자동색인기 성능시험을 위한 Test Set 개발

A Development of the Test Set for Estimating the Retrieval Performance of an Automatic Indexer

김성혁(Sung-Hyuk Kim)*, 서은경(Eun-Gyoung Seo)**, 이원규(Won-Gyu Lee)***,
김명철(Myoung-Cheol Kim)***, 김영환(Young-Whan Kim)****, 김재군(Jae-Kun Kim)****

□ 목 차 □

초 록

다양한 정보를 신속, 정확하게 제공할 수 있는 정보검색시스템은 선진국에서 일찍이 개발되어 현재 우리나라에서도 한국어 데이타베이스를 검색할 수 있는 정보검색시스템이 실험적으로 또는 상업적으로 개발되고 있다. 이에 따라 개발된 시스템의 실행 가능성 테스트(feasibility test)가 계속 부수적으로 수행되어 왔으나 평가 테스트들의 객관성 부족으로 인하여 개발된 정보검색시스템의 성능 또한 논쟁이 되어왔다.

이에 본 연구는 한국어 정보검색시스템과 자동색인기의 객관적인 성능 평가를 위하여 실험데이타 컬렉션을 개발하였다. 실험데이타 컬렉션은 정보과학회논문지, 한국정보과학회 1993 Proceedings, 정보관리학회지에 수록된 1,053개의 논문으로 구성되었다. 입력된 모든 데이타는 국문 및 영문 저자, 서명, 서지사항, 초록, 분류번호, 색인어 등 18개의 access point를 지니며, 한국어 문헌 Set 구축이외에 Test Set과 관련된 질의문을 작성하였고 질의문에 해당하는 적합문헌을 제시해 주었다

ABSTRACT

According to the development of various information retrieval system suitable for Korean database, many researchers have realized the need of IR Test Collection which can be readily used for evaluating a retrieval system. Therefore, This study developed the TEST SET which helps objectively evaluating the retrieval performance of an Hangul Automatic Indexer or Korean Information Retrieval System. The developed Test Set has four files such as: 1) Korean Document Set(*. all); 2) Natural Language Query Set(KTset.nql); 3) Boolean Query Set(Ktset.bql); 4) Query-Relevance Judgment Set(KTset.rel).

* 숙명여자대학교 문헌정보학과 ** 한성대학교 문헌정보학과 *** 한국문예진흥원 전산개발부
**** 한국통신 인공지능연구실

1. 서 론

인간의 생활은 의사결정의 연속이라 할 수 있다. 이러한 의사결정에 필요한 정보는 인간의 뇌속에 들어있을 수도 있고, 외부로부터 유입될 수도 있다. 정보의 유형과 양이 제한적이었던 시대에는 의사결정에 필요한 정보를 찾는데 별 어려움이 없었다. 그러나 정보화 사회가 성숙되어 가면서 인간의 정보요구가 다양화, 전문화, 개성화되어가고 있을 뿐만 아니라 정보의 양은 기하급수적으로 증가하고 이런 정보를 수록한 매체가 다양해짐에 따라, 이제는 의사결정에 필요한 정보를 찾는다는 것은 백사장에서 바늘을 찾는 것과 같을 정도로 어려운 상태이다. 따라서 적합한 정보를 신속, 편리하게 검색을 해줄 수 있는 시스템 또는 매캐니즘에 대한 이용자의 요구는 점점 증가하였다.

일찍부터 선진국에서는 이러한 요구에 부응하여 의사결정에 필요한 정보를 신속, 정확하게 제공하는 정보검색시스템을 개발하였고, 현재 모든 도서관 및 정보센타에서 이용하고 있다. 우리나라는 1980년대에 들어와서 몇몇 정보검색시스템을 개발하였지만 여러 측면에서 초보적인 단계의 시스템이라고 할 수 있었다. 무엇보다도 우리글로 된 대규모의 데이타베이스가 형성되지 못한 관계로 개발된 정보검색시스템은 주로 해외 데이타베이스를 탐색하는 시스템이어서 많은 제한적 요소를 가지고 있었다. 그러나 1980년대 중반부터 우리글로 된 데이타베이스의 구축이 본격화 되면서 이를 검색할 수 있는 정보검색시스템의 개발이 활발히 진행되고 있다.

일반적으로 정보검색시스템은 '색인'과 '검색'이라는 두가지 측면을 가지고 있기 때문에 이 두가지 면에서 개발되어 왔다. 오늘날 정보검색시스템에서 수행되고 있는 대부분의 정보검색 방법은 개념검색이 아니라 코드검색으로 입력된 문헌을 표현한 색인어와 질의어를 표현된 색인어가 정확하게 일치할 때 해당 문헌을 검색하는 방법을 말한다. 이러한 정보검색시스템에서의 색인어 역할은 정보원과 정보 요구자 사이에 위치하여 특정한 주제의 문헌들을 선별해 주고 선별된 자료의 소재를 지시하여 주는 것으로 색인의 성능이 직접적으로 정보시스템의 성능에 영향을 미친다고 볼 수 있다. 따라서 많은 정보학자들은 검색의 성능을 높이기 위하여 색인어 또는 색인 기법에 대한 연구를 계속해서 수행하고 있다. 특히, 최근에는 데이타베이스에 입력되는 문헌의 제목, 초록, 본문 등을 컴퓨터가 분석하여 논문의 주제를 나타내는 색인어를 추출하는 자동색인 기법(예: 서은경, 1993; 최기선, 1991)이나, 문헌의 내용과 정보요구를 나타내는 색인언어를 우리가 일상 사용하는 자연언어로 표현하여 검색하는 방법에 대한 연구가 계속 진행되고 있다.

또한 '색인' 측면이외에 정보검색 기법, 모델링, 시스템 개발 등 '검색' 측면을 중심으로 한 연구들도 많이 수행되고 있다. 특히, 퍼지집합 검색, 가중치에 의한 검색, 매칭함수에 의한 검색, 확률검색 등 여러 가지 검색기법을 비교 평가하는 연구(예: 김현희 & 배금표, 1993; 이준호, 1993)와 용어절단 탐색, 제한 탐색, 비교탐색, 본문 탐색, 관련용어 탐색 등과 같은 탐색기법을 연구가 활발히 진행되었으며, 보다 최근에는 인공지능의 응용에 관한 연구가 활발해짐에 따라 비록 실험적이지만 여러 종류의 지능형

정보검색시스템이 개발되고 있다(예: 김성혁, 1992; 정영미, 1991).

이와같이 새로운 기법을 이용한 정보검색시스템을 개발하고 평가하는 연구 또는 기존 기법들을 비교 평가하는 연구들이 계속 나오고 있으나 평가 테스트 또는 실행 가능성 테스트(feasibility test)들의 객관성 부족으로 인하여 개발된 정보검색시스템 성능 또는 비교연구의 적합성이 계속해서 논쟁 되어왔다. 따라서 정보검색시스템의 평가에 관한 문제는 이 분야의 연구자들에게 주요 관심사로 부각되었다. 정보검색시스템의 평가에 영향을 미치는 요인들은 다양하다. 컴퓨터의 성능, 검색소프트웨어의 기능, 색인의 수준, 질의어 처리 수준, 검색된 문헌의 적합성, 이용자의 수준 등이 일반적으로 이용하는 요인들이다. 이 요인들 중에서 항상 논쟁되온 이슈는 적합성 판정, 이용자 만족도 등과 같은 주관적으로 평가해야 하는 요인으로 현재까지도 많이 논의되어 왔으나(Froehlich, 1994; Hersh, 1994; Su, 1994; Saracevic, 1988 & 1975), 객관적 평가를 위해 실험데이타 컬렉션의 개발은 지금까지 우리나라에서는 간과되어 왔다. 즉 개발된 한국어 정보검색시스템의 성능 또는 한글을 대상으로 개발된 색인어 기법, 검색기법, 탐색기법 등을 객관적으로 평가할 수 있는 객관적인 한국어 실험데이타 컬렉션구축에 관한 연구는 없다는 것이다. 외국의 경우, 다양한 주제분야, 데이타의 양, 개발된 정보검색시스템의 성격 등에 적합한 실험데이타 컬렉션들이 1970년대 부터 개발되어 검색시스템의 평가에 많이 이용되고 있다. 국내의 경우, 정보검색시스템을 객관적으로 평가할 수 있는 실험데이타 컬렉션의 필요성은 인식하고 있지만 아직까지 개발이 되지 않은 실정이다. 한 주제에 대한 충분한 양의 문헌, 그 문헌을 주제전문가가 분석하여 만들어진 색인어, 탐색 평가를 위하여 구축된 질의문과 탐색문, 각 질의문에 대한 문헌의 적합성 판정 등이 담겨진 테스트 컬렉션은 정보검색시스템, 자동색인기, 다양한 검색 기법 등을 평가하는데 이제는 필수적이라 볼 수 있다.

이에 본 연구는 한국어 정보검색시스템 또는 한글 자동색인기의 객관적인 성능 평가는 물론 성능평가를 용이하게 하기 위하여 분류번호와 색인어가 포함된 한국어 문헌데이타와 문헌 데이타 set에 맞는 질의문과 적합 문헌을 제공하는 Test Set을 구축하였다.

실험데이타 컬렉션은 국문 및 영문 저자, 서명, 서지사항, 초록, 분류번호, 색인어 등 18개의 access point로 탐색할 수 있는 문헌 Set, 자연언어 질의문 set, 불리안 질의문 set, 질의문-적합문헌 set으로 구성되었다. 입력된 데이타는 정보과학회논문지, 한국정보과학회 1993 Proceedings, 정보관리학회지에 수록된 1,053 개의 논문이다.

2. 정보검색 테스트 컬렉션

정보검색시스템의 성능 평가를 위해서 미국을 중심으로 1970년대 초부터 테스트 컬렉션(Test Collection)을 준비하여 많은 연구자들의 실험에 도움을 주어왔다. 처음으로 만들어진 테스트 컬렉션은 CRANFIELD TEST COLLECTION으로 Aslib-Cranfield Project에 사용되었던 기체역학과 항공공학 분야의 1,398 개의 문헌을 수록한 컬렉션이다. 특히 각 문헌에 대하여 단일색인어,

복합어로 구성된 개념어, Engineer's Joint Council Thesaurus에서 추출한 통제어로 구성된 세 개의 색인어 그룹이 있으며, 탐색시 각 색인어는 가중치에 의하여 그 사용 범위가 좁혀지거나 또는 색인어의 어간 탐색이 가능하여 그 범위를 넓힐 수도 있도록 하였다. 그 이후 자연과학 문헌을 중심으로 여러 종류의 테스트 컬렉션이 만들어 졌다.

대표적인 테스트 컬렉션으로는 프로그램 디버깅으로 주로 사용되는 ADI, 컴퓨터 공학 분야의 CACM, 문헌정보학 분야의 CISI와 LISA, 우주항공 분야의 CRAN, 전자 전기 분야의 INSPEC, 의학 분야의 MED와 NLM, 물리학 분야의 NPL, Rutgers 대학의 문헌을 수록한 RIRD, Cornell 대학과 Massachusetts 대학의 논문을 수록한 TIME 등이 있다. 대부분의 컬렉션은 5 개의 화일로 구성되어 있다. 첫째번 화일은 서명, 저자, 초록, 색인 등 모든 데이타가 수록된 원래 문헌의 텍스트(*.all)로 구성되어 있으며, 두번째 화일은 자연어로 표현된 질의문 화일(*.qry)이고, 세번째 화일은 문헌 벡터(npl.dvr)와 질의문 벡터(npl.qvr)로 구성된 벡터 화일(*.npl)이다. 또한 네번째 화일은 불리안 질의문 화일(*.bln)로 보통 간단한 질의문 화일(*.bl1)과 확장 질의문 화일(*.bl2)로 구성되어 있으며 컬렉션을 구성하는 마지막 화일은 각 질의문에 대하여 적합 문헌과 그 적합 정도를 알려주는 적합성 평가 정보 화일(*.rel)이다. 예를 들어 CACM 문서 컬렉션의 화일의 구성은 다음과 같다: 1) cacm.all (문서 원본); 2) cacm.qry (자연어로 기술된 질의문); 3) cacm.bln (불리안 질의문: cacm.bl1, cacm.bl2); 4) cacm.rel (적합성 평가 정보); 5) cacm.db0 (cacm의 확장 버전).

다음 〈표 1〉은 정보검색 성능 평가를 하는데 가장 많이 사용되는 테스트 컬렉션인 CACM, CISI, CRAN, MED, INSPEC의 문헌수, 질의문의 수 등을 자세히 살펴보았다. Cornell대학에서 처음 작성하고 Rebert Korfhage가 확장을 한 CACM 테스트 컬렉션은 1958년에서 1979년 사이에 Communication of the Association for Computing Machinary에 출판된 3,204 개의 문헌을 수록한 것으로 질의문는 52 개 정도 있다. CISI 테스트 컬렉션은 1969년에서 1977년 사이에 출판된 정보학과 도서관학 분야를 다룬 1,460 개의 문헌과 76 개의 질의문을 수록한 것이며, CRAN은 Aslib-Cranfield Project에 이용된 1,398 개의 문헌과 225 개의 질의문을 수록한 테스트 컬렉션이다. INSPEC은 가장 큰 장서규모를 가지는 Test Set으로 전자학, 전자공학, 컴퓨터 공학분야의 문헌 12,684개를 수록한 것으로 질의문의 수는 77 개이다. 마지막으로, MED 테스트 컬렉션은 국립 의학 도서관에서 받은 의학자료 중 1,033개의 문헌을 선택하여 구축된 set이다.

이외에 물리학분야의 11,492개의 문헌과 100 개의 질의문으로 구성된 NPL(National Physical Laboratory)도 잘 이용되는 테스트 컬렉션으로 다른 컬렉션과 다소 틀리다. 즉, 입력된 문헌는 원래의 자연언어 형태의 원본문이 아니라 색인이라 할 수 있듯이 키워드의 집합으로 구성되었고 질의문은 한 개의 탐색어로 구성되어 있는 특징이 있다. 따라서 NPL의 이용은 특수한 경우에 이루워지고 있다.

또한 정보검색 분야의 논문이 어느 정도 테스트 컬렉션을 이용했는 지를 알아보기 위하여

〈표 1〉 CACM, CISI, CRAN, MED, INSPEC의 데이타의 통계적 특성

문 헌 컬 렉 션		CACM	CISI	CRAN	MED	INSPEC
문헌의 총수		3,204	1,460	1,398	1,033	12,684
어간형태의 용어의 수		4,522	5,019	3,763	6,927	14,255
문헌당 평균 어간의 수		20.22	45.20	53.13	51.60	30.01
문헌 길이의 표준편차		21.21	19.38	22.53	22.78	14.27
색인어의 수	최 대	2.94	5.29	5.81	5.88	3.86
	최 소	1.00	1.00	1.00	1.00	1.00
	평 균	1.23	1.39	1.54	1.51	1.36
색인어의 문헌 빈도수	최 대	904.7	573.0	775.7	310.7	3724.5
	최 소	14.0	2.9	3.3	1.3	14.9
	평 균	236.5	123.2	173.0	59.9	722.1
질의문 컬렉션		CACM	CISI	CRAN	MED	INSPEC
질의문의 수		52	76	225	30	77
어간형태의 용어의 수		324	657	585	241	576
질의문당 평균 어간의 수		10.67	22.59	9.17	10.10	15.81
질의문 길이의 표준 편차		6.43	19.49	3.19	6.03	8.66
탐색어의 수	최 대	1.98	3.38	1.28	1.53	2.64
	최 소	1.00	1.00	1.00	1.00	1.00
	평 균	1.14	1.25	1.03	1.08	1.21
탐색어의 문헌빈도수	최 대	754.1	581.2	532.5	190.1	3371.5
	최 소	17.6	21.9	31.3	8.5	45.0
	평 균	205.6	186.2	1 7.2	59.1	752.3
질의문당 평균 관련 문헌		15.8	33.0	8.2	23.2	49.8

정보학 분야의 대표적인 학술잡지인 *Journal of the American Society for Information Science(JASIS), Information Processing and Management(IPPI), Journal of Documentation (JD)*에서 1970년 부터 현재까지 테스트 컬렉션을 이용하여 실험을 한 논문을 찾아보았다. 그 결과 총 42 개의 논문이 있었고 각 논문은 최소 1 개에서 최대 6 개까지의 테스트 컬렉션을 사용하였다. 대다수의 논문이 "색인 기법," "검색 기법," "적합성 판정 방법"에 관련된 것으로 보다 상세한 주제는 〈표 2〉에 나타나 있다.

다른 한편으로 위의 42 개 논문이 사용한 테

스트 컬렉션의 종류와 어느 정도 사용되었는 지를 알아보았다. 가장 먼저 구축된 CRANFIELD (CRAN) Test Collection이 가장 많이 이용되었고 그 뒤를 MED, CISI, CACM, INSPEC 등이 자주 이용된 사실이 나타났다(참조 〈표 3〉).

3. KT Test Set 개발

3.1 한국어 문헌 Set (KTset.all) 구성

컴퓨터공학과 정보학 분야의 테스트 컬렉션을 구축하기 위하여 먼저 한글, 영문 초록이 다

〈표 2〉 테스트 컬렉션을 사용한 학술 논문

주제분야	수록된 잡지와 논문 수	총 논문 수
검색기법	JASIS (4) IPM (1) JD (2)	7
검색시스템평가	JD (1)	1
색인 기법	JASIS (8) IPM (4) JD (4)	16
색 인 어	JASIS (1) IPM (3)	4
적합성 가중치	JASIS (1) IPM (2)	3
적합성 평가	IPM (1) JD (2)	3
탐색 기법	IPM (1)	1
탐 색 어	JASIS (2) IPM (1) JD (1)	4
피드백 기법	JASIS (2) JD (1)	3

〈표 3〉 학술 잡지에 나타난 테스트 컬렉션의 종류와 이용된 수

컬렉션	수	컬렉션	수	컬렉션	수
AIP SPIN	1	EVANS	2	MED	11
AIR	2	HARDING	2	NPL	3
CACM	8	INSPEC	7	SMART	1
CF	4	ISILT	2	TIME	5
CISI	8	KEEN	3	UKCIS	4
CRAN	23	LISA	4		

담긴 학술잡지 세 개를 선정하였다. 선정된 학술잡지는 정보과학회논문지, 한국정보과학회 Proceedings, 정보관리학회지이며 수록된 총 논문은 1,053 개이다. 서지사항, 분류번호, 색인어 등이 태그(tag)를 이용하여 구분되었고 각 논문에 나타난 수식이나 특수문자 등은 TeX형식으로 입력하여 이기종간에 불편없이 호환될 수 있도록 하였다. 가능한한 원자료에 충실하였으며, 원자료의 오류가 분명하다고 인정되는 것에 대해서는 수정을 가했다.

문헌분류는 CRCS를 기준으로 복수 분류를 허용하였다. 즉, 한 논문에 2 ~3 개의 분류번호가 주어졌다. CRCS(Computing Reivew Classification Structure)는 *Computing Reviews*라는 잡지에 사용된 분류표로 *Computing Reviews*의 편집위원에 의하여 구성된 것이다. 컴퓨터 분야의 연구 동향을 구분하기 위해 구성된 CRCS는 보는 관점에 따라 실세계의 주제 체제를 충분히 반영하고 있지 못하다는 단점을 안고 있으나 다른 어떤 분류표를 이용할 때보다 컴퓨터공학 분야를 상세히 분류할 수 있다는 장점을 가지고 있다. 그러나 정보관리학회지의 내용을 분류하기에는 그 한계성이 있어 정보학 분야를 분류할 수 있는 X 필드를 별도로 첨가하였다. X 필드의 분류체계는 ISA(Information Science Abstract)에서 사용하고 있는 분류체계를 이용하였다.

색인어의 선정은 수작업으로 이루워졌다. 일반적으로 수작업 색인은 색인자의 주제 지식을 이용하여 색인자가 직접 각 문헌에 해당하는 색인어를 할당/책정하는 작업을 뜻한다(Kemp, 1988). 따라서 색인작업은 참여 연구원 4명이 이 분야의 주제전문가라는 전제 아래 분담하여 이루워졌다. 색인어의 불일치성과 색인자의 편견, 실수를 가급적 최소화시키기 위하여 색인어를 각 논문의 초록과 서명에서 자연어 형태로 추출한 다음, 토론과 협의를 걸쳐 가장 적합한 색인어를 선택하였다. 또한 같은 내용의 영어 어휘 또는 영어 약자로도 자주 쓰이는 용어는 같이 선택하였다.

각 문헌은 고유번호, 국문/영문 저자, 국문/영문 서명, 서지사항, 국문/영문 초록, 분류번호, 색인어 등 18 개의 access point를 가지고 있으며, access point를 이용하여 모든 필드를 검색할 수 있도록 구성되었다. 한국어 문헌 Set는 네 개의 화일로 구성되어 있는데 그 내용은 다음과 같다.

- kiss.all: 403건 (KISS: Korean Information Science Society)
  1985년에서 1993년까지 정보과학회논문지에 수록된 논문들을 입력한 화일.
- kissps.all: 212건(KISSPS: Korean Information Science Society Proceedings Spring)
  한국정보과학회 Proceedings 1993년 춘계편에 수록된 논문들을 입력한 화일.
- kisspf.all: 322건 (KISSPF: Korean Information Science Society Proceedings Fall)
  한국정보과학회 Proceedings 1993년 추계편에 수록된 논문들을 입력한 화일.
- ksim.all: 116건 (KSIM: Korean Society of Information Management)
  창간호(1984)에서 부터 1993년까지 정보관리학회지에 수록된 논문들을 입력한 화일.

다음은 한국어 문헌 Set에 수록된 문헌의 예와 입력할 때 사용된 태그를 제시한 것이다. 또

한 본 연구에서 구축된 문헌 Set과 CACM에서 구축된 문헌 Set를 비교하기 위하여 CACM의 문헌 Set의 예도 제시하였다. 두 개의 예에서 볼 수 있듯이 CACM은 훨씬 적은 수의 access point를 가지며 색인어의 수도 훨씬 작은 것으로 보인다.

〈예 1: 한국어 문헌 Set〉

〈id〉0985
〈title〉문헌정보학 영역의 지능형 정보시스템에 관한 고찰
〈author〉김성혁
〈affiliation〉숙명여자대학교 문헌정보학과
〈language〉한국어
〈journal〉정보관리학회지
〈issn〉1013-0799
〈year〉1992
〈volume〉9
〈number〉1
〈pages〉165-181
〈abstract〉본 연구는 차세대 정보시스템으로 정착되어가는 지능형 정보시스템의 등장배경과 개념, 그리고 지능형 정보시스템 구축에 필요한 기술인 객체지향시스템, 전문가시스템, 하이퍼미디아 등을 고찰하였다.
나아가 이들 기술이 문헌정보학의 영역에 적용되어 개발된 시스템을 중심으로 지능형 정보시스템을 소개하였고 앞으로의 전망에 대해 기술하였다.
〈etitle〉A Study on Intelligent Information System in the Field of Library and Information Science
〈eauthor〉
〈eabstract〉The purpose of this study is to review the concept and background of intelligent information system which will be fixed on he next generation's information system and to describe the object orientation system, expert system, and hypermedia which are the core technologies in the design of intelligent information system. Furthermore, intelligent information systems which are developed using these technologies in the field of library and information science, and the future prospects on the intelligent information system are described in detail.
〈classification〉
H.3.4.3
X.5.2
X.5.11
〈keywords〉
문헌정보학
지능형 정보시스템
정보시스템
객체지향시스템
전문가시스템
하이퍼미디아
〈notes〉

화일에 사용된 Tag의 의미

Tag	Tag 내용
〈id〉	고유번호
〈title〉	제목
〈author〉	저자, 공저자
〈affiliation〉	소속기관(저자, 공저자)
〈language〉	수록언어
〈journal〉	수록지
〈issn〉	국제표준 연속간행물 번호
〈year〉	수록년도
〈volume〉	수록권
〈number〉	수록호

〈pages〉	페이지 (시작-끝)
〈abstract〉	한글초록
〈etitle〉	영문제목
〈eauthor〉	영문저자, 공저자
〈eabstract〉	영문초록
〈classification〉	분류 (CRCS에 근거 1개 이상 가능)
〈keywords〉	색인어
〈notes〉	기타정보

〈예 2: CACM의 문헌 Set〉

title--PROLOG in 10 Figures
abstract--In the fall of 1981, a Japanese report officially initiated the quest for fifth-generation computers that would encompass the functions of knowledge processing and artificial intelligence. The conceptual understanding behind Prolog - Japan's language of choice for these activities - are presented here in a way that suggests why Prolog of a similar language might be considered a model for designing the computers of the future.
journal--CACM December, 1985
author--Alain Colmerauer
keys--languages
categories--D.3.1, D.3.2, I.2.5
end--CA851201 SM Jan 24, 1986 8:39PM

3.2 자연어 질의문 Set(KTset.nql)과 불리안 질의문 Set(KTset.bql) 구성

구축된 문헌을 검색하기 위하여 본 연구는 자연어 질의문 Set을 작성하였다. 30 개의 질의문으로 구성되어 있는 자연어 질의문 Set는 탐색자의 실제 정보요구를 선정하여 이루워진 것이 아니라 참여 연구원이 본 연구에서 구축한 문헌 Set에 적합한 질의문을 인위적으로 작성하여 구성되었다. 또한 한 주제에 질의문이 몰리는 것을 배제시키기 위하여 입력된 모든 논문을 주제로 분류한 다음, 그 주제에 해당되는 논문의 분량에 비례하여 질의문을 분포하였다. 보통 한 질의문에서 주 탐색어가 2 개에서 4 개 정도 추출될 수 있도록 하여 질의문이 너무 일반적이거나 특정적인 특징을 갖지 못하게 하였다. 다음의 〈예 3〉은 본 연구에서 작성한 자연어 질의문 Set 중 10 개를 보여주고 있다.

〈예 3: 자연어 질의문 Set〉

〈nlq〉 1 실시간 데이타베이스에 관한 연구
〈nlq〉 2 멀티미디어 데이타베이스에 관한 연구는?
〈nlq〉 3 객체지향 데이타베이스를 다룬 논문은?
〈nlq〉 4 분산 환경의 데이타베이스에 대한 연구 결과에는 어떤 것이 있나요?
〈nlq〉 5 인공지능 분야중 퍼지이론을 응용한 연구
〈nlq〉 6 인공지능중에서 필기체 문자인식을 다룬 것
〈nlq〉 7 인공지능을 갖고 있는 로보트
〈nlq〉 8 중형 컴퓨터의 구조에 대한 설계 및 구현에 대한 연구
〈nlq〉 9 프로그래밍 언어에 대한 컴파일러의 작성에 대한 연구
〈nlq〉10 분산 운영체제

불리안 질의문란 각 자연어 질의문에 나타난 키워드(탐색어)를 불리안 논리연산을 이용하여 그 논리적 관계를 표현한 탐색문을 말한다. 본

연구에서는 두 종류의 불리안 질의문 Set를 작성하였다. 먼저 KTset.bq1은 자연어 질의문에 나타난 주 탐색어만을 이용하여 논리연산으로 변환시킨 것이고, KTset.bq2는 KTset.bq1에 나타난 탐색어의 동의어 및 관련어 또는 영어 표기 등을 이용하여 작성된 확장 불리안 질의문 Set를 말한다. 다음은 KT와 CACM에서 작성된 불리안 질의문 Set이 예를 보여주고 있다.

〈예 4: 불리안 질의문 Set1(KTset.bq1)〉

〈bq1〉 = '데이타베이스' and '실시간'
〈bq2〉 = '데이타베이스' and '멀티미디어'
〈bq3〉 = '데이타베이스' and '객체지향'
〈bq4〉 = '데이타베이스' and '분산'
〈bq5〉 = '인공지능' and '퍼지'
〈bq6〉 = '인공지능' and '필기체' and '문자인식'
〈bq7〉 = '인공지능' and '로보트'
〈bq8〉 = '컴퓨터 구조' and '중형'
〈bq9〉 = '프로그래밍 언어' and '컴파일러'
〈bq10〉 = '운영체제' and '분산'

〈예 5:불리안 질의문 Set2(KTset. bq2)〉

〈bq1〉 = ('데이타베이스' or 'Database' or 'DB') and ('실시간' or 'real time' or 'real-time')
〈bq2〉 = ('데이타베이스' or 'Database' or 'DB') and ('멀티미디어' or 'multimedia')
〈bq3〉 = (('데이타베이스' or 'Database' or 'DB') and('객체지향' or 'object oriented')) or 'OODB' or 'Object Oriented Database'
〈bq4〉 = (('데이타베이스' or 'Database' or 'DB') and('분산' or 'Distributed') or 'DDB' or 'Distributed Database'
〈bq5〉 = ('인공지능' or 'AI' or 'Artificial Intelligence') and ('퍼지' or 'Fuzzy')
〈bq6〉 = ('인공지능' or 'AI' or 'Artificial Intelligence') and ('필기체' or '필기' or 'Hand Written') and ('문자인식' or 'Character Recognition')
〈bq7〉 = ('인공지능' or 'AI' or 'Artificial Intelligence') and ('로보트' or 'Robot')
〈bq8〉 = ('컴퓨터 구조' or 'Computer Architecture') and ('중형' or 'Mid-range')
〈bq9〉 = ('프로그래밍 언어' or 'Programming Language') and ('컴파일러' or 'Compiler')
〈bq10〉 = ('운영체제' or 'OS' or 'Operating System') and ('분산' or 'Distributed')

〈예 6: CACM의 자연어 및 불리안 질의문 Set〉

〈nlq1〉 What articles exist which deal with TSS (Time Sharing System), an operating system for IBM computers?
〈#bq1.1〉= #or ('tss', #and ('ibm', #and ('time', 'sharing')));
〈#bq1.2〉= #or (#and('ibm', 'tss'), #and ('ibm','time','sharing','system'));

3.3 적합문헌 Set(KTset.rel) 구성

각각의 질의문에 대하여 적합한 문헌을 분야별 전문가(참여 연구원)들에 의하여 선정되었다. 적합성이란 질문과 정보자료간의 내용상의 일치 정도나 이용자에게 그 자료가 얼마나 유요한 것인가를 나타내는 용어로 정의내릴 수 있다. 따라서 적합성은 이용자 측면을 강조한 주관적 개념과 이용자 지식 상태는 고려하지 않고 주제 전문가가 판정하는 객관적 개념으로 설명할 수 있는데 실제 정보요구를 다루지 않은 본 연구에서는 객관적 개념으로서의 적합성을 근거로 하여 적합문헌을 선정하였다. 즉 살톤과 맥길(Salton and McGill, 1983)이 말한 적합성의 정의 ("정보자료가 질문에 적합한 내용을 다루고 있는 정도")에 따라 네 명의 주제전문가가 각 질의문에 대한 문헌의 적합성을 판정하였다. 주제전문가는 모든 문헌을 분석하여 각 질의문에 대한 적합 문헌을 선정하였고 네 명이 모두 일치하였을 때는 1 값을, 세 명이 일치하였을 때는 0.75 값을, 두 명이 일치하였을 때는 0.50 값을, 그리고 한 명만 적합하다고 보았을 때는 0.25 값을 주었다.

적합문헌 Set(KTset.rel)은 질의문 번호, 적합문헌번호, 적합 정도를 나타내는 값으로 구성되었고 다음은 KT의 적합문헌 Set과 CACM의 적합문헌 Set의 예를 제시하고 있다. CACM은 적합 정도를 나타내주는 필드는 있으나 실제, 값은 주지않고 있다.

〈예 7: KT와 CACM의 적합문헌 Set〉

〈KTset.rel〉			〈cacm.rel〉		
질의문	적합문헌 번 호	적 합 정 도	질의문	적합문헌 번 호	적 합 정 도
1	0174	0.75	1	1410	0.0
1	0351	0.75	1	1572	0.0
1	0352	1.00	1	1605	0.0
1	0353	1.00	1	2020	0.0
1	0572	1.00	1	2358	0.0
1	0574	1.00	2	2434	0.0
1	0575	1.00	2	2863	0.0
1	0617	1.00	2	3078	0.0
1	0704	0.25	2	0588	0.0
1	0705	0.50	2	0615	0.0
1	0706	0.50	2	0618	0.0
1	0707	0.50	2	0619	0.0
1	0713	0.50			
1	0731	1.00			

3.4 CRCS 번역

CRCS는 1982년 컴퓨터 분야의 연구 동향을 쉽게 파악하기 위하여 *Computing Reviews*에서 작성한 분류표이다. CRCS는 컴퓨터 분야를 11개의 큰 주제로 나눈 뒤, 세 개의 계층의 트리 구조를 가지고 있어 보다 주제를 상세히 분류할 수 있는 장점이 있다. 본 연구는 테스트 컬렉션을 구축하는 것 이외에 CRCS 분류표의 보급과 그 이용을 용이하게 하기 위하여 1993년에 개정한 CRCS를 번역하였다. 번역 또한 컴퓨터 분야의 주제전문가 두명이 일차적으로 번역한

다음, 다시 네명의 협의를 걸쳐 이루워졌다. 일반화된 한글 표기를 찾기 위하여 전기통신용어 사전, 영한/일한 컴퓨터 용어 큰사전 등 컴퓨터 분야의 전문용어 사전을 이용하였고 용어의 통일을 가져오도록 노력하였다. 다음은 CRCS의 두 계층을 번역한 예이다.

〈예 8: 한국어 CRCS 분류표 (CRCS.kor)〉

A. General literature
A.0 general
A.1 introductory and survey
A.2 reference
(dictionaries, encyclopedias, glossaries)
A.m miscellaneous

B. Hardware
B.0 general
B.1 control structures and microprogramming (D.3.2)
B.2 arithmetic and logic structures
B.3 memory structures
B.4 input/output and data communication
B.5 register-transfer-level implementation
B.6 logic design
B.7 integrated circuits
B.m miscellaneous

A. 일반문헌
A.0 일반
A.1 입문서 및 현황조사 (개관)
A.2 참고문헌
(예: 사전, 백과사전, 용어집)
A.m 기타

B. 하드웨어
B.0 일반
B.1 제어구조 및 마이크로 프로그래밍
B.2 연산 및 논리구조
B.3 메모리 구조 (기억장치)
B.4 입/출력 및 데이타 통신
B.5 레지스터-변환-수준의 구현
B.6 논리설계
B.7 집적회로
B.m 기타

4. Test Set의 분석

4.1 Test Set 데이타 및 색인 데이타의 분석

본 연구에서 구현한 문헌 Set의 데이타는 정보과학 및 문헌정보학 분야의 1,053 개의 학술잡지 논문에 대한 초록, 색인어, 분류 번호, 그 외 모든 서지적 사항으로 구성되었다. 따라서 Test Set은 1053 개의 레코드로 구성되었으며, 다음과 같은 18 개의 access point를 가지고 있다: id, title, author, affiliation, language, journal, issn, year, volume, number, page, abstract, etitle, eauthor, eabstract, classification, keywords, notes. 또한 학술잡지 논문은 1985년 이후에 출판된 403 개의 정보과학회논문지, 534 개의 1993년도 한국정보과학회 Proceedings, 창간호부터 1993년 현재까지 출판된 116 개의 정보관리학회지 논문으로 구성되었다.

Test Set 데이타를 분석한 결과, 학술잡지 논문의 초록에 평균 77 개의 어절이 있는 것으로 나타났으며, 표제는 평균 7개의 어절로 구성되어 있음을 알 수 있었다. 〈표 4〉에서 보면, 세 그룹의 문헌들의 표제 길이(표제를 구성하는 어절의 수)가 거의 비슷하나, 초록의 길이는 약간 차이가 있는 것으로 나타났다. 즉, 정보관리학회의 논문의 초록이 다른 학술잡지 논문보다 비교적 짧을 뿐 아니라, 가장 긴 초록 그리고 가장 짧은 초록이 모두 정보관리학회 논문임을 알 수 있었다. 예상외로, Proceedings 논문의 초록 길이가 학술지 논문의 초록 길이보다 긴 것으로 나타났다.

이와 같은 특성을 지닌 Test Set 데이타를 분석하여 각 문헌에 대한 수작업 색인어 리스트를 형성하였다. 〈표 5〉는 선정된 색인어의 통계적 특징을 보여주고 있다. 즉 1,053 개의 초록에서 색인자들이 전체 11,305 개의 색인어를 추출하여 한 초록당 평균 11 개의 색인어를 선택하였음을 알 수 있다. 정보과학회논문지에서 평균적으로 가장 많은 색인어(11 개)가 추출된 반면, 정보관리학회지에서 가장 작은 수의 색인어(9 개)가 추출되었다. 긴초록에서 보다 많은 색인어가 추출되는 일반적인 현상과는 다소 틀리게 나타났지만, 정보과학회논문지와 한국정보과학회 Proceedings의 초록 길이가 현저한

〈표 4〉 Test Set의 표제 및 초록에 관한 분석

	KISSPAP		KISS93		KSIM		TOTAL	
	표제	초록	표제	초록	표제	초록	표제	초록
문헌의 수	403	403	534	534	116	116	1053	1053
어절의 수	2,909	28,581	3,773	41,927	705	6,248	7,387	76,756
평균어절의 수	7.22	70.92	7.07	78.51	6.08	53.86	7.39	76.76
최소어절의 수	4	19	3	26	3	9	3	9
최대어절의 수	21	146	15	162	16	176	21	176

〈표 5〉 Test set의 색인어의 분석

	KISPAP	KISS93	KSIM	TOTAL
초록수	403	534	116	1,053
총색인어의 수	4,448	5,756	1,101	11,305
초록당 평균	11.03	10.78	9.49	10.74
최소 색인어 수	4	5	4	4
최대 색인어 수	29	23	10	29
범 위	25	13	14	25

〈표 6〉 단일 및 복합 색인어의 분석

	KISSPAP		KISS93		KSIM		TOTAL	
	총 수	평 균	총 수	평 균	총 수	평 균	총 수	평 균
색인어의 수	4,448	11.03	5,756	10.78	1,101	9.49	11,305	10.74
단일어 색인어 수	1,405	3.48	1,412	2.64	554	4.78	3,371	3.20
두어절로된 색인어 수	1,955	4.85	2,424	4.54	414	3.57	4,793	4.55
세어절로된 색인어 수	815	2.02	1,352	2.53	115	0.99	2,282	2.17
네어절로된 색인어 수	273	0.68	568	1.06	8	0.16	859	0.82

차이가 나타나지 않았으므로 이와 같은 현상이 일어날 수 있다고 볼 수 있다. 또한 한 초록당 최소 색인어수는 4로 나타났으며, 최대 색인어의 수는 29 개로 최대 범위가 25임을 알 수 있다.

다음은 선정된 색인어의 구문적 특성을 살펴보기 위하여 명사구로 된 색인어와 단일어 색인어의 통계를 구하여 보았다. 〈표 6〉에서 볼 수 있듯이 단일어 색인어의 수(30%) 보다는 두 어절로 된 색인어 수(42%)가 훨씬 많았으며 세 어절로 된 색인어도 20%나 차지하고 있다. 특히 정보과학회 논문지에서는 평균적으로 5 개의 두 어절로 된 색인어가 선정된 반면, 정보관리학회지에서는 단일어가 평균적으로 5 개씩 선정되었다. 또한 proceedings 논문에서는 평균적으로 한 문헌에서 한 개의 네 어절로 된 색인어를 찾아볼 수 있지만 정보관리학회지에서는 네 어절로 된 색인어가 거의 선정되지 않았음을 알 수 있다.

또한 선정된 색인어의 언어 구성을 살펴보았

다. 총 색인어의 71%가 한글로만 구성된 용어이며 영어로만 구성된 색인어가 22%정도로 나타났다. 특히 proceedings 논문에서는 한 문헌당 평균 세 개는 영어로 쓰여진 색인어인 반면, 정보관리학회지의 논문에서는 영어로 쓰여진 색인어가 거의 없는 것을 알 수 있다(참조 〈표 7〉).

4.2. Test Set의 속성 분석

한국어 문헌 Set의 속성을 알아보기 위하여, 이번에는 저자, 소속기관, 분류번호의 수와 문헌의 출판된 년도를 분석하였다. 평균적으로 한 문헌당 저자수는 3 명이며, 그 저자들이 소속한 기관의 수는 한 기관이며, 총 문헌에 대한 분류번호의 수는 1,905 개로서, 한 초록당 평균 2개의 분류번호를 가진 것으로 나타났다.

세그룹의 학술지에 대해 보다 상세히 분석해 보면 다음과 같다. 한국정보과학회 Proceedings 논문의 저자수는 평균적으로 세 명이며, 이들이 속한 기관의 수는 1.3으로 나타난 반면, 정보과학회논문지의 논문의 평균 저자수는 2 명이며

〈표 7〉 색인어의 언어 구성

	KISSPAP		KISS93		KSIM		TOTAL	
	총 수	평 균	총 수	평 균	총 수	평 균	총 수	평 균
색인어의 수	4,448	11.03	5,756	10.78	1,101	9.49	11,305	10.74
한글로만 구성된 색인어 수	3,270	8.11	3,694	6.92	1,015	8.75	7,979	7.58
영어로만 구성된 색인어 수	760	1.89	1,676	3.14	40	0.34	2,476	2.35
혼합되어 구성된 색인어 수	418	1.04	386	0.72	46	0.40	850	0.81

〈표 8〉 Test Set의 저자, 소속기과, 분류번호의 수 분석

	KISPAP	KISS93	KSIM	TOTAL
문헌의 수	403	534	116	1,053
총 저자의 수	957	1679	156	2,792
초록당 평균 저자의 수	2.37	3.14	1.34	2.65
총 기관의 수	692	696	145	1,533
초록당 평균 기관의 수	1.72	1.30	1.25	1.46
총 분류번호의 수	626	1036	243	1,905
초록당 평균분류번호의 수	1.55	1.94	2.09	1.81

소속기관 또한 2 개로 나타났다. 따라서 대다수의 정보과학회논문지의 저자는 다른 소속기관에 속한 연구자와 공저하는 것을 알 수 있었고, 정보관리학회지의 평균 저자의 수는 1.34로 대다수가 혼자 작업하여 발표하는 것으로 나타났다. 또한 정보관리학회지의 논문은 CRCS와 ISA 분류표를 이용하여 분류됨에 따라 각 문헌당 2 개이상의 분류번호를 가진 것으로 보인다 (참조 〈표 8〉).

다음 〈표 9〉는 본 연구에서 구축된 문헌 Set을 년도별로 분석한 것이다. 1993년에 출판된 proceedings 논문이 입력되었기 때문에 1993년의 논문이 가장 많이 수록되었으며, 데이타를 입력할 당시 정보관리학회지 2호판은 출판되지 않아 1993년 봄에 출판된 7 개의 논문만이 수록되었다.

4.3 Test Set의 분류 번호 분석

1,053개의 Test Set 문헌은 CRCS와 ISA Classification Scheme을 이용하여 분류 번호가 매겨졌다. 특히 ISA가 사용한 분류표는 정보관리학회지의 논문을 분류하기 위하여 사용된 것으로 다른 학술잡지 논문에 이 분류표를 적용시키지 않았다.

예상대로, A항 즉 General Literature(일반문헌)에 해당되는 논문은 하나도 없었지만, 그 밖의 주제에 골고루 분포되어 있는 것으로 나타났다. 특히, 정보과학회논문지의 많은 논문이 "Computing Methodologies"(계산방법론)에 관한 것이며, 한국정보과학회 Proceedings의 대다수 논문이 "Software"(소프트웨어)와 "Computing Methodologies"(계산방법론)에 관한 논문인 반면, 정보관리학회지의 대다수 논문의 주제가

〈표 9〉 문헌 Set의 년도별 분석

	KISSPAP	KISS93	KSIM	TOTAL
1984			10	10
1985	30		12	42
1986	22		13	35
1987	31		14	45
1988	52		11	63
1989	53		10	63
1990	59		13	72
1991	52		12	64
1992	42		14	56
1993	62	534	7	603
TOTAL	403	534	116	1053

"Information System"(정보시스템)임을 알 수 있다. 따라서 정보관리학회지의 주된 관심이 수학을 기본으로 하는 computing 이론보다는 정보검색, 자동문헌분석, 시스템 자동화인 것이 확연히 나타났다(참조 〈표 10〉과 부록).

4.4 질의어와 적합성 판별에 대한 분석

어떠한 목적에서든지 정보를 얻고자 하는 이용자는 반드시 정보요구을 가지게되며, 데이타베이스를 검색하기 위해서는 자연어 형식의 정보요구를 보다 명확한 개념어 또한 키워드로 변환하게 된다. 본 연구에서는 30 개의 자연어 형태의 질의어를 만들었고, 모든 탐색은 현재 온라인 검색시스템에서 가장 많이 채택하고 있는 검색기법인 불리안 검색을 한다는 전제로, 자연언어질의를 불리안 연산식을 이용한 검색식(불리안 질의)으로 변환하였다. 불 논리(Boolean Logic)를 이용한 탐색문은 정보요구를 나타내는 용어인 탐색어와 이 탐색간의 논리적인 관계로 구성되므로 정보요구를 비교적 정확하고 간단하게 표현할 수 있다.

본 연구에서는 두가지 형태의 불리안 질의를 만들었는데, 그 하나는 자연언어질의를 표현하고 있는 기본 개념만을 이용하여 만든 것이고, 또 하나는 기본 개념을 확장할 수 있는 용어(예: 유사어, 관련어, 영문 표기 등)들 까지 포함한 확장 불리안 질의문이다. 〈표 11〉에서 볼 수 있듯이 모든 질의는 최소 2 개 이상 최대 4 개까지의 기본 개념을 가지고 있고 총 기본 개

〈표 10〉 Test Set의 분류번호 분석

	KISSPAP	KISS93	KSIM	TOTAL
A. 일반문헌	0	0	0	0
B. 하드웨어	29	40	0	65
C. 컴퓨터시스템 구성	92	143	1	224
D. 소프트웨어	147	286	0	416
E. 데이타	14	21	4	34
F. 계산 이론	23	31	0	53
G. 계산 수학	16	30	0	45
H. 정보시스템	100	169	97	345
I. 계산 방법론	186	273	19	459
J. 컴퓨터 응용	12	31	1	41
K. 컴퓨팅 일반	7	12	3	21
X. 정보학	0	0	118	118
TOTAL	626	1036	243	1,905

〈표 11〉 Test Set에 대한 질의어 및 적합성 분석

QUERY #	기본개념의 수	확장어의 수	적합문헌의 수
# 1	2	4	14
2	2	2	25
3	2	5	33
4	2	5	11
5	2	3	17
6	3	5	22
7	2	3	8
8	2	2	10
9	2	2	15
10	2	3	21
11	2	4	9
12	2	3	34
13	3	4	9
14	3	6	13
15	4	8	11
16	2	5	10
17	3	5	12
18	3	4	4
19	2	3	32
20	2	4	6
21	3	4	22
22	2	6	16
23	4	6	12
24	2	5	23
25	4	6	9
26	3	7	5
27	2	4	8
28	3	3	4
29	2	8	5
30	3	8	2

념의 수는 75 개이며 60%의 질의문이 두 개의 기본 개념으로 표현되었음을 알 수 있다. 또한 확장어의 수는 2 개에서 8 개까지의 분포를 보이고 있으며, 총 확장어의 수는 137 개이며 55%의 질의문이 세 개에서 다섯 개의 확장어를 가지고 있음이 나타났다.

정보검색시스템의 평가는 일반적으로 검색효율, 신속성, 경제성의 세가지 측면에서 수행된다(정영미, 1992). 이용자의 정보요구 만족도를 측정하는 평가 기준인 검색효율은 시스템 평가의 가장 기초적인 기준이며 많은 시스템이 검색 효율성에 의거하여 평가되고 있다. 이와 같은 검색효율 측정에 있어서 가장 중요한 개념은 "적합성"이다. 적합성이란 정보자료가 질문에 적합한 내용을 다루고 있는 정도를 말하는 것으로 일반적으로 적합성 판정을 가장 힘든 작업이라 한다. 즉, 질의문에 비추어 각 문헌에 대한 특정한 질문에 대해 검색된 같은 문헌이라도 적합성 판정자에 따라 적합문헌으로 평가될 수도 있고 부적합문헌으로 평가될 수도 있기 때문이다(Bruce, 1994). 그러나 검색효율을 측정하는 실험에서는 질문과 관련된 주제분야의 전문가들로 하여금 적합성 평가를 하게 함으로써 평가의 일관성과 아울러 객관성을 유지하도록 노력한다.

본 연구에서는 적합성 판정에 일관성과 객관성을 유지하기 위하여 네 명의 주제전문가가 각 질의문에 대한 적합문헌을 선택하였고, 어느 정도 일치하였는지 그 통계를 추출하였다. 30 개의 질의문에 대한 총 적합 문헌은 422 문헌이고 평균 14 개의 적합문헌이 선택되었다. 한 질의문에 대한 최대 적합문헌의 수는 34인 반면, 최소 2개인 질의문도 있었다. 특히 "데이타베이

스 AND 객체지향", "정보통신 AND 프로토콜", "병렬처리 AND 알고리즘"에 관한 적합문헌은 30 개가 넘은 반면, 퍼지집합이론과 검색효율에 관한 논문은 2 개 뿐이었다 (참조 〈표 11〉).

네 명의 주제전문가의 적합성 판별의 일치성을 분석한 결과, 84%의 완전일치율을 보였다. 네 사람 중 한 명만 부적합하다고 한 문헌(0.75 값을 가진 문헌)의 수는 49 개로 12%를 보인 반면, 두사람이 적합하다고 보고 나머지 두사람은 부적합하다고 본 문헌(0.50)의 수는 16개로 3%를 차지했으며, 세사람이 부적합하다고 보고 한 사람만 적합하다고 본 문헌(0.25)의 수는 3개(0.7%)로 매우 미미하였다.

5. 결 론

우리글 정보검색시스템과 자동색인기의 객관적인 성능 평가를 위한 실험데이타 컬렉션을 개발한다는 것은 정보검색분야를 한 단계 도약시키는 것이다. 지금까지의 시스템에 대한 성능평가는 시스템을 개발한 팀에 의해 만들어진 실험데이타 컬렉션에 의한 평가이었다고 할 수 있다. 따라서 평가의 기준이 주관적인 측면이 강하였다고 볼 수 있다. 최근들어 우리글 자료를 처리하는 정보검색시스템 및 자동색인기의 개발이 활발히 진행되고 있는 시점에서 시스템을 개발한 팀이 아닌 연구자들에 의해 실험데이타 컬렉션이 만들어졌다는 것은 그 의의가 크다고 할 수 있다.

실험데이타 컬렉션은 정보과학회논문지, 한국정보과학회 1993년 Proceedings, 정보관리학회지에 수록된 1,053개의 논문으로 구성되어 있다. 실험데이타 컬렉션의 입력포맷은 국문 및 영문 서지사항, 분류번호, 색인어 등을 수록할 수 있도록 하였으며, 원문에 나와있지 않은 서지사항은 공란으로 두었다. 색인작업은 원문에 나온대로 따랐지만 명백한 철자오류는 수정을 가하였다. 분류번호와 색인어의 숫자는 제한을 두지 않았지만 분류번호는 5 개 이내로 부여하였다. CRCS분류표가 전산학 위주로 되어있어 문헌정보학 분야 논문들의 분류를 위하여 CRCS분류표에 ISA를 참조하여 'X' 부분을 추가하였다. CRCS분류표의 한국어 번역은 모든 용어를 한국어로 번역하였지만 우리글로 표현하기가 부적당한 것은 영어로 표기하였다.

질의어를 작성하여 질의어에 해당하는 적합한 문헌을 실험데이타 컬렉션에서 추출하여 적합성을 판정하였다. 질의어는 자연언어 질의어, 불리안 질의어, 불리안 확장질의어로 각각 구성하였다. 적합성의 판정은 해당분야의 주제전문가가 수행하였다. 검색된 문헌의 적합성 판정은 주제전문가의 주관성이 개입될 수 있기 때문에 질의어와 수작업으로 검색된 문헌만을 검토하여 판정하도록 함으로써 주관적인 요인을 배제시키도록 하였다.

본 연구에서 개발한 실험데이타 컬렉션이 우리글 정보검색시스템과 자동색인기의 성능을 평가하는 표준 데이타 컬렉션으로 정착시키기 위해서는 데이타 건수의 확장, 주제분야의 확장, 데이타 컬렉션의 무상보급, 데이타 컬렉션의 정

기적인 수정 및 보완, 적합성 판정을 위한 다양한 주제전문가의 참여 등이 수반되어야 할 것이다. 또한 본 연구에서 개발된 Test Set의 모든 문헌이 구축되는 테스트 컬렉션의 성격 또는 목적에 맞게 선정된 것이 아니라 각 년도에 해당되는 모든 문헌을 입력하여 생성된 것으로 구축된 Test Set이 보다 정제되지 못한 단점을 갖고 있다. 따라서 Test Set를 확장시킬 경우 논문집 선정은 물론 기사 논문의 선정의 기준 등을 새로 설정하여 KT Test Set이 그 고유한 특성을 가진 테스트 컬렉션으로 발전시켜야 할 것이다.

참고문헌

김성혁, "문헌정보학 영역의 지능형 정보시스템에 관한 고찰," 정보관리학회지. 9(1), 1992: 165-180.

김현희, 배금표, "퍼지정보검색시스템의 검색효율에 관한 연구," 정보관리학회지. 10(1), 1993: 31-52.

이준호, 김원용, 이윤준, 김명호, "퍼지 집합모델의 검색 효율 개선을 위한 퍼지 연산자의 분석," 정보관리학회지. 10(1), 1993: 53-63.

서은경, "구문 통계적 기법을 이용한 한국어 자동색인에 관한 연구," 정보관리학회지. 10(1), 1993: 97-124.

전기통신용어사전, 대전: 한국전자통신연구소, 1985

정보과학용어사전편찬위원회, 영한/일한 컴퓨터용어 큰사전. 서울: 성안당, 1993

정영미, "우리말 정보자료를 처리하는 지능형 정보검색시스템의 설계," 정보관리학회지. 8(2), 1991: 3-31.

정영미, 정보검색론. 개정판. 서울: 구미무역 출판부, 1992.

최기선, "구문 및 의미분석을 통한 한국어 자동색인," 정보관리학회지. 8(2), 1991: 96-107.

Bruce, Harry, W., "A Cognitive View of the Situational Dynamism of User-Centered Relevance Estimation," *JASIS.* 45(3), 1994: 142-148.

Froehlich, Thomas J., "Relevance Reconsidered-Towards an Agenda for the 21st Century: Introduction to Special Topic Issue on Relevance Research," *JASIS.* 45(3), 1994:124-134.

Hersh, William, "Relevance and Retrieval Evaluation: Perspectives from Medicine," *JASIS.* 45(3), 1994: 201-206.

Information Science Abstract. New York: New York, 1993.

Kemp, D.A. *Computer-Bases Knowledge Retrieval.* London: Aslib, 1988.

Salton, G. and McGill, M.J. *Introduction to Modern Information Retrieval.* NY: McGraw-Hill, 1983.

Saracevic, T., "RELEVANCE: A Review of and a Framework for the Thinking on the Nortion in Information Science," *JASIS.* 26(6), 1975: 331-343.

Saracevic, T., "A Study of Information Seeking and Retrieving III.: Searchers, Searches,

and Overlap," *JASIS.* 39(3), 1988: 197-216.

Su, Louise T., "The Relevance of Recall and Precision in User Evaluation," *JASIS.* 45(3), 1994: 207-217.

부 록

분 류	KISS	KISS93	KSIM	TOTAL	분 류	KISS	KISS93	KSIM	TOTAL
A.0 일반문헌 일반					H.0정보시스템일반				
A.1 입문서					H.1 모델 및 원리			5	5
A.2 참고문헌					H.2 DB관리	84	68	2	154
A.m 기타					H.3정보 저장/검색	10	25	82	117
B.0 하드웨어 일반					H.4정보시스템응용	2	13		15
B.1 제어구조	3	1		4	H.5정보인터페이스	4	63	8	75
B.2 연산,논리구조	1	1		2	I.0계산방법론일반				
B.3 메모리구조	2	12		14	I.1대수 조작	2			2
B.4 데이타통신	8	11		19	I.2인공지능	104	124	15	243
B.5 레지스터	3	7		10	I.3컴퓨터그래픽스	19	41		60
B.6 논리설계	4	1		5	I.4영상처리	22	38	1	61
B.7 집접회로	8	7		15	I.5패턴인식	31	41	2	74
B.m 기타					I.6시뮬레이션	2	8		10
C.0 컴퓨터 시스템		4		4	I.7텍스트 처리	6	21	1	28
C.1 프로세서	18	35		53	I.m기타				
C.2 컴퓨터 통신	66	80	1	147	J.0컴퓨터응용일반				
C.3 특수 목적	2	4		6	J.1사무데이타처리	3	2		5
C.4 시스템 성능	6	2		8	J.2물리과학및공학	3	3		6
C.5 시스템 구현		18		18	J.3생명과학및의학	1	2	1	4
C.m 기타					J.4사회및행동과학		1		1
D.0소프트웨어일반					J.5예술및인문과학		2		2
D.1프로그래밍기술	26	79		105	J.6컴퓨터이용기술	4	4		8
D.2소프트웨어공학	48	76		124	J.7기타 응용	1	16		17
D.3프로그래밍언어	37	28		65	J.m기타		1		1
D.4 운영체제	36	103		139	K.0 Computing일반				
D.m 기타					K.1 컴퓨터 산업				
E.0 데이타 일반					K.2 컴퓨팅의 역사		1		1
E.1데이타 구조	4	5		9	K.3 컴퓨터와 교육		8	2	10
E.2데이타저장표현	1	2		3	K.4 컴퓨터와 사회		1		1
E.3 데이타 암호화	1	6		7	K.5 법적인 면			1	1
E.4코딩과정보이론	5	5	4	14	K.6정보시스팀관리	6	1		7
E.5화일	3	3		6	K.7 컴퓨팅 전문업				
E.m 기타					K.8 개인 컴퓨팅	1	1		2
F.0 계산이론 일반					K.m 기타				
F.1 추상장치	2	8		10	X.0 정보학 일반				
F.2 알고리즘 분석	10	20		30	X.1정보학 및 문헌			11	11
F.3프로그램의논리	11	2		13	X.2연구 방법론			14	14
F.4 수학적 논리		1		1	X.3정보생성및보급			7	7
F.m 기타					X.4정보인식맡기술			14	14
G.0 계산수학 일반		12		12	X.5정보처리및제어			31	31
G.1 수치해석	9	15		22	X.6정보시스팀			28	28
G.2 이산수학	7			7	X.7정보관리			13	13
G.3 확률/통계		3		3					
G.4수학소프트웨어									
G.m 기 타									