 질의들이며, 재현율이 80% 이상인 수정질의에 비하여 배제율이 2.6배 이상으로 현저하게 높았다.

(5) 재현율이 10%미만인 수정질의 집단은 요소 D와 E′을 NOT으로 결합한 경우로서, 배제율이 93.3% 이상으로 가장 높게 나타났다.

3.4.3 실험결과의 분석

정보필터링 시스템에서 최적의 질의는 가능한 한 적합문헌을 초기검색 수준으로 유지하면서 부적합문헌을 검색하지 않도록 초기 재현율 수준에서 배제율을 높이는 질의라 할 수 있다. 여기에서는 OQ에 결합가능한 수정질의 결합요소 6가지를 설정한 후 결합요소의 추출방법에 따라 중복제거법에 의한 수정질의 17개와 저빈도제거법에 따른 수정질의 10개의 검색결과를 기초로 재현율, 정확률, 배제율 각각에 대하여 검색효율을 비교, 분석하여 최적의 수정질의식을 도출하고자 하였다.

검색성능이 가장 높은 수정질의식을 찾기 위한 방법으로는 기준 검색성능 척도 향상률 1위에는 2점을 주고, 그 외의 검색성능 척도는 향상률 1위에 1점을 부여하여 합산하는 방식과 4가지 검색성능 척도 향상률을 모두 더한 후 4로 나눈 평균향상률 값을 구하여 비교하는 방식을 사용하였다.

향상률 순위척도=4가지 척도의 순위향상률

$$\text{평균향상률} = \frac{\text{4가지 척도의 향상률의 합}}{4}$$

OQ와 수정질의의 검색성능 향상률을 비교하기 위하여 OQ의 검색성능은 배제율을 0으로 설정한 후 값을 구하였는데, 그 이유는 초기 검색결과의 부적합 문헌이 전체 부적합문헌 총 수이며, 검색되지 않은 부적합 문헌이 0건이라는 전제로 실험을 수행하였기 때문이다.

3.4.3.1 재현율을 중심으로 한 최적의 수정질의식

이 실험에서는 초기검색결과에 대한 적합성 판정 결과 적합문헌으로 판정된 문헌을 전체 적합문헌으로 정했기 때문

〈표 8〉 재현율 80% 이상인 수정질의식의 검색성능 향상률 비교

질의 형태		검색 문헌수	적합 문헌수	부적합 문헌수	재현율 (향상률)	정확률 (향상률)	배제율	F 척도 (향상률)	순위척도값	평균 향상률
OQ		414	148	266	100.0	35.7	0	64.3		
중복 제거법	OQ NOT A	371	148	223	100.0 (0)	39.9 (11.8)	16.2	68.3 (6.2)	2	8.6
	OQ NOT E NOT A	272	140	132	94.6 (-5.4)	51.5 (44.3)	50.4	75.2 (17.0)	3	26.6
	OQ NOT E	307	140	167	94.6 (-5.4)	45.6 (27.7)	37.2	71.1 (10.6)	0	17.6
	OQ AND (B OR C) NOT A	335	140	195	94.6 (-5.4)	41.8 (17.1)	26.7	68.1 (5.9)	0	11.1
	OQ AND (B OR C)	354	140	214	94.6 (-5.4)	39.5 (10.6)	19.5	66.2 (3.0)	0	6.9
	OQ AND B NOT A	325	131	194	88.5 (-11.5)	40.3 (12.9)	27.1	64.7 (0.6)	0	7.3
	OQ AND B	344	131	213	88.5 (-11.5)	38.1 (6.7)	19.9	62.9 (-2.2)	0	3.2
	OQ NOT F NOT A	270	128	142	86.5 (-13.5)	47.4 (32.8)	46.6	69.0 (7.3)	0	18.3
	OQ NOT F	296	128	168	86.5 (-13.5)	43.2 (21.0)	36.8	66.1 (2.8)	0	11.8
저빈도 제거법	OQ AND (B OR C′) NOT A	326	131	195	88.5 (-11.5)	40.2 (12.6)	26.7	64.6 (0.5)	0	7.1
	OQ AND (B OR C′)	345	131	214	88.5 (-11.5)	38.0 (6.4)	19.5	62.8 (-2.3)	0	3.0
	OQ AND C′ NOT A	290	127	163	85.8 (-14.2)	43.8 (22.7)	38.7	66.3 (3.1)	0	12.6
	OQ AND C′	308	127	181	85.8 (-14.2)	41.2 (15.4)	32.0	64.4 (0.2)	0	8.4

에 초기검색결과의 재현율을 100으로 하였으며, 두 가지 실험결과 재현율이 80%이상인 수정질의 집단과 그 이하인 수정질의 집단이 확연히 구분되었다. 재현율을 중심으로 한 최적의 수정질의식을 찾기 위하여 수정질의의 검색결과에 대한 재현율을 중심으로 재현율이 80%이상인 수정질의들을 다른 검색성능 척도들과 비교하여 재현율 중심의 최적의 수정질의식을 찾고자 하였다. 〈표 8〉은 재현율 80%이상인 수정질의들의 검색성능 향상률을 비교한 표이다.

〈표 9〉 재현율 80% 이상인 수정질의식 중 요소 A가 없는 수정질의식의 검색성능 향상률 비교

질의 형태	검색 문헌수	적합 문헌수	부적합 문헌수	재현율 (향상률)	정확률 (향상률)	배제율	F척도 (향상률)	순위척 도값	평균 향상률
OQ	414	148	266	100.0	35.7		64.3		
OQ NOT E	307	140	167	94.6 (-5.4)	45.6 (27.7)	37.2	71.1 (10.6)	5	17.6
OQ AND (B OR C)	354	140	214	94.6 (-5.4)	39.5 (10.6)	19.5	66.2 (3.0)	2	6.9
OQ AND B	344	131	213	88.5 (-11.5)	38.1 (6.7)	19.9	62.9 (-2.2)	0	3.2
OQ NOT F	296	128	168	86.5 (-13.5)	43.2 (21.0)	36.8	66.1 (2.8)	0	11.8
OQ AND (B OR C′)	345	131	214	88.5 (-11.5)	38.0 (6.4)	19.5	62.8 (-2.3)	0	3.0
OQ AND C′	308	127	181	85.8 (-14.2)	41.2 (15.4)	32.0	64.4 (0.2)	0	8.4

재현율 80% 이상의 수정질의 13개 중 순위척도와 평균향상률 모두 수정질의 OQ NOT E NOT A가 모두 가장 높은 값을 보여 주었다. 이 수정질의는 재현율을 제외한 정확률, 배제율, F 척도에서 검색성능 향상률이 1위를 나타내어 3점으로 재현율만 1위인 OQ NOT A보다 높았다. 또한 평균향상률에서도 26.6%를 나타내어 향상률이 가장 높은 수정질의임이 나타났다.

단순히 재현율이 가장 높은 수정질의는 OQ NOT A로서 적합문헌 148건이 모두 검색되기는 하였으나, 정확률과 배제율이 높지 않았기 때문에 최적의 수정질의가 되지 못하였다.

분류기호가 없는 정보원을 대상으로 하는 검색시스템에서는 요소 A를 사용할 수 없다. 〈표 9〉는 재현율 80% 이상의 수정질의들 중 요소 A가 결합되지 않은 수정질의식들의 검색성능 향상률을 비교한 표이다.

여기에서는 수정질의 OQ NOT E가 순위척도는 5점, 평균향상률은 17.6%로 두 방법에서 1위를 나타냈다.

이상의 분석결과를 볼 때 재현율을 중심으로 한 수정질의식 중에서는 분류기호가 있는 정보를 다루는 정보필터링 시스템의 경우 요소 E(적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터를 제거하고 남은 부적합문헌 집단 최고빈도 디스크립터)와 요소 A(부적합문헌 집단의 분류기호에서 적합문헌 집단의 분류기호를 제거한 후의 부적합문헌 분류기호)를 NOT으로 결합한 OQ NOT E NOT A가 최적의 수정질의식이고, 분류기호가 없는 정보를 다루는

〈표 10〉 정확률 60% 이상인 수정질의식의 검색성능 향상률 비교

질의 형태	검색 문헌수	적합 문헌수	부적합 문헌수	정확률 (향상률)	재현율 (향상률)	배제율	F 척도 (향상률)	순위 척도값	평균 향상률
OQ	414	148	266	35.7	100.0		64.3		
OQ AND C NOT E NOT A	80	70	10	87.5 (145.1)	47.3 (-52.7)	96.2	55.1 (-14.3)	3	43.6
OQ AND C NOT E	81	70	11	86.4 (142.0)	47.3 (-52.7)	95.9	54.9 (-14.6)	0	42.7
OQ AND C NOT A	93	75	18	80.6 (125.8)	50.7 (-49.3)	93.2	57.2 (-11.0)	2	39.7
OQ AND C	94	75	19	79.8 (123.5)	50.7 (-49.3)	92.9	57.1 (-11.2)	1	39.0
OQ NOT E′ NOT A	80	51	29	63.8 (78.7)	34.5 (-65.5)	89.1	40.1 (-37.6)	0	16.2
OQ AND C′ NOT E′ NOT A	61	38	23	62.3 (74.5)	25.7 (-74.3)	91.4	31.3 (-51.3)	0	10.1

수정질의식 중에서는 OQ NOT E가 최적의 수정질의식임이 밝혀졌다.

3.4.3.2 정확률을 중심으로 한 최적의 수정질의식

중복제거법을 사용한 수정질의에서 정확률이 80% 이상으로 높은 수정질의들은 재현율 80% 미만인 수정질의들 중에 4개가 있다. 그러나 저빈도제거법을 사용한 수정질의에는 정확률이 80% 이상인 수정질의가 하나도 없으며, 가장 높은 정확률이 63.8%이다. 저빈도제거법에 의한 수정질의는 중복제거법에 비하여 정확률이 상대적으로 낮지만 다른 검색성능 척도에서는 높은 향상률을 보일 수 있으므로, 정확률을 중심으로 한 분석에서는 중복제거법에서 정확률이 80%이상인 4개의 수정질의와 저빈도제거법에서 정확률이 60% 이상인 수정질의 2개를 함께 분석하였다.

〈표 10〉은 정확률 60% 이상인 수정질의의 검색성능 향상률을 나타낸 것이다. 여기에서는 수정질의식 OQ AND C NOT E NOT A가 순위척도와 평균향상률 모두에서 1위를 나타내었다. 그러나 평균향상률에 있어서는 순위척도에서 0점을 나타낸 OQ AND C NOT E의 평균향상률 42.7%와 큰 차이가 없었다.

6개의 분석대상 수정질의중 분류기호 NOT A를 사용하지 않은 수정질의는 2개가 있으며, 두 수정질의의 검색성능 향상률은 〈표 11〉과 같다.

〈표 11〉 정확률 60% 이상인 수정질의식 중 요소 A가 없는 수정질의식의 검색성능 향상률 비교

질의 형태	검색 문헌수	적합 문헌수	부적합 문헌수	정확률 (향상률)	재현율 (향상률)	배제율	F척도 (향상률)	순위 척도값	평균 향상률
OQ	414	148	266	35.7	100.0		64.3		50.0
OQ AND C NOT E	81	70	11	86.4 (142.0)	47.3 (-52.7)	95.9	54.9 (-14.6)	3	42.7
OQ AND C	94	75	19	79.8 (123.5)	50.7 (-49.3)	92.9	57.1 (-11.2)	2	39.0

수정질의 OQ AND C NOT E가 OQ AND C와 순위척도와 평균향상률에서 비슷한 향상률을 보이고 있다.

이상의 분석결과를 볼 때 정확률을 중심으로 한 수정질의식 중에서는 분류기호가 있는 정보를 다루는 정보필터링 시스템에서는 요소 C(적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터를 제거하고 남은 적합문헌 집단 최고빈도 디스크립터)를 AND로 결합하고, 요소 E(적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터를 제거하고 남은 부적합문헌 집단 최고빈도 디스크립터)와 요소 A(부적합문헌 집단의 분류기호에서 적합문헌 집단의 분류기호를 제거한 후의 부적합문헌 분류기호)를 NOT으로 결합한 OQ AND C NOT E가 최적의 수정질의식이고, 분류기호가 없는 정보를 다루는 수정질의식 중에서는 OQ AND C가 최적의 수정질의식임이 밝혀졌다.

3.4.3.3 배제율을 중심으로 한 최적의 수정질의식

정보필터링에서 배제율은 시스템의 성능을 나타내는 중요한 척도이다. 중복제거법과 저빈도제거법 모두에서 배제율이 80%이상인 수정질의는 14가지이며, 이 수정질의들의 재현율은 모두 80% 이하이었다.

〈표 12〉는 배제율 80% 이상인 수정질의의 검색성능 향상률을 비교한 표이다. 순위척도에서는 수정질의식 OQ NOT (D OR E) NOT A와 OQ AND C NOT A가 2점으로 1위인 반면, 평균향상률을 중심으로 한 수정질의식에서는 OQ AND C NOT E NOT A가 43.6%로 1위를 나타내었다.

배제율이 80% 이상인 수정질의식 중 요소 A를 사용하지 않은 수정질의식은 7개이며, 〈표 13〉과 같다. 수정질의 OQ AND C NOT E는 순위척도와 평균향상률 모두에서 1위를 나타내어 배제율을 중요시하는 정보필터링 시스템

〈표 12〉 배제율 80% 이상인 수정질의식의 검색성능 향상률 비교

질의 형태	검색 문헌수	적합 문헌수	부적합 문헌수	배제율	재현율 (향상률)	정확률 (향상률)	F 척도 (향상률)	순위 척도값	평균 향상률
OQ	414	148	266		100.0	35.7	64.3		
OQ NOT (D OR E) NOT A	17	9	8	97.0	6.1 (-93.9)	52.9 (48.2)	8.4 (-86.9)	2	-8.9
OQ NOT (D OR E′) NOT A	18	9	9	96.6	6.1 (-93.9)	50.0 (40.1)	8.3 (-87.1)	0	-11.1
OQ AND C NOT E NOT A	80	70	10	96.2	47.3 (-52.7)	87.5 (145.1)	55.1 (-14.3)	1	43.6
OQ AND C NOT E	81	70	11	95.9	47.3 (-52.7)	86.4 (142.0)	54.9 (-14.6)	0	42.7
OQ NOT D NOT A	26	10	16	94.0	6.8 (-93.2)	38.5 (7.8)	9.1 (-85.8)	0	-19.3
OQ AND C NOT A	93	75	18	93.2	50.7 (-49.3)	80.6 (125.8)	57.2 (-11.0)	2	39.7
OQ NOT (D OR E′)	27	9	18	93.2	6.1 (-93.9)	33.3 (-6.7)	8.1 (-87.4)	0	-23.7
OQ NOT (D OR E)	27	9	18	93.2	6.1 (-93.9)	33.3 (-6.7)	8.1 (-87.4)	0	-23.7
OQ AND C	94	75	19	92.9	50.7 (-49.3)	79.8 (123.5)	57.1 (-11.2)	1	39.0
OQ AND C′ NOT E′ NOT A	61	38	23	91.4	25.7 (-74.3)	62.3 (74.5)	31.3 (-51.3)	0	10.1
OQ NOT D	36	10	26	90.2	6.8 (-93.2)	27.8 (-22.1)	8.8 (-86.3)	0	-27.9
OQ AND C′ NOT E′	65	38	27	89.8	25.7 (-74.3)	58.5 (63.9)	31.0(-5 1.8)	0	6.9
OQ NOT E′ NOT A	80	51	29	89.1	34.5 (-65.5)	63.8 (78.7)	40.1 (-37.6)	0	16.2
OQ NOT E′	93	51	42	84.2	34.5 (-65.5)	54.8 (53.5)	38.9 (-39.5)	0	8.2

에서 분류기호가 없는 정보원의 필터링에 최적의 수정질의식임을 알 수 있다.

3.4.4 검색성능 척도별 최적의 수정질의식 분석

이제까지 위에서 분석하여 도출한 최적의 수정질의식을 각 중심이 되는 검색성능 척도별로 나타내면 〈표 14〉와

〈표 13〉 배제율 80% 이상인 수정질의식 중 요소 A가 없는 수정질의식의 검색 성능 향상률 비교

질의 형태	검색 문헌수	적합 문헌수	부적합 문헌수	배제율	재현율 (향상률)	정확률 (향상률)	F 척도 (향상률)	순위 척도값	평균 향상률
OQ	414	148	266		100.0	35.7	64.3		
OQ AND C NOT E	81	70	11	95.9	47.3 (-52.7)	86.4 (142.0)	54.9 (-14.6)	3	42.7
OQ NOT (D OR E′)	27	9	18	93.2	6.1 (-93.9)	33.3 (-6.7)	8.1 (-87.4)	0	-23.7
OQ NOT (D OR E)	27	9	18	93.2	6.1 (-93.9)	33.3 (-6.7)	8.1 (-87.4)	0	-23.7
OQ AND C	94	75	19	92.9	50.7 (-49.3)	79.8 (123.5)	57.1 (-11.2)	2	39.0
OQ NOT D	36	10	26	90.2	6.8 (-93.2)	27.8 (-22.1)	8.8 (-86.3)	0	-27.9
OQ AND C′ NOT E′	65	38	27	89.8	25.7 (-74.3)	58.5 (63.9)	31.0(-51.8)	0	6.9
OQ NOT E′	93	51	42	84.2	34.5 (-65.5)	54.8 (53.5)	38.9 (-39.5)	0	8.2

같다.

재현율을 중심척도로 사용할 최적의 수정질의식은 분류기호 요소 A를 NOT으로 연결하여 사용할 경우에는 수정질의식 OQ NOT E NOT A이고, 분류기호 요소 A를 사용하지 않을 경우에는 수정질의식 OQ NOT E이다

따라서 재현율 중심의 정보필터링 시스템에서는 초기질의 OQ에 적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터를 제거하고 남은 부적합문헌 집단 최고빈도 디스크립터(E)를 NOT로 연결한 OQ NOT E가 최적의 수정질의식임을 알 수 있다.

정확률을 중심척도로 사용한 수정질의의 경우 분류기호 요소 A를 NOT으로 연결하여 사용할 경우에 최적의 수정질의식은 OQ AND C NOT E NOT A이고, 분류기호 요소 A를 사용하지 않을 경우에는 최적의 수정질의식이 OQ AND C NOT E이다.

따라서 정확률 중심의 정보필터링 시스템에서는 OQ에 적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터를 제거하고 남은 적합문헌 집단 최고빈도 디스크립터(C)를 AND로 연결하고, 적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터를 제거하고 남은 부적합문헌 집단 최고빈도 디스크립터(E)를 NOT로 연결하는 것이

〈표 14〉 중심 검색성능 척도별 최적의 수정질의식

중심 검색성능 척도		재 현 율	정 확 률	배 제 율
분류기호 요소(A) 사용	순위척도 순에 의한 최적 수정질의	OQ NOT E NOT A	OQ AND C NOT E NOT A	OQ NOT (D OR E) NOT A OQ AND C NOT A
	평균 향상률에 의한 최적 수정질의	OQ NOT E NOT A	OQ AND C NOT E NOT A	OQ AND C NOT E NOT A
분류기호 요소(A) 미사용	순위척도 순에 의한 최적 수정질의	OQ NOT E	OQ AND C NOT E	OQ AND C NOT E
	평균 향상률에 의한 최적 수정질의	OQ NOT E	OQ AND C NOT E	OQ AND C NOT E
공통요소		OQ NOT E	OQ AND C NOT E	OQ AND C NOT E

최적의 수정질의식임을 알 수 있다.

배제율을 중심척도로 사용한 수정질의의 경우 분류기호 요소 A를 NOT으로 연결하여 사용할 경우에 순위척도에 의한 최적의 수정질의식은 OQ NOT (D OR E) NOT A와 OQ AND C NOT A 두 가지이고, 분류기호 요소 A를 사용하지 않을 경우에는 순위척도와 평균향상률에 의한 최적의 수정질의식은 OQ AND C NOT E이다.

따라서 배제율 중심의 정보필터링 시스템에서도 정확률 중심의 수정질의와 마찬가지로 초기질의 OQ에 적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터를 제거하고 남은 적합문헌 집단 최고빈도 디스크립터(C)를 AND로 연결하고, 적합문헌 집단과 부적합문헌 집단에 동시 출현한 디스크립터를 제거하고 남은 부적합문헌 집단 최고빈도 디스크립터(E)를 NOT로 연결하는 OQ AND C NOT E가 최적의 수정질의식임을 알 수 있다.

4 결 론

이 연구에서는 정보필터링 시스템의 성능을 높이기 위하여 이용자의 질의에 대하여 최신 정보원들을 검색하고, 그 결과에 대한 이용자의 적합성 평가를 반영하는 정보필터링 시스템을 구현한 후, 이 정보필터링 시스템 상에서 온라인 이용자 피드백을 통해 초기질의를 수정한 다음 수정질의에 의하여 재검색하는 실험을 수행하였다. 실험 결과 정보필터링 성능을 최적화할 수 있는 수정질의식을 제시하였다.

수정질의 생성의 첫 번째 방법은 적합문헌 집단과 부적합문헌 집단 양측에 발생하는 분류기호와 디스크립터를 동

시에 삭제하고 남은 분류기호와 디스크립터를 사용하여 수정질의를 만드는 '중복제거법'을 사용하였고, 두 번째 방법은 적합문헌 집단과 부적합문헌 집단에 동시에 출현한 디스크립터나 분류기호를 적합문헌 집단에서의 출현빈도와 부적합문헌 집단에서의 출현빈도를 비교하여 낮은 쪽을 제거하는 '저빈도제거법'을 사용하였다.

중복제거법을 이용한 수정질의의 검색결과를 초기질의의 검색결과와 비교한 결과 다음과 같은 특징이 드러났다.

(1) 17개의 수정질의는 재현율 80% 이상인 집단, 47%~51%인 집단, 10% 미만인 집단의 세 수정질의 집단으로 구분되었다.
(2) 배제율을 기준으로 90%가 넘는 수정질의 집단은 재현율이 51%미만의 수정질의 집단과 일치하며, 배제율이 90% 미만인 수정질의 집단은 재현율이 80% 이상인 집단과 일치하였다.
(3) 재현율이 80% 이상인 수정질의 집단(배제율이 51% 미만인 수정질의 집단)은 요소 B를 AND로 결합한 경우, 요소 A를 NOT으로 결합한 경우, 요소 E를 NOT으로 결합한 경우, 요소 F를 NOT으로 결합한 경우이다. 이 수정질의 집단은 OQ에 비하여 정확률이 모두 높게 나타났다.
(4) 재현율이 47%~51%인 수정질의 집단(배제율이 90% 이상인 수정질의 집단중 일부)은 요소 C를 AND로 결합한 경우이다. 이 집단은 집단은 OQ에 비하여 정확률이 모두 2배 이상 향상된 결과를 보인다.
(5) 재현율이 10%미만인 수정질의 집단(배제율이 90% 이상인 수정질의 집단중 일부)은 요소 D를 NOT으로 결합한 경우이다.
(6) 재현율 100%인 경우는 요소 A를 NOT으로 결합한 수정질의 한 가지이다.

저빈도제거법을 이용한 수정질의의 검색결과와 초기 검색결과의 비교 결과 다음과 같은 특징이 드러났다.

(1) 17개의 수정질의는 재현율 80% 이상인 집단, 25%~35%인 집단, 10% 미만인 집단의 세 수정질의 집단으로 구분되었다.
(2) 배제율을 기준으로 80%가 넘는 수정질의 집단은 재현율이 35%미만의 수정질의 집단과 일치하며, 배제율이 39% 미만인 수정질의 집단은 재현율이 80% 이상인 집단과 일치하였다.
(3) 재현율이 80% 이상인 수정질의 집단(배제율이 39% 미만인 수정질의 집단)은 요소 C′을 AND로 결합한 경우이다. 이 수정질의 집단은 OQ에 비하여 정확률이 모두 높게 나타났다.

(4) 재현율이 35% 미만인 수정질의 집단은 요소 E′을 NOT으로 결합한 수정 질의들이며, 재현율이 80% 이상인 수정질의에 비하여 배제율이 2.6배 이상으로 현저하게 높았다.

(5) 재현율이 10%미만인 수정질의 집단은 요소 D와 E′을 NOT으로 결합한 경우로서, 배제율이 93.3% 이상으로 가장 높게 나타났다.

재현율, 정확률, 배제율을 각각 가장 중요한 검색성능 척도로 삼아실험 결과를 분석한 결과 분류기호 요소 A를 NOT으로 결합하여 사용한 수정질의들 중에서는 재현율 중심의 경우 OQ NOT E NOT A가, 정확률 중심의 경우 OQ AND C NOT E NOT A가 최적의 수정질의식으로, 배제율 중심의 경우 OQ AND C NOT E NOT A가 최적의 수정질의식인 것으로 나타났다.

분류기호 요소 A를 사용하지 않은 수정질의들 중에서는 재현율 중심의 경우 OQ NOT E가, 정확률 중심의 경우 OQ AND C NOT E가, 배제율 중심의 경우도 OQ AND C NOT E가 최적의 수정질의식으로 밝혀졌다.

이 연구에서는 정보필터링 시스템을 통하여 제공된 초기 검색결과에 대한 이용자의 적합성 피드백을 바탕으로 새로운 수정질의를 만들고, 이 수정질의를 통하여 검색효율을 높이려는 목적으로 불논리에 기반한 최적의 수정질의식을 도출하였다. 이 연구에서 도출된 수정질의식들은 실제 시스템을 통하여 이용자에게 시스템이 수정한 새로운 검색식으로 제시될 수도 있고, 시스템 내부에서 자동으로 수정질의로 저장하여 후속 검색에서부터 적용할 수도 있다.

향후 이 연구를 통하여 도출된 새로운 수정질의식을 이용한 후속검색과 그 결과에 대한 이용자의 재피드백 과정을 통한 수정질의식의 검증 및 재수정 검색에 대한 연구를 통하여 수정질의에 대한 검증연구가 있어야 할 것이다. 또한 불논리 검색의 한계점인 출현하지 않은 용어 또는 색인되지 않은 용어로 검색이 불가능한 문제점과 질문-문헌간 유사도에 근거한 검색의 문제점을 극복할 수 있도록 질의확장 검색과 연계된 연구도 수행되어야 할 것이다.

참 고 문 헌

박지연. 2001. 『질의확장에 의한 단락 검색의 성능 향상에 관한 연구』. 석사학위논문. 연세대학교 대학원, 문헌정보학과.

Callan, J. 1996. "Document filtering with inference networks." *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 262-

269.

Croft, W. B., Harper, D. J. 1979. "Using probabilistic models of document retrieval without relevance information." *Journal of Documentation*, 35(4) : 285-295.

Denning, P. 1982. "ACM president's letter on electronic junk." *Communications of the ACM*, 25(3) : 163-165.

Dillon, M., Desper, J. 1980. "The use of automatic relevance feedback in boolean retrieval systems." Journal of Documentation, 36(3) : 197-208.

Fidel, R., Crandall, M. 1997. "Thr role of subject access in information filtering." *Proceedings of the 1997 Clinic on Library Applications of Data Processing*. 16-27.

Fisher, G., Stevens, C. 1991. "Information access in complex, poorly structured information spaces." *Chi'91 Conference Proceedings of ACM*, New York. 63-70.

Greenberg, J. 2001. "Optimal query (QE) processing methods with semantically encoded structured thesauri terminology." *Journal of the American Society for Information Science and Technology*, 52(6) : 487-498.

Ide, E. 1971. "New experiments in relevance feedback." In *The Smart System-Experiments in Automatic Document Processing*, NJ : Prentice Hall.

Kay, J., Kummerfeld, R. J. 1996. "User model based filtering and customisation of web pages." *UM'96 Workshops*. [online]. [cited 2002.10.03] <http://www.cs.su.oz.au/~bob/um96-paper.html>

Luhn, H. P. 1958. A business intelligent machine." *IBM Journal of R&D*, 2 : 314-319.

Malone, T. W., Grant, K. R., Turbak, F. A. 1986, "The Information Lens : an intelligent system for information sharing in organizations." *CHI'86 Proceedings of ACM*. 1-8

Mandala, R., Takenobu, T. and Tanaka, H. 2000. "Query expansion using hetrogeneous thesauri." *Inoformation Prosessing & Management*, 36(3)

: 361-378.

Qui, Y., Frei, H. 1993. "Concept based query expansion." *ACM SIGIR '93.* 160-169.

Ram, A. 1992. "Natural language understanding for information filtering system." *Communications of the ACM,* 35(12) : 80-81.

Rocchio, J. J. 1971. "Relevance feedback in information retrieval." In G. Salton (ed). *Smart Retrieval System.* NJ : Prentice Hall.

Rodriguez-Mula, H. G., Garcia-Monila and A. Paepcke. 1998. "Collaborative value filtering on the web." *Computer Networks and ISDN Systems 30.* 736-738.

Salton, G. 1981. "The estimation of term relevance weights using relevance feedback." *Journal of Documentation,* 37(4) : 194-214.

Stadnyk I. and Kass, R. 1992. "Modeling users' interests in information filters." *Communications of the ACM,* 35(12) : 49-50.

Stevens, C. 1992. "Automating the creation of information filters." *Communications of the ACM,* 35(12) : 48.

Xu, J., Croft, W. B. 1996. "Query expansion using local and global document analysis." *Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval.* 4-11.

Yang, Y. 1999. "An evaluation of statistical approaches to text categorization." *Information Retrieval,* 1 : 69-90.