□ 기술해설 □

정보검색 시스템 평가 및 테스트 컬렉션 개발

연구개발정보센터 조영환·박혁로
숭실대학교 이준호*

1. 서 론

지난 30년 동안 과학과 기술 분야의 급속한 발전은 수많은 주제들에 대하여 방대한 양의 정보가 생성되는 정보화 사회를 탄생시켰다. 원하는 정보에 대한 정확하고 빠른 접근은 정보화 사회를 살아가는 현대인들에게 성공의 여부를 결정짓는 중요한 요소이다. 그러나 수많은 주제들에 대한 대용량의 정보로부터 한정된 시간 내에 원하는 정보를 발견하는 것은 매우 어려운 일이다. 이러한 문제를 해결하기 위해 1960년대 초에 컴퓨터를 이용하여 원하는 정보를 찾도록 도와주는 정보검색(Information Retrieval)이라는 연구 분야가 확립되었다.

지금까지 컴퓨터를 이용하여 대용량의 문서를 효율적으로 검색할 수 있는 정보검색 시스템에 관한 많은 연구들이 수행되어 왔다. 정보검색 시스템의 사용은 원하는 정보에 대한 접근을 용이하게 함으로써 다양한 분야의 정보들에 대한 수집 시간과 노력을 단축시킨다. 특히 관리할 정보의 양이 기하급수적으로 증가하고 있는 정보화 시대에서 효율적인 정보검색 시스템에 대한 요구는 더욱 절실하다.

외국에서는 이미 1960년대 초에 일괄 처리 방식에 의한 MEDLARS 시스템이 구축된 이후 STAIRS, BASIS, BRS, TOPIC 등과 같은 많은 정보검색 시스템들이 상용화되어 현재까지 사용되고 있다. 국내에서는 1990년대 중반까지도 정보검색 시스템에 대한 인식이 매우 부족한 상태이었다. 그러나 네트워크의 발달로 인터넷 및 인트라넷을 통한 정보 서비스가 활성화되면서, 정보검색 시스템에 대한 중요성이 인식되었으며, 이에 따라 한글의 특성을 보다 잘 반영할 수 있는 정보검색 시스템들이 개발되기 시작하였다. 특히, 웹페이지들을 검색할 수 있는 심마니, 네이버, 정보탐정, 까치네와 같은 정보검색 시스템들이 모두 국내 기술로 개발된 것은 주목할 만하다.

이처럼 정보검색에 대한 관심이 고조되고 정보검색 시스템에 대한 개발이 국내의 기술로 이루어지고 있음에도 불구하고, 지금까지 정보검색에 대한 연구 및 개발에 투자된 시간과 노력이 외국에 비하여 상대적으로 적었기 때문에, 향후 정보의 양이 보다 급속히 증가할 경우 현재 국내의 정보검색 기술로서 외국의 시스템들과 경쟁하는데는 많은 어려움이 예상된다. 따라서 대용량의 데이타를 대상으로 하는 정보검색 기술에 대한 확보가 시급히 요구되며, 이를 위해서는 정보검색 시스템의 성능들을 비교할 수 있는 평가 항목들에 대한 조사와, 정보검색 시스템의 성능을 평가할 수 있는 테스트 컬렉션의 개발이 선행되어야 한다.

2. 시스템 평가를 위한 기본 모형

정보검색 시스템은 문서의 등록과 저장 그리고 검색의 기능을 모두 포함한다. 그러므로 정보검색과 질의 재구성 알고리즘의 평가를 위하여 제안된 정확률과 재현율[1]만으로 정보검색 시스템을 평가하기에는 부족함이 있다. 반면에 상용의 정보검색 시스템 혹은 패키지에서 제공하는 부가적인 기능인 Web 인터페이스, 에이전트 기능, Z39.50 프로토콜, 디렉토리 서비스

*종신회원

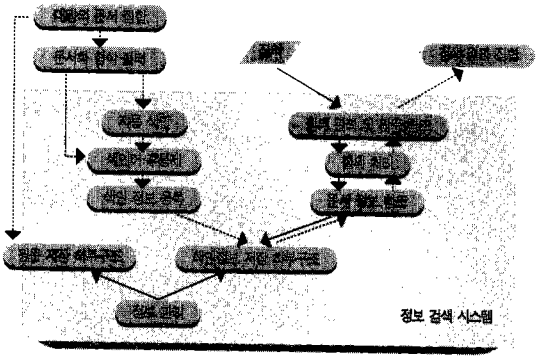

그림 1 정보검색엔진의 기본 모형

등은 시스템 도입자의 목적에 의존되는 것이고 지속적인 추가가 가능한 것이므로 평가 항목의 범위로는 적합하지 않다.

상용의 정보검색 시스템을 도입하기 위해서는 1) 색인어 추출의 품질 및 색인에 소요되는 시간, 2) 대량의 데이타에 대한 적재 경험 및 적재 시간, 3) 시스템 구성을 위한 S/W, H/W, Media, Interaction의 지원, 4) 지식 관리, 계정 관리 등의 검색 보조기능에 대한 검토가 필요하다.

본 논문에서 대상으로 하는 평가 항목은 정보검색 시스템의 기본 기능 즉, 문서를 등록, 저장 그리고 검색하는 기능으로 제한하여 각각의 부분에 어떠한 평가 항목이 필요한 지를 검토하고자 한다. 이러한 검토는 각각의 처리에 대한 기능적인 평가 항목과 성능적인 평가항목으로 나누어 고려한다. 이를 위하여 정보검색 시스템의 기본 엔진과 처리의 과정을 다음의 그림 1과 같이 정의하기로 한다. 그림 1에서 점선은 데이타의 흐름을 실선은 질의의 흐름을 의미한다.

정보검색 시스템을 구축하는 관점에서 작업의 순서에 따라 평가항목을 나누어 보면 1) 문서의 등록에 관련된 평가 항목, 2) 정보 질의에 관련된 평가 항목, 3) 정보 관리에 관련된 평가 항목 등의 3가지로 구분하여 나누어진다. 그러나 현실적으로 정보검색 시스템에서의 한국어 처리는 대단히 중요한 역할을 하므로 4) 색인어 추출에 관련된 평가 항목을 추가하여 본 논문에서는 4가지 부분에서 요구되는 기능적인 점검 항목과 성능적 평가 항목을 설정하도록 한다.

3. 기능적 점검 항목

정보검색 시스템의 기능에 대하여 기본적으로 갖추어야 하는 필수적 기능과 응용 시스템에 따라 요구되는 부가적인 기능으로 구분하여 검검항목을 설정한다.

3.1 색인어 추출에 관련된 점검 항목

정보검색 시스템은 저장된 문서에 대한 문서 표현 수단으로 색인어와 색인어의 위치를 이용하고 사용자는 색인어와 이들간의 관계를 통하여 검색하기를 원하는 문서집합을 표현한다. 그러므로 저장할 문서에 대한 내부 표현의 기본이 되는 색인어의 추출은 정보검색 시스템에서 매우 중요한 평가 기준으로 고려되야 한다.

- 필수적 기능
  - 한자를 한글로 변환하는 기능
  - 다양한 종류의 코드로 된 원문의 처리
  - 분야별 명사 사전의 제공
  - 시스템/사용자/불용어 사전의 제공
  - 복합명사 분리 기능
  - 한영 혼합 색인어 추출 기능
  - 품사별로 색인어를 선택하는 기능
  - 색인어의 타입을 정의하는 기능
  - 미등록어 추정 기능
- 부가적 기능
  - 띄어쓰기 오류를 포함한 경우에 대한 처리
  - 사전 표제어를 표준 색인어로 변환하는 기능

색인어의 추출기능에서 중요한 것은 정보검색의 대상이 되는 분야에 적정한 색인어를 추출하는 것이다. 특히 한국어의 경우에는 색인어의 대상이 띄어쓰기의 단위가 되는 어절로는 적합하지 않기 때문에 응용분야에 적합한 명사 혹은 동사 및 형용사의 추출 여부는 매우 중요한 점검 요소이다.

3.2 문서 등록에 관련된 점검 항목

문서를 등록하는 과정은 해당 문서에 대하여 적절한 형태의 색인을 추출하여 각 색인어가 발생한 문서에 대한 정보를 작성하는 부분이다. 이때에는 원문에 대한 형식의 정의와 이의 검사 그리고 실제의 작업에 대하여 개념적 처

리 기능을 제공하여야 한다.

- 필수적 기능
  - 논리적 문서 모음(Collection)의 개념이 제공
  - 원문 화일 포멧에 대한 검사
  - 제공되는 원문에 대한 필터의 종류
  - 다양한 색인어 추출 방식 지원
  - 필드별로 색인어 추출 방식을 지정하는 기능
  - 추출된 색인어를 후통제하는 기능
  - On-Line 방식으로 문서를 등록/수정하는 기능
- 부가적 기능
  - 동시에 여러 프로세스가 문서를 등록하는 기능
  - 문서의 등록과 검색이 동시에 이루어지는 기능
  - DB의 이상적 사태에 대한 동일성 유지 기능
  - 논리적 문서 모음에서 이질적인 문서 형식을 지원

문서의 등록기능에서 중요한 것은 다양한 오퍼레이션을 제공하는 것이다. 필요한 Primitive Operator를 제공함으로써 이들의 조합으로 원하는 기능을 구현하는 방법을 지원하는 것과 더불어 "논리적 문서 집합"과 같이 미리 정의된 고급 수준의 개념을 제공하는 것이 시스템을 평가하는 매우 중요한 점검항목이 된다. 반면에 HWP 파일 필터 등과 같이 제공되는 문서 필터는 시스템의 전체적인 구조에 미치는 영향이 작다.

3.3 정보 질의에 관련된 점검 항목

정보 질의에 대한 평가는 사용자의 측면에서 "검색하기를 원하는 문서 집합에 대한 표현" 즉, 질의어에 대한 풍부함에 기초되어야 한다. 이와 더불어 표준 질의언어의 지원과 다회성 질의를 지원하는 기능도 고려되어야 한다.

- 필수적 기능
  - 질의 표현과 저장된 문서 표현의 일치
  - 표준 질의어 형식을 지원
  - 좌/우 절단된 색인어 지원
  - 색인어간의 위치 관계 연산자 지원
  - 정형 필드와 비정형 필드 지원
  - 문서집합에 대한 가중치 부여 기능
  - 검색 결과를 질의에 따라 순서화하는 기능
  - 이전의 검색 결과 집합에 대한 history 지원
  - 이전의 질의문을 피드백하는 기능
  - 색인어 열람 기능
  - 임의의 건수로 결과를 제한하는 기능
  - 결과에 검색어가 특정 마크로 표시되는 기능
- 부가적 기능
  - 검색된 문서의 요약 기능
  - 질의에 사용된 단어의 의미 이해
  - 질의를 고려하여 문서를 요약하는 기능
  - 사용할 시소러스를 선택하는 기능

정보 질의 기능에서 중요한 것은 질의자의 의도를 적절히 반영하는 것이다. 그러므로 질의를 작성하는 작업에 도움을 주는 기능과 질의에 포함된 단어와 시스템 내부의 단어를 일치시키는 작업이 수행되어야 한다.

3.4 정보 관리에 관련된 점검 항목

정보 관리에 대한 평가는 관리자의 측면에서 시스템의 구성과 저장된 문서의 관리에 대한 편의성에 기초되어야 한다. 문서의 표현에 관련하여 어휘를 통제하고 저장된 문서의 내용을 변경하는 등의 작업을 지원하는 것도 필요하다.

- 필수적 기능
  - 사용자/불용어 사전 관리기 지원
  - 시소러스 관리기 지원
  - 저장된 문서의 On-line Update 기능
  - 색인, 저장, 검색, 관리에 대한 지침서 제공
  - 구조적으로 정돈된 API를 제공
  - 쉬운 인스톨을 제공
  - 시스템의 모니터링 기능 지원
  - 변경중인 문서에 대한 정보 제공
- 부가적 기능
  - 사용자의 계정 관리 기능 지원
  - 사용자의 세션에 대한 로그작성
  - 시스템 프로세스의 부하 조정 기능 지원
  - 색인 정보의 관리 기능 지원

정보 관리 기능에서 중요한 것은 관리자가 효율적이고 안정적으로 작업할 수 있도록 지원하는 것이다. 그리고 불필요한 색인어에 대한 통제를 가하거나 요청이 많은 문서 집합의 검색 프로세스에 우선권을 부여하는 등의 유연성을 제공하는 것도 중요하다.

4. 성능적 평가 항목

정보검색 시스템의 성능적 평가는 시스템 운영에 대한 예측을 충족시키는지에 대한 판단의 측면에서 다루어져야 한다. 본 논문에서는 시스템의 기본적인 기능에 관하여 적절한 평가 단위를 설정한다. 이러한 성능적 평가 항목에 대한 객관성을 갖추기 위해서는 5장 이후에서 설명될 테스트 콜렉션을 사용하여야 한다.

4.1 색인어 추출에 관련된 평가 항목

정보검색 시스템에서 문서에 대한 표현으로 명사 등의 특정 단어를 사용하지 않고 N-gram 방법을 도입하는 것도 가능하다. 그러나 본 논문서는 형태소 분석의 방법을 적용하여 색인어를 추출하는 것에 초점을 두어 평가 항목을 설정한다.

- 대량 색인어 추출 속도
  - 방법 : 1 Mega Byte문서를 처리하는 시간
- 소량 색인어 추출 속도
  - 방법 : 1 KiloByte 문서를 처리하는 시간
- 사전 표제어의 크기
  - 방법 : 품사별 단어의 개수
- 색인의 정확성
  - 방법 : 전문가의 색인어와 자동 색인어 비교
- 색인어 추출의 견고성
  - 방법 : 최악의 경우 테스트 문서에 대한 색인

색인의 정확성을 위하여 테스트 콜렉션에는 색인어 목록이 존재하여야 한다. 이때의 평가 방법은 Precision과 Recall을 사용한다. 그리고 색인어 추출의 견고성에 대한 테스트를 위하여 깨어진 글자와 1000자 이상의 문자열 등의 경우를 모두 고려한 테스트 문서가 준비되어야 한다.

4.2 문서 등록에 관련된 평가 항목

문서 등록에 관련된 성능적 평가는 초기의 색인 정보 작성에 드는 비용적인 측면과 추후에 지속적인 관리에 드는 비용의 측면에 주안점을 두어야 한다.

- 색인어별 문서 등록 속도
  - 방법 : 10만 keyword에 대한 등록 속도
- 건수별 문서 등록 속도
  - 방법 : 1K 크기의 10만 문서에 대한 등록 속도
- 최대 동시 문서 등록 프로세스의 수
  - 방법 : 프로세스의 수와 등록 속도의 증가 비율
- 건수의 증가에 대한 등록 시간의 감소율
  - 방법 : 초기 10%부터 점진적인 속도 증가 함수
- 문서 수정 속도
  - 방법 : 1개의 문서를 N번 수정하는 시간
- 원문 대비 색인 정보의 비율
  - 방법 : 색인필드의 전체 크기와 색인정보 저장공간
- 최대로 저장이 가능한 전체 DB의 크기
  - 방법 : 최대로 저장이 가능한 1Kbyte 문서의 갯수

문서 등록의 평가 항목은 정보 질의의 점검 항목과 연관되어 고려되어야 한다. 예를 들어 근접 검색을 충실히 지원하기 위해서는 문서내의 위치 정보가 필요하므로 원문 대비 색인 정보의 비율이 증가하게 된다. 그러므로 위의 평가 항목을 적용하기 위한 전제조건으로 동일한 정보 질의의 점검 항목을 공유하여야 한다.

4.3 정보 질의에 관련된 평가 항목

정보 질의에 대한 성능적 평가는 사용자의 편리성에 기초되어야 한다. 이는 결과의 충실도와 검색 속도의 측면으로 나누어 고려할 수 있다.

- 단일 키워드에 대한 검색 속도
  - 방법 : 결과가 각각 10건/10000건인 색인어 100회
- 검색된 문서 중에서 질의에 적합한 문서의 비율

• 방법 : 결과의 top 20에 포함된 정도
—적합한 문서 중에서 검색된 문서의 비율
• 방법 : 결과의 top 20에 포함된 정도
—최대 동시 검색 프로세스의 수
• 방법 : 프로세스와 검색 속도의 증가 비율

정보 질의에 대한 충실도의 평가는 정보검색과 질의 재구성 알고리즘의 평가를 위하여 제안된 정확률과 재현율을 도입하여야 한다. 그러나 검색 반응 속도에 대하여는 테스트 콜렉션의 질의에 대하여 다양한 환경에서 검색 시간을 측정하여야 한다.

5. 테스트 컬렉션의 중요성

정보검색 시스템의 검색 효과(Retrieval Effectiveness)를 향상시키기 위하여 색인어 가중치, 자연어 처리, 적합서 피드백 등을 이용한 다양한 검색 기법들이 개발되고 있다. 정보검색 분야의 연구에 있어서 특징적인 사항중의 하나는 많은 경우에 직관적 통찰에 의해 개발된 검색 기법들이 검색 효과의 향상을 가져오지 않는다는 것이다. 이에 대한 예로서 시소러스를 사용함으로써 기대했던 만큼의 검색 효과의 향상을 얻는데 실패하여 왔음을 들 수 있다[3]. 따라서 개발중인 검색 기법의 성능을 평가할 수 있는 테스트 컬렉션은 검색 기법의 개발에 있어서 필수적인 요소이다.

정보검색에 있어서 실험은 오랜 역사를 지니고 있다. 정보검색에 대한 연구는 Cranfield I [4]이라고 불리는 색인에 대한 실험과 함께 시작되었으며, 그 후로 30년이 넘는 동안 실험은 검색 기법의 개발에 있어서 필수적인 요소로 인식되어 왔다. Cranfield II[5]에서는 컴퓨터에 의한 자동 색인과 사람에 의한 수작업 색인을 비교하였으며, 자동 색인에 대한 긍정적인 연구 결과는 정보검색에 대한 연구를 활성화시키는 계기가 되었다. 1960년대 말에 생성된 Cranfield 컬렉션은 1,400개의 문서와 225개의 질문으로 구성되어 있으며, 그 후로 많은 사람들에 의해 활발히 이용되어 왔다. 이 외의 테스트 컬렉션으로는 CACM 컬렉션[6], NPL 컬렉션[7] 등이 있으며, 특히 1992년도부터 매년 개최되고 있는 TREC(Text REtrieval Conference)[3]에서는 100만건이 넘는 대용량 테스트 컬렉션을 구축하고 있으며, 해마다 테스트 컬렉션을 크기를 증가시키고 있다.

앞에서 설명된 바와 같이 외국에서는 다양한 테스트 컬렉션들이 개발되어 정보검색에 대한 연구에 많이 이용되어 왔다. 그러나 이러한 테스트 컬렉션들에 포함된 문서들은 모두 한글과는 특성이 매우 다른 영어로 작성되어 있다. 예를 들면, 영어는 단어의 구분이 분명하여 색인 과정이 단순한데 비하여, 한글은 띄어쓰기의 자유로움과 조사의 발달로 인하여 색인 과정에 어려움을 지니고 있다. 따라서 한글 문서들로 구성된 테스트 컬렉션은 한글 정보검색 연구를 위한 필수적인 요소이다.

6. 한글 테스트 컬렉션 개발 현황

국내의 경우 한글 문서들로 구성된 테스트 컬렉션의 필요성은 인식되어 있으나, 테스트 컬렉션 구축에 따른 어려움으로 인하여 개발이 미미한 실정이다. 지금까지 개발된 한글 테스트 컬렉션은 KT 컬렉션[8]과 KRIST 컬렉션[9]이 있다. KT 컬렉션은 정보과학회 논문지, 한국정보과학회 1993 Proceedings, 정보관리학회지에 수록된 1,053개의 논문들과 30개의 매우 단순한 질문을 포함하고 있다. KRIST 컬렉션은 연구개발정보센터 소유의 연구보고서 데이타베이스를 이용하여 개발되었으며, 32개의 필드로 구성된 13,515개의 문서와 30개의 질의, 그리고 각 질의에 대한 적합 문헌 리스트로 구성되어 있다.

7. 테스트 컬렉션 개발 방법

정보검색용 테스트 컬렉션은 일반적으로 문서 집합, 질의 집합 그리고 각각의 질의에 대한 적합 문헌 리스트로 구성된다. 다음에서는 테스트 컬렉션을 구성하는 문서들의 선정과 질의 작성, 그리고 적합 문헌 리스트를 생성하는 방법에 대하여 기술한다.

7.1 문서 집합

검색의 대상이 되는 문서 집합은 테스트 컬

렉션 구축에 있어서 가장 기본적인 요소이다. 문서 집합의 구성에 있어서 유의해야 할 사항은 다양한 분야의 문서들로 문서 집합을 구성해야 한다는 것이다. 특히, 정보검색 기술의 기본이 되는 가중치 기법들 중 일부 가중치 기법은 특정 크기의 문서들에 높은 유사도 값을 부여하는 특성을 지니고 있기 때문에, 가중치 기법의 성능 평가를 위해서라도 다양한 크기의 문서들로 문서 집합을 구성하는 것이 바람직하다.

7.2 질의 작성

질의의 내용을 특정 분야로 편중시킬 경우, 개발된 테스트 컬렉션을 일반적인 정보검색 시스템의 성능 평가에 활용하기 어렵기 때문에, 질의의 내용은 가능한 한 일부 분야에 치우치지 않고 여러 분야에 골고루 분포시키는 것이 바람직하다. 이를 위하여 연구개발정보센터의 지원하에 이화여자대학교에서 수행하였던"질의 작성 및 적합문헌 선정"에 관한 연구[10, 11]를 살펴보고자 한다.

이 연구에서는 한국경제신문에 게재된 신문기사 211,216건을 대상으로 50개의 질의를 작성하고, 각각의 질의에 대하여 연구개발정보센터에서 선정한 1,000개의 적합 문헌 후보들에 대하여 적합성 여부를 판별하였다. 이때 작성된 50개의 질의는 다음과 같은 방법으로 작성되었다.

테스트 컬렉션의 문서 집합이 신문기사이므로, 질의가 작성될 분야를 한국언론연구원·한국조사기자회(1992)의 「기사자료표준분류표」를 참조하여 총류, 정치, 경제, 산업, 사회, 사건·사고, 문화, 과학, 스포츠, 국제와 같이 10개 분야로 대분류하였다. 그리고 이러한 10개의 분야들을 총류(6), 정치(7), 경제(5), 산업(5), 사회(6), 사건·사고(6), 문화(4), 과학(4), 스포츠(3), 국제(4)의 50 주제 그룹으로 세분하였으며, 이 50 주제 그룹중 분야가 광범위한 것은 임의로 다시 세분하여 최종적으로 65 주제 그룹을 생성하였다. 이때 질의는 질의의 내용이 65주제들에 고르게 분포되도록 작성되었다.

7.3 적합 문헌 후보 리스트 생성

각각의 질의에 대한 적합 문헌 후보 리스트의 생성은 테스트 컬렉션 구축에 있어서 가장 중요한 요소이다. 적합 문헌 리스트의 생성을 위한 가장 초보적인 방법은 각각의 질의에 대하여 테스트 컬렉션의 문서 집합에 포함된 모든 문서들을 읽고 적합성 여부를 판단하는 것이다. 그러나 이 방법은 문서의 수가 많은 경우에 대단히 많은 시간을 요구하며, 대용량 테스트 컬렉션의 개발에 있어서는 현실적으로 불가능하다.

한편, 서로 다른 정보검색 시스템들은 질의와 문서에 대하여 서로 다른 표현 기법을 사용하고, 또한 서로 다른 특성의 검색 기법들을 사용하기 때문에, 동일 질의에 대하여 서로 다른 집합의 문서들을 검색하는 것으로 알려져 있다. 이러한 특성을 이용하여 다수의 정보검색 시스템들을 사용하여 검색을 수행하고, 각각의 시스템에 의해 높은 순위를 부여받은 문서들에 대하여 적합성 여부를 판단하는 방법이 제안되었다[3]. 풀링 방법(Pooling Method)이라고 불리는 이 방법은 주어진 질의에 대하여 테스트 컬렉션을 구성하는 모든 문서들의 적합성 여부를 판별하는 대신에, 정보검색 시스템들에 의해 검색된 일부의 문서들에 대해서만 적합성 여부를 판별하기 때문에, 대용량 테스트 컬렉션 구축시 적합 문헌 리스트 생성에 효과적인 방법으로 알려져 있다.

7.4 적합 문헌 선정

적합 문헌 선정 작업은 각각의 질의에 대하여 생성된 적합 문헌 후보들을 사람이 직접 검토함으로써 이루어진다. 이때 한 사람보다는 두 사람이 적합 문헌 선정 작업을 수행하는 것이 바람직하며, 이 경우 두 사람의 적합성 판정 결과가 일치하지 않는 경우도 발생한다. 일반적으로 적합 문헌 선정 과정에서 A라는 사람이 판정한 적합 문헌과 B라는 사람이 판정한 적합 문헌들 중에서 약 30% 정도의 문헌들이 일치하지 않는 것으로 알려져 있다. 이 경우, 적합 문헌 리스트에는 A와 B의 적합 문헌들의 합집합 또는 교집합 중 어느 것을 사용해

도 무관한 것으로 알려져 있다.

8. 결 론

네트워크의 발달로 인터넷 및 인트라넷을 통한 정보 서비스가 활성화되면서, 정보검색 시스템에 대한 중요성이 인식되었으며, 이에 따라 한글의 특성을 보다 잘 반영할 수 있는 정보검색 시스템들의 개발이 국내의 기술로 이루어지고 있다. 그러나, 향후 정보의 양이 보다 급속히 증가할 경우 현재 국내의 정보검색 기술로서 외국의 시스템들과 경쟁하는데는 많은 어려움이 예상된다. 본 연구에서는 국내의 정보검색 시스템 개발의 방향을 제시하기 위하여, 정보검색 시스템에 대한 사용자들의 요구사항을 정리하였다.

한편, 정보검색에 대한 연구에 있어서 테스트 컬렉션은 수행중인 정보검색 연구 결과의 우수성을 입증하기 위한 필수적인 요소로 인식되어 왔다. 그러나 최근 한글 정보검색에 대한 관심이 급속히 확산되었음에도 불구하고, 한글 정보검색용 테스트 컬렉션의 부족으로 인하여 한글 정보검색에 대한 연구에 어려움을 겪고 있다. 따라서 대용량 한글 정보검색용 테스트 컬렉션의 개발이 시급하며, 본 연구에서는 대용량 테스트 컬렉션 개발을 위한 방법을 기술하였다.

참고문헌

[1] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw~Hill, Inc., 1983.

[2] J.H. Lee and J.S. Ahn,"Using N~Grams for Korean Text Retrieval," ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 216~224, 1996.

[3] D. Harman,"Overview of the 1st Text Retrieval Conference," ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 36~48, 1993.

[4] C.W. Cleverdon,"Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems," College of Aeronautics, Cranfield, England, 1962.

[5] C.W. Cleverdon, J. Mills, and E.M. Keen,"Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test Results," Aslib Cranfield Research Project, Cranfield, England, 1966.

[6] E. Fox,"Characteristics of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts," Technical Report TR 83~561, Cornell University, Computer Science Department, 1983.

[7] K. Sparck Jones and C.A. Webster, "Research in Relevance Weighting," British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1979.

[8] 김성혁, 서은경, 이원규, 김명철, 김영환, 김재군,"자동색인기 성능시험을 위한 Test Set 개발," 정보관리학회지, Vol. 11, No. 1, pp. 82~101, 1994.

[9] 이준호, 최광남, 한현숙, 김종원, 남성원, "정보 검색 연구를 위한 KRIST 테스트 컬렉션의 개발," Vol. 12, No. 2, pp. 225~232, 1995.

[10] 연구개발정보센터,"질의작성 및 적합문헌 선정," 1997.

[11] 연구개발정보센터, "질의작성 및 적합문헌 선정," 1998.

조 영 환

1989 연세대학교 전산과학과(학사)
1991 한국과학기술원 전산학과(석사)
1997 한국과학기술원 전산학과(박사)
1997～현재 연구개발정보센터 선임연구원
관심분야 : 정보 검색 인터페이스, 자연어처리, 대화시스템
E-mail:choyh@ns.kordic.re.kr

박 혁 로

1987 서울대학교 컴퓨터공학과(학사)
1988 한국과학기술원 전산학과(석사)
1995～현재 연구개발정보센터 선임연구원
1997 한국과학기술원 전산학과(박사)
관심분야 : 정보검색, 자연어 처리, 한국어 정보 처리
E-mail:hrpark@ns.kordic.re.kr

이 준 호

1987 서울대학교 컴퓨터공학과(학사)
1989 한국과학기술원 전산학과(석사)
1993 한국과학기술원 전산학과(박사)
1993～1994 한국과학기술원 인공지능연구센터 연구원
1994～1995 코넬대학교 전산학과 방문연구원
1994～1997 연구개발정보센터 선임연구원
1997～현재 숭실대학교 컴퓨터학부 조교수
관심분야 : 정보검색, 정보시스템, 데이타베이스
E-mail : joonho@computing.soongsil.ac.kr