정보 검색 연구를 위한 KRIST 테스트 컬렉션의 개발

Developing the KRIST Test Collection for Researches in Information Retrieval

이준호(Joon-Ho Lee)* · 최광남(Kwang-Nam Choi)** · 한현숙(Hyun-Sook Han)*** · 김종원(Jong-Weon Kim)*** · 남성원(Seong-Won Nam)***

□ 목 차 □

초 록

정보 검색에 대한 연구를 위해 테스트 컬렉션은 필수적인 요소로 인식되어 왔다. 외국의 경우, 다양한 테스트 컬렉션들이 개발되어 정보 검색에 대한 연구에 이용되어 왔다. 그러나 국내의 경우, 최근에 한글 정보 검색에 대한 관심이 확산되었음에도 불구하고 정보 검색용 테스트 컬렉션에 대한 부족으로 인하여 한글 정보 검색에 대한 연구에 어려움을 겪고 있다. 본 연구에서는 연구개발정보센터 소유의 KRIST 데이타베이스를 기반으로 하여 개발된 KRIST 테스트 컬렉션에 대하여 기술한다. KRIST 테스트 컬렉션은 과기처 연구보고서에 대한 서지 레코드 13,515건과 30개의 자연어 질의 그리고 각 질의에 대한 적합 문헌리스트로 구성된다.

ABSTRACT

It has been known that test collections play an important role for researches in information retrieval. A variety of test collections have been created in foreign countries, and have been heavily used by researchers. Although research interests in Hangul information retrieval have been rapidly grown up in Korea these days, lack of Hangul test collections makes it difficult to develop retrieval techniques for Hangul texts. This study describes the development of the KRIST test collection. The KRIST test collection consists of 13,515 bibliographic records, 30 queries and a list of relevant documents to the queries.

* 연구개발정보센터 연구개발부 선임연구원
** 연구개발정보센터 연구개발부 연구원
*** 연구개발정보센터 정보사업부 선임연구원

■ 논문접수일 : 1995년 11월 1일

1. 서 론

지난 30년동안 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 요소가 되었다. 그러나 대용량의 데이터로부터 주어진 시간 내에 원하는 정보를 발견하는 것은 매우 어려운 일이다. 이러한 문제점을 해결하기 위해 1960년도 초에 컴퓨터를 이용하여 지정된 정보를 검색하는 정보 검색(Information Retrieval)이라는 연구 분야가 확립되었다.

지금까지 컴퓨터를 이용하여 대용량의 문서를 효율적으로 검색할 수 있는 정보 검색 시스템에 관한 많은 연구가 이루어져 왔다. 정보 검색 시스템의 사용은 원하는 정보에 대한 접근을 용이하게 함으로써 여러 분야에 있어서 정보 수집에 대한 시간과 노력을 단축시키게 된다. 특히 관리할 정보의 양이 기하급수적으로 증가하고 있는 정보화 시대인 오늘날에는 효율적인 정보 검색 시스템에 대한 요구는 더욱 절실하다.

정보 검색 시스템의 검색 효과(Retrieval Effectiveness)를 향상시키기 위하여 색인어 가중치, 자연어 처리, 적합성 피드백 등을 이용한 다양한 검색 기법들이 개발되고 있다. 정보 검색 분야의 연구에 있어서 특징적인 사항중의 하나는 많은 경우에 직관적 통찰에 의해 개발된 검색 기법들이 검색 효과의 향상을 가져오지 않는다는 것이다. 이에 대한 예로서 구(Phrase) 또는 시소러스(Thesaurus)를 사용함으로써 기대했던 만큼의 검색 효과의 향상을 얻는데 실패하여 왔음을 들 수 있다(Harman 1993). 따라서 개발중인 검색 기법의 성능을 평가할 수 있는 테스트 컬렉션은 검색 기법의 개발에 있어서 필수적인 요소이다.

정보 검색에 있어서 실험은 오랜 역사를 지니고 있다. 정보 검색에 대한 연구는 Cranfield Ⅰ(Cleverdon 1962)이라고 불리는 색인에 대한 실험과 함께 시작되었으며, 그 후로 30년이 넘는 동안 실험은 검색 기법의 개발에 있어서 필수적인 요소로 인식되어 왔다. Cranfield Ⅱ (Cleverdon et al. 1966)에서는 컴퓨터에 의한 자동 색인과 사람에 의한 수작업 색인을 비교하였으며, 자동 색인에 대한 긍정적인 연구 결과는 정보 검색에 대한 연구를 활성화시키는 계기가 되었다. 1960년대 말에 생성된 Cranfield 컬렉션은 1400개의 문서와 225개의 질문으로 구성되어 있으며, 그 후로 많은 사람들에 의해 활발히 이용되어 왔다. 이 외의 테스트 컬렉션으로는 CACM 컬렉션(Fox 1983), NPL 컬렉션(Sparck Jones & Webster 1979) 등이 있다.

앞에서 설명된 바와 같이 외국에서는 다양한 테스트 컬렉션들이 개발되어 정보 검색에 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 띄어 쓰기의 자유로움과 조사의 발달로 인하여 색인 과정에 어려움을 지니고 있다(이준호 외 1995). 따라서 한글 문서들로 구성된 테스트 컬렉션은 한글 정보검색 연구를 위한 필수적인 요소이다.

국내의 경우 한글 문서들로 구성된 테스트

컬렉션의 필요성은 인식되어 왔으나, 테스트 컬렉션 구축에 따른 어려움으로 인하여 개발이 미미한 실정이다. 단지 최근에 개발된 테스트 컬렉션으로서 KT 컬렉션(김성혁 1994)이 있다. KT 컬렉션은 정보과학회논문지, 한국정보과학회 1993 Proceedings, 정보관리학회지에 수록된 1,053개의 논문들과 30개의 매우 단순한 질문을 포함하고 있다. 본 연구에서는 연구개발정보센터 소유의 연구보고서 데이타베이스를 이용하여 개발된 KRIST 컬렉션에 대하여 기술하고자 한다. KRIST 컬렉션은 32개의 필드로 구성된 13,515개의 문서와 30개의 질의, 그리고 각 질의에 대한 적합 문헌 리스트로 구성되어 있다.

질의에 대한 적합 문헌 리스트의 생성은 테스트 컬렉션 개발에 있어서 가장 중요한 요소이다. 적합 문헌 리스트의 생성을 위한 가장 초보적인 방법은 각각의 질의에 대하여 테스트 컬렉션에 포함된 모든 문서들을 읽고 적합성 여부를 판단하는 것이다. 그러나 이 방법은 문서의 수가 많은 경우에 대단히 많은 시간을 요구한다. 본 연구에서는 적합 문헌 리스트의 생성을 위하여 다음과 같은 방법을 사용하였다. Cornell 대학에서 개발된 SMART 시스템(Salton & McGill 1983)을 한글 문서를 검색할 수 있도록 수정하고, 각각의 질의에 대하여 검색을 수행하여 적합 문헌 후보들을 선정하였다. 즉, SMART 시스템에 의해 높은 순위를 부여받은 문서들을 적합 문헌 후보들로 선정하였다. 그리고 후보로 선정된 문서들에 대해서만 적합성 여부를 판단하였다. 이러한 방법은 대용량 테스트 컬렉션의 개발을 목표로 하는 TREC(Text REtrieval Conference)(Harman 1993)에서 사용된 방법과 유사성을 지니고 있다.

본 논문의 구성은 다음과 같다. 2장에서 한글 문서를 검색할 수 있도록 수정된 SMART 시스템에 대해 설명한다. 수정된 시스템은 적합성 여부의 판단 대상이 되는 적합 문헌 후보들을 선정하기 위해 사용된다. 3장에서 문서와 질의의 형태에 대하여 기술한다. 4장에서 적합 문헌 리스트의 생성 과정을 자세히 설명하고, 마지막으로 5장에서 결론을 맺는다.

2. SMART 시스템을 이용한 한글 문서의 검색

SMART 시스템은 하버드와 코넬 대학에서 35년 간에 걸쳐 개발된 정보 검색 시스템이다. SMART 시스템은 벡터 공간 모델(Salton 1989)을 기반으로 하며, 문서와 질의 모두 다음과 같은 벡터 형태로 표현된다.

$$d_i = (w_{i1}, w_{i2}, \cdots, w_{in})$$

여기에서 d_i는 문서 또는 질의를 표현하고, w_{ik}는 문서 d_i에서 색인어 t_k의 가중치이다. 특정 문서에 나타나지 않는 색인어들에 대해 가중치 0이 할당된다. SMART 시스템은 영어 문서의 검색을 목적으로 개발되었다. 그러나 영어는 단어의 구분이 분명하여 색인 과정이 단순한데 비하여, 한글은 띄어 쓰기의 자유로움과 조사의 발달로 인하여 색인 과정에 어려움을 지니고 있다. 본 연구에서는 다음과 같은 색인 방법을 사용함으로써 한글 텍스트를 벡터 형태로 변환하였다(이준호 외 1995).

1. 빈칸, 마침표, 쉼표, 따옴표 등을 구분자로 하여 어절들을 추출한다.
2. 불용어 리스트를 이용하여 색인어로서 가치가 없는 불용어들을 제거한다. 한글에서는 단어에 다양한 종류의 조사나 어미 등이 붙을 수 있고, 복합어와 동사의 활용이 다양하므로 불용어 선정에 신중을 기해야 한다.
3. 나머지 어절들에 대해 최장 일치법을 이용하여 조사, 어미, 접미사 등의 결합으로 생성된 비색인 분절을 제거한다.
4. 색인 분절에 대해 bigram 방법을 적용한다. bigram이란 인접한 2개의 음절을 말한다. 예를 들면, '프로그래밍'이란 어절에 대해 bigram은 '프로', '로그', '그래', '래밍'이다. 색인 분절의 음절수가 2보다 큰 경우에는 색인 분절을 여러개의 bigram들로 분리하고, 작은 경우에는 색인 분절 전체를 하나의 bigram으로 취한다.
5. 각각의 bigram을 색인어로 선정하고 가중치를 부여한다.

문서 또는 질의에 대한 벡터들이 형성되면, 이후의 검색 과정은 벡터들의 연산에 의해 이루어진다. 문서 d가(w_{d1}, w_{d2}, ⋯, w_{dn})로 표현되고, 질의 q가(w_{q1}, w_{q2}, ⋯, w_{qn})로 표현되었을 때, 문서 d와 질의 q사이의 유사도를 의미하는 문서 d의 문서값은 다음과 같이 두 벡터들의 내적으로 계산된다.

$$Sim(d,q) = \sum_{i=1}^{n} w_{di} \times w_{qi}$$

문서값은 색인어들의 가중치에 의해 결정되기 때문에, 가중치 부여 기법은 검색 효과에 영향을 미치는 중요한 요소이다(Lee 1995; Salton & Buckley 1988). 본 연구에서는 색인어에 가중치를 부여하기 위하여 다음과 같은 공식들을 사용하였다.

$$(atc)\,w_{ik} = \frac{(0.5+0.5\,\frac{tf_{ik}}{\max tf}) \cdot \ln \frac{N}{n_k}}{\sum_{j=1}^{n} [(0.5+0.5\,\frac{tf_{ij}}{\max tf}) \cdot \ln \frac{N}{n_j}]^2}$$

$$(atc)\,w_{ik} = (0.5+0.5\,\frac{tf_{ik}}{\max tf}) \cdot \ln \frac{N}{n_k}$$

$$(atc)\,w_{ik} = \frac{tf_{ik} \cdot \ln \frac{N}{n_k}}{\sum_{j=1}^{n} (tf_{ij} \cdot \ln \frac{N}{n_j})^2}$$

3. 문서와 질의 형태

정보 검색용 테스트 컬렉션은 일반적으로 문서집합, 질의집합 그리고 각 질의에 대한 적합 문헌 리스트로 구성된다. 이들 중 검색의 대상이 되는 문서집합은 테스트 컬렉션 구축에 있어서 가장 기본적인 요소이다. 본 연구에서는 문서집합으로 연구개발정보센터 소유의 KRIST 데이타베이스를 사용하였다. KRIST 데이타베이스는 다음과 같은 과기처지원 연구보고서에 대한 서지 레코드 13,515건을 포함한다.

- 과학기술정책관리연구소주관 특정연구개발사업 (1982-1992): G7 프로젝트, 국책연구사업, 첨단요소기술개발사업, 중소기업지원

사업, 국제공동연구사업, 연구기획평가사업
- 한국과학재단지원 특정목적기초지원사업: 핵심전문연구(1978~1992), 특정기초연구(1986~1992)

다음은 KRIST 데이타베이스에 포함된 하나의 레코드를 보여주며, Table 1은 레코드 작성에 사용된 태그들의 의미를 보여준다.

〈DOCID〉 65
〈AB〉 ^a 한국대륙붕 석유자원 탐사의 성패는 정확한 시추위치 선정에 달려있음이 과거 10개년간에 걸친 탐사 결과 밝혀지고 있다. 성공적 시추의 관건은 막대한 량의 기존 탐사자료에 대한 재평가와 중요 구역에 대한 재정밀탐사에 달려 있음도 아울러 인식되고 있다. 이러한 일은 주요관련 기술의 자립화에 의거하는 길만이 첩경이 될 수 있기 때문에, 그 동안 축척 확보한 기술, 경험, 고급인력을 토대로 대륙붕석유자원 탐사기술을 자체개발하고 이를 실용화시키기로 하였다. 대륙붕 석유자원 탐사기술은 현장조사, 자료처리 및 해석, 시추, 평가에 이르기까지 일련의 상이한 각종 기술로 이루어지나, 이중 핵심이 되는 탄성파탐사자료 전사처리기술과 석유근원암 분석평가기술에 국한, 연구를 추진하였고 제1차년도 연구내용과 범위는 다음과 같다. 탄성파탐사자료 전산처리기술 연구는 이론정리, 전산처리 표준과정의 선정, 연구용 컴퓨터 및 소프트웨어의 선정 및 발주, 도입예정 컴퓨터 및 소프트웨어를 이용한 탄성파기록단면 작성, 소프트웨어의 시험적 자체개발의 단계로 수행하였다. 석유근원암 분석연구는 당연구소 보유 IFP Rock Eval를 이용, 한국대륙붕 시추암편 sample에 대한 분석을 실사하고 외국에서의 분석결과와 비교하므로써, 자체분석치에 대한 신뢰도를 검토하였다. 아울러 석유근원암 평가를 시험적으로 실시하였다.
〈AN〉 KRDC00000081
〈EQ〉 ^a 컴퓨터 CDC CYBER ^a ROCK EVAL
〈EW〉 ^a oil exploration over continental shelf ; ^a seismic stack section ; ^a source rock evaluation
〈ID〉 ^a 탐사개발연구실
〈IN〉 ^a IA14 ^b 한국자원연구소 ^c Korea Institute of Geology Mining and Materials ^d 042)868-3060 ^e 대전시 유성구 가정동 30번지
〈JP〉 ^a 한국동력자원연구소 ^b 책임연구원 ^c 구자학 ^d 물리탐사 ^e 박사; ^a 한국동력자원연구소 ^b 책임연구원 ^c 조동행 ^d 물리탐사 ^e 박사; ^a 한국동력자원연구소 ^b 책임연구원 ^c 곽영훈 ^d 지질학 ^e 박사; ^a 한국동력자원연구소 ^b 선임연구원 ^c 서상용 ^d 물리탐사 ^e 박사; ^a 한국동력자원연구소 ^b 선임연구원 ^c 이상규 ^d 물리탐사 ^e 박사
〈KW〉 ^a 대륙붕석유탐사 ; ^a 탄성파기록단면 작성 ; ^a 석유근원암 평가
〈NM〉 ^a 구자학 ^c 320325-1023511
〈RA〉 ^a 108268
〈RN〉 82-10-0106-00-00
〈RR〉 ^a 82.05.01-83.04.31

본 연구에서는 구축된 문서를 검색하기 위하여 30개의 자연어 질의를 작성하였다. 박사학위를 소지한 3명의 전문가들이 자신들의 연구 분야와 관련된 10개씩의 질의를 작성하였으며, 주제 분야는 생명과학, 의용전자공학, 기계공학이다. 다음은 본 연구에서 작성된 자연어 질의중에서 각 분야별로 하나씩을 보여주고 있다.

〈QID〉 1
〈TI〉 리포좀과 약물수송체
〈AB〉 약물을 체내에 수송하는 담체로써 리포좀을 이용하는 것과 관련된 문서들을 검색하고자 한다. 리포좀의 물리화학적 성질을 이용하여 최적으로 약물을 리포좀 내부로 포집시키는 효율성, 저장기간의 연장 등에 대한 것과 필요한 부위

로만 약물을 선택적으로 보낼 수 있도록 하는 설계문제, 체내에서의 안전성 및 안정성, 혈중농도, 수송하고 있는 약물의 약효 변화등에 대한 논의가 예상된다.
〈KW〉 리포좀, 약물수송체, 약물 타겟팅, 혈중농도
〈QID〉 11
〈TI〉 디지털 신호처리 칩을 이용한 칼만필터 알고리즘의 구현
〈AB〉 칼만필터 알고리즘을 하드웨어로 구현하는 연구와 관련된 문서를 검색하고자 한다. 특히 고속의 연산처리가 가능한 디지털 신호처리칩을 이용한 시스템의 구현에 관한 문서를 검색하고자 한다. 시뮬레이션은 관계가 없다.
〈KW〉 디지털 신호처리 칩, 칼만필터 알고리즘, 고정소숫점, 부동소숫점

〈QID〉 21
〈TI〉 일반 곡선 좌표계를 이용한 비압축성 유동 해석
〈AB〉 복잡한 계산영역에 있어서 일반곡선 좌표계를 이용한 비압축성 유동 해석에 관하여 기술하는 문서를 검색하고자 한다. 계산영역에 대한 격자 생성과 좌표변환을 통한 지배방정식의 변환이 이루어져야 하며, 운동방정식의 대류항에 대한 고정도 스킴에 관한 토론 및 레이놀즈수의 증가시 난류의 적절한 모델링이 필요하다. 해의 수렴을 가속시키기 위한 행렬해법의 벡터화와 고속화를 위한 프로그램의 벡터화 및 병열 처리가 논의될 수 있다.
〈KW〉 일반 곡선 좌표, 격자 생성, 계산 스킴, 난류, 벡터화, 행렬해법, 대류항 처리

4. 적합 문헌 리스트 생성

각각의 질의에 대한 적합 문헌 리스트의 생성은 테스트 컬렉션 구축에 있어서 가장 중요한 요소이다. 적합 문헌 리스트의 생성을 위한 가장 초보적인 방법은 각각의 질의에 대하여 테스트 컬렉션에 포함된 모든 문서들을 읽고 적합성 여부를 판단하는 것이다. 이 방법은 문서의 수가 많은 경우에 대단히 많은 시간을 요구한다. 한편, 다수의 검색 시스템을 사용하여 검색을 수행하고, 각각의 시스템에 의해 높은 순위를 부여받은 문서들에 대하여 적합성 여부를 판단하는 방법이 제안되었다(Harman 1983). 이 방법은 풀링 방법(pooling method)이라고 불리며, 테스트 컬렉션 구축시 적합 문헌 리스트 생성에 효과적인 방법으로 알려져 있다. 풀링 방법을 적용하기 위해서는 다수의 검색 시스템이 요구된다.

그러나 다수의 검색 시스템의 확보에 어려움이 있기 때문에, 본 연구에서는 한글문서를 검색할 수 있도록 수정된 SMART 시스템을 이용하여 다음과 같이 적합 문헌 리스트를 생성하였다.

① 국문연구과제명(TI), 국문초록(AB), 국문키워드(KW)에 대해 색인을 수행하여 문서와 질의 벡터를 생성한다. 문서와 질의 모두에 대하여 가중치 공식 atc를 사용하고, 상위 순위 100개의 문서를 검색한다.

② 검색된 100개의 문서에 대해 적합성 여부를 판단함으로써 1차 적합 문헌 리스트를 생성한다. 적합성 여부의 판단은 각각의 해당 질의를 작성한 전문가에 의해 수행되었다.

③ 1차 적합 문헌 리스트를 이용하여 다음과 같은 두번의 적합성 피드백 검색을 수행한다. 적합성 피드백 방법으로는 Rocchio(Salton & Buckley 1990)가 사용되었다. 첫째, 문서와 질의 모두에 대하여 가중치 공식 atc를 사용하여 상위 순위 100개의 문서를 검색한다. 둘째,

문서에 대해 가중치 공식 atn을 사용하고, 질의에 대해 가중치 공식 ntc를 사용하여, 상위 순위 100개의 문서를 검색한다.

④ 두번의 적합성 피드백 검색에 의해 검색된 문서들의 합집합에 대하여 적합성 여부를 판단함으로서 최종 적합 문헌 리스트를 생성한다.

1차 적합 문헌 리스트의 생성을 위해 질문당 100씩, 총 3,000개의 문서에 대해 적합성 여부가 판단되었고, 이들 중 250개의 문서가 적합 문헌으로 판명되었다. 최종 적합 문헌 리스트의 생성을 위해서는 총 3,715개의 문서에 대해 적합성 여부가 판단되었고, 이들중 314개의 문서가 적합 문헌으로 판명되었다.

<표 1> KRIST 레코드에서 태그의 의미

〈AN〉	제어번호
〈DT〉	자료유형
〈RN〉	과제관리번호
〈RP〉	총연구기간
〈RR〉	당해연구기간
〈SN〉	차수
〈TI〉	국문연구과제명
〈TO〉	영문연구과제명
〈NM〉	연구책임자국문명
〈ENM〉	연구책임자영문명
〈IN〉	국문소속기관명
〈EIN〉	영문속속기관명
〈ID〉	소속부서
〈JN〉	참여기업
〈PR〉	선행연구과제
〈IC〉	국제공동연구
〈CR〉	위탁연구
〈RA〉	정부출연금
〈RC〉	기업부담금
〈PJ〉	기수행국내연구
〈JP〉	참여연구원국문명
〈EJP〉	참여연구원영문명
〈EQ〉	사용기자재
〈AB〉	국문초록
〈EA〉	영문초록
〈PN〉	면수
W〉	국문키워드
〈EW〉	영문키워드
〈PD〉	발행일자
〈MN〉	발주기관관리번호
〈N1〉	기타1
〈N2〉	기타2

5. 결 론

정보 검색에 대한 연구에 있어서 테스트 컬렉션은 수행중인 연구 결과의 우수성을 입증하기 위한 필수적인 요소로 인식되어 왔다. 그러나 최근 한글 정보 검색에 대한 관심이 급속히 확산되었음에도 불구하고, 한글 정보 검색용 테스트컬렉션의 부족으로 인하여 한글 정보 검색에 대한 연구에 어려움을 겪고 있다. 본 연구에서는 연구개발정보센터 소유의 KRIST 데이타베이스를 기반으로 하여 개발된 KRIST 테스트 컬렉션에 대하여 기술하였다. KRIST 테스트 컬렉션은 13,515건의 서지 레코드와 30개의 자연어 질의 그리고 각 질의에 대한 적합 문서 리스트로 구성된다.

참 고 문 헌

김성혁 (1994). 자동색인기 성능시험을 위한

Test Set 개발. 정보관리학회지, 11(1), 82-101.

이준호 외 (1995). 한글 문서를 위한 효과적인 색인 방법. 제2회 한국정보관리학회 학술대회, 11-14.

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

Cleverdon, C.W., Mills, J. & Keen E.M. (1966). Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test Results. Aslib Cranfield Research Project, Cranfield, England, 1966.

Fox, E. (1983). 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.

Harman, D. (1993). Overview of the 1st text retrieval conference. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 36-48.

Lee, J.H(1995). Combining multiple evidence from different properties of weighting schemes. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrival, 180-188.

Salton, G. & McGill, M.J. (1983). Introduction to Modern Information Retrieval, McGraw-Hill, Inc.

Salton, G. & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513-523.

Salton, G. (1989). Automatic Text Processing the Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley Publishing Co., Reading, MA.

Salton, G. & Buckley, C.(199). Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4), 288-297.

Sparck Jones, K. & Webster, C.A. (1979). Research in Relevance Weighting. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge.