개선된 불리언 검색시스템 사례연구
Retrieval Effectiveness in Hybrid Boolean System

명순희 (사무자동화과)
Soon-Hee Myoung (Dept. of Office Automation)

Key Words: information retrieval, Boolean model, term weighting, relevance rankings, TARGET system, fuzzy set model, probabilistic retrieval

ABSTRACT: Conventional Boolean retrieval system has long been regarded as an efficient information retrieval model owing to the fact that it is relatively easy to implement and processing time is fairly short. However, the lack of a mechanism presenting the closeness to or similarity with the query of retrieved documents has been a major defect of the system. Various research outcomes to complement the Boolean search model while preserving its efficiency are summarized. Several searches were conducted on DIALOG using a hybrid Boolean feature for relevance retrieval, and the results are compared with those of plain Boolean search operation.

1. 서론

컴퓨터를 활용한 정보검색 기술은 1960년대 이래 급속히 발전되어 전통적으로 수작업 방식에 의존하던 정보검색 분야에 혁신을 가져왔다. DBMS(database management system)는 대부분 기관에서 대량의 수치정보나 고도로 정규화된 문자정보를 효율적으로 저장하고 검색하는데 매우 광범위하게 이용되어 왔으며, 최근에는 DBMS에서 다루는 정형화된 데이터와는 달리 비정형 데이터의 관리 및 활용에 대한 요구가 증가하면서 효율적인 정보 검색 시스템(information retrieval system; IRS)에 대한 관심 또한 커지고 있다. 신문제작, 출판의 전산화, 워드 프로세싱의 발달 등으로 인하여 종전과 같은 입력 및 저장 비용을 들이지 않고도 이용 가능한 텍스트 데이터의 양은 증가일로에 있으며, 인터넷의 대중화에 따라 웹문서나 멀티미디어 데이터에 대한 검색요구도 커지고 있다.

이러한 데이터를 본고에서는 문헌(document)으로 지칭하는데 이는 시스템 상에서 저장과 검색의 단위가 되는 데이터 객체이다. 정보검색 시스템의 목적은 이용자의 정보요구에 적합한 콘텐트를 담고 있는 데이터 객체를 추출하는 것이며, 사용자 질의와 대응되는 데이터 객체에는 전자출판물을 비롯하여 멀티미디어 데이터, 웹 문서 등, 전자화된 데이터가 광범위하게 포함되기 때문이다. 그러나 IRS는 텍스트 데이터베이스에 많이 이용되므로 편의상 검색 대상 데이터 객체 전반에 걸쳐 문헌이라고 쓰기도 한다. 텍스트 데이터의 경우에는 본문 또는 초록과 색인어 등, 본문의 대용물(surrogates)을 탐색대상으로 한다.

비정형정보의 양적인 증가뿐 아니라 정보검색 환경은 정보검색사와 같은 중간자에서 이용자 중심으로 바뀌어 정보검색 시스템은 직접조작을 통한 검색을 보다 용이하게 할 수 있는 사용자 인터페이스를 제공할 필요가 있다. 과거에 구축되어 현재까지 이용되는 대부분의 정보검색 시스템들이 새로운 검색 기술의 수용에 소극적인데 그 이유는 이미 기존의 시스템에 대량의 자료가 축적되어 있고 시스템의 재구축 비용이 상당하기 때문이다. 따라서 개선된 시스템은 기존의 불리언 모델의 장점을 최대한 유지하면서 검색 성능을 향상시킬 수 있도록 순위부여

모델과 불리언 모델을 통합하려는 연구가 시도되었다. 본 논문에서는 개선된 불리언 모델의 순위부여 방법에 관한 이론적 배경을 검토하고, 현재 DIALOG 시스템 상에 구현된 TARGET 프로그램을 이용, 적합성 검색(relevance retrieval)을 실시하여 기존의 불리언 방식과 비교하였다.

2. 불리언 검색모델의 문제점

정보검색 이론의 발전과 환경의 변화에도 불구하고 현재 운용되고 있는 대부분의 정보검색 시스템은 역화일(inverted file) 구조에 전통적인 불리언 검색 방식이 사용되고 있다. 불리언 검색방식은 데이터가 엄격히 정형화된 DBMS에 유용한 검색방식으로 대용량의 비정형 텍스트 데이터를 다루는 정보 검색 시스템에서 불리언 방식을 채택한 것은 적합치 않았다는 의견도 제기되었다.(1) 수치정보나 정형화된 문자정보를 저장.처리하는 DBMS에서의 결정적 검색과는 달리 IRS에서는 질의와 검색될 데이터의 형태가 매우 상이하여 매칭 프로세스가 훨씬 복잡하며 검색이 확률론적이라는 사실을 간과했다는 것이다. 그러나 초기 IRS 개발 당시로서는 화일 조작 기술에 한계가 있었으며, 보통 검색을 전담하는 중간자가 시스템과 이용자 사이에서 요구분석과 질의표현 등을 담당하였으므로 특별히 대안을 요하지 않았다는 반론과 함께 기존의 불리언 검색시스템의 개선 방안이 제기되었다.(2)

대량의 문헌집합을 검색하기 위한 상용서비스를 제공해온 DIALOG, MEDLARS 등, 대부분 IRS들은 최근까지도 대부분 집합이론에 기초한 불리언 논리검색 기법을 이용하고 있다. 이들 IRS의 문헌은 일반 데이터베이스에 비하여 훨씬 구조화되어 있지 못하며 문헌 내용에 대해 추론하기도 어렵다. 따라서 IRS에서 질의는 주제를 나타내는 문자열을 포함하는 연산으로 제한된다. 즉 ($Term_1$ OR $Term_2$) AND $Term_3$의 조건을 만족하는 문헌집합을 생성하는 정도이다. 이와 같이 불리언 검색 시스템에서 문서의 내용은 색인어로 대표되고, 불리언 질의는 검색어, 즉 키워드를 불리언 논리 연산자 AND, OR, NOT으로 연결하여 표현된다.

대부분 상용정보시스템들은 빠른 검색과 검색효율을 위해 사전화일, 색인화일과 문헌화일로 구성되는 역화일 구조를 이용한다. 검색 엔진은 시스템에 입력된 질의를 구성용어와 구성 연산자로 파싱하고, 각 질의용어는 색인화일에서 합치되는 색인어를 찾아 그 용어에 대응하는 문헌식별자의 집합을 검색한 다음, 문헌식별자 목록을 불리언 연산자에 따라 결합하여 문헌집합을 되돌려준다. 대부분 논리연산에 더하여 검색어 사이에 위치해도 허용되는 단어의 개수를 지정하는 인접연산과, 용어절단 기능(truncation) 등을 이용할 수 있다.(3)

질의에 포함된 색인어들의 불리언 논리연산을 통해 질의조건을 완전히 충족시키는 문헌만을 검색하게 되는 기능은 매우 단순하면서도 질의 조건에 따라서는 높은 검색성능을 보여주기도 하므로 널리 이용된다. 질의조건을 충족하는 문헌집합만이 제시되므로 이용자는 결과물에 대해서는 검색된 이유를 알 수 있다. 효율적인 질의 구성 능력에 따라서는 재현율과 정확도 면에서 좋은 성능을 기대할 수 있다. 또한 불리언 검색 모델은 앞서 지적한대로 구현하기가 비교적 용이하고 검색 속도가 빠르다는 점에서 대부분의 운영환경에서 이용되고 있다.

그러나 불리언 검색 시스템에 대해서는 다음과 같은 문제점이 지적되었다.

1. 불리언 연산의 해석이 너무 엄격하다. 하나의 문헌은 색인어 벡터 (t_1, t_2, t_3, ..., t_n)로 표현할 수 있는데 불리언 모델에서 t_i의 값은 용어의 유무에 따라 각각 1과 0의 값이 부여된다. 불리언 연산의 결과 참값인 1로 나타나는 문헌만 검색되는데 전체문헌을 검색되는 문헌과 나머지 문헌으로 엄격히 상호 배타적인 체계로 분리함으로써 부분 매칭이 되는 관련문헌은 모두 누락된다. 불리언 연산은 융통성 있게 해석하는 것이 검색효율을 향상시키는 것으로

나타났다.[4]

2. 검색결과로는 질의어와 색인어의 합치여부만 알 수 있을 뿐 질의내용과의 합치 정도에 관한 정보는 전무하다. 검색 용어의 중요도나, 질의와 문헌과의 유사성 등을 계산, 검색자에게 제시하는 기능이 전무하다.

3. 단순한 매칭 결과로 문헌 집합이 생성되고 이에 대한 적절한 여과장치가 없으므로 검색결과물이 과다하거나 과소하게 되는 점도 시스템의 효율성을 저해한다.

4. 이밖에도 전산이용환경의 변화도 들 수 있다. 검색자가 필요로 하는 자료를 검색하기 위해서는 논리연산자를 이용한 정교한 검색식을 구성하여 적용함으로써 검색 결과의 정확율과 재현율을 높일 수 있는데, 정보검색 시스템에 대한 사용자의 직접 탐색이 증가하는 환경에서 모든 이용자가 이와 같은 검색을 수행할 것으로 기대할 수는 없다. 정확율은 검색된 문헌집합에 대한 적합문헌의 비율이고, 재현율은 적합문헌에 대한 검색문헌의 비율로서 검색의 효율성을 나타낸다.

텍스트의 경우 인접 링크 등을 색인어와 병행 활용하는 등 다양한 방안이 연구되어 있다. 대부분의 연구는 가중치를 이용한 순위부여 방법을 이용하면 단순한 논리연산 결과로 주어진 문헌 집합에 비하여 정확율이 향상됨을 보여주고 있다.[5]

3. 개선된 불리언 모델

불리언 모델의 대안은 벡터모델, 퍼지 검색 및 확률 검색 모델 등의 순위 부여 모델들이다. 따라서 기존의 불리언시스템을 유지하면서 문제점을 보완하기 위한 연구는 대부분 불리언 모델에 이들 순위부여 방법을 결합시킨 것들이다. 경험적 접근방법의 하이브리드 불리언 모델과, 확률, 퍼지이론에 기초한 이론적 모델에 관해 간략히 고찰한다.

3.1 하이브리드 불리언 모델

전통적인 불리언 모델의 논리연산 기능과 계산의 효율성을 유지하는 한편, 통계적 용어가중치를 이용한 순위부여 기능을 부여하였다. 검색용어와 색인어의 합치여부에 따라 검색이 결정되는 점은 불리언 모델과 동일하나 합치되는 용어의 수에 따라, 또는 합치되는 용어의 가중치에 따라 적합도가 높은 문헌부터 이용자에게 제시되는 것이다. 문헌 D_i의 검색값(retrieval status value)은 다음과 같다.

$$RSV(D_i) = \sum_{j=0}^{q} w_{ij} \quad for\ i = 1, 2, \ldots, n \qquad (1)$$

w_{ij}는 질의용어와 합치되는 색인어의 문헌내 가중치이며, 질의용어 T_j에 대하여 j=1, 2, ..., q 동안 가중치를 합산한다.

용어가중치를 계산하기 위해서는 우선 색인어가 자동 추출되어야한다. 텍스트 또는 검색대상이 되는 데이터는 용어의 스테밍(stemming), 용어 절단, 불용어 목록 검사, 시소러스 참조 등의 텍스트 처리를 거쳐 색인어로 추출된다. 스테밍은 단어를 어미변화에 무관하게 검색되도록 공통 어근 형태로 축소시키는 절차이고, 시소러스 참조는 색인어를 동의어, 유사어 등 관련어들도 함께 확대검색하기 위해 제공되는 장치이다. 추출된 색인어는 단위문헌 내에

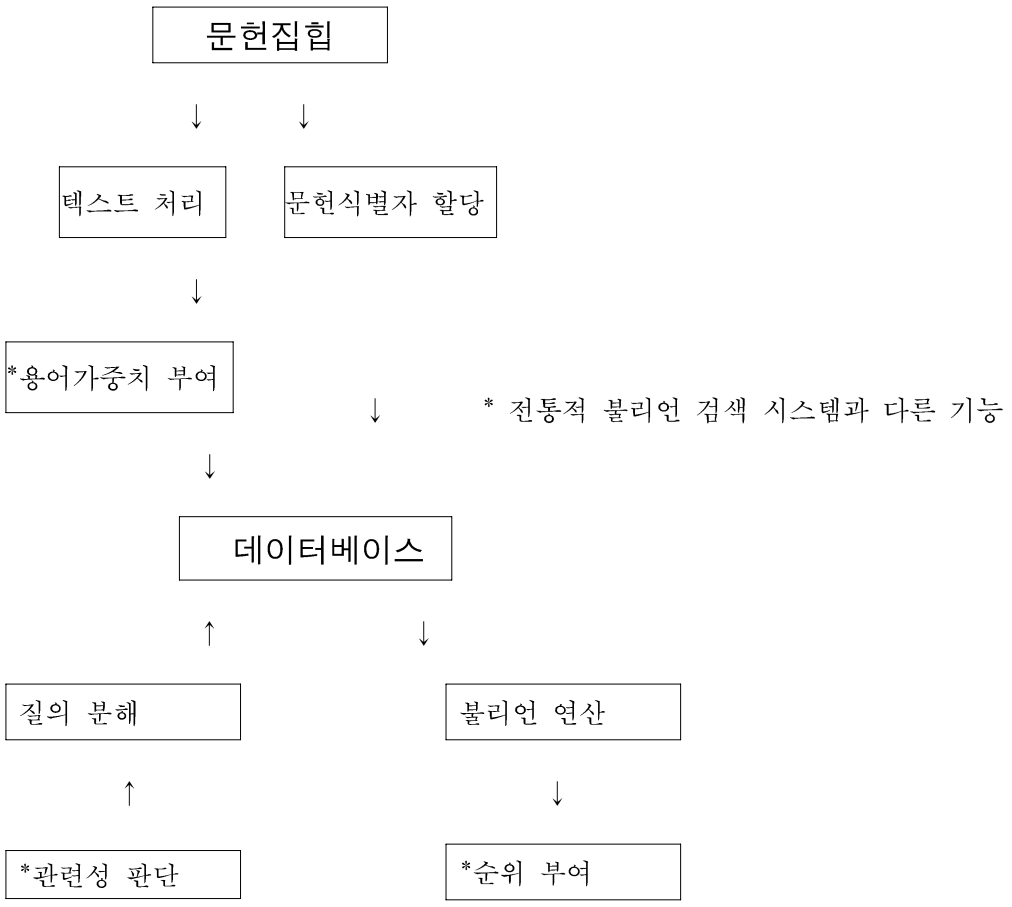
나타난 용어의 도수(tf; term frequency)와 문헌집합 전체에 나타난 도수(df; document frequency) 정보를 이용하여 용어의 중요도를 나타내는 가중치를 계산한다. 가중치는 Zipf의 법칙을 이용하여 전체 문헌집합에 널리 분포된 용어는 가중치를 줄이고 (inverse df) 특정문헌에 집중적으로 나타난 용어에 높은 가중치가 주어지는 것이다. 용어가중치 w_{ij}를 계산하는 표준적인 방법은 다음과 같다.[6]

$$w_{ij} = tf \cdot idf = tf_i \cdot \log\left[\frac{n}{df_i}\right] \quad (2)$$

여기에 문헌의 길이가 길수록 검색가능성이 높아지는 것을 방지하기 위하여 값을 0과 1사이에 두고자 식(1)을 다음과 같이 정규화 하였다.[7]

$$W_{ij} = tf_{ij} \cdot \frac{idf_i}{\sqrt{\sum_{k=0}^{t}(tf_{ik} \cdot idf_k)^2}} = \frac{W_{ij}}{\sqrt{\sum_{k=0}^{t} W_{ik}^2}} \quad (3)$$

하이브리드 모델에서는 위와 같은 용어가중치를 이용하므로 가중치 계산으로 인한 처리속도 증가로 인한 검색시간 부담을 줄이려면 용어가중치 정보를 역색인 파일에 함께 저장해야하는데 이 경우 색인화일의 크기가 늘어나는 것을 감수해야한다.



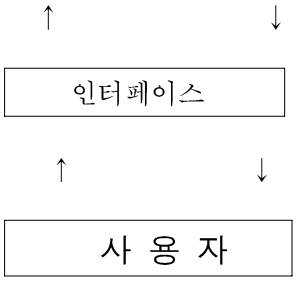


Fig.1 Hybrid Boolean model

보다 복잡한 유형의 하이브리드 불리언 모델은 SIRE 시스템과 같이 불리언 모델과 벡터모델의 특성을 결합한 것이다. 즉 벡터 공간모델을 이용하여 질의와 문헌간의 유사도를 계산하는 것이다. 벡터공간모델은 정보검색분야가 이진가중치에 근거한 단순 불리언 검색방법을 탈피하여 지능화하는데 많은 공헌을 하였다. 검색질의와 문헌간의 유사도는 질의 용어의 가중치를 요소로 하는 질의 벡터와 문헌의 내용을 대표하는 색인어의 가증치를 요소로 하는 문헌벡터가 문헌집합의 공간에서 위치한 거리로 표현된다.

$$\text{Cosine}(D_1 \cdot Q) = \sum_{i=0}^{t} tf_{ij} \cdot \frac{w_{qt}}{\sqrt{\sum_{j=0}^{t} tf_{ij}^2 \cdot \sum_{i=0}^{t} w_{qj}^2}} \qquad (4)$$

w_{qj} : 질의 용어 T_j의 가중치

이상과 같이 계산된 용어 가중치나 유사도에 따라 내림차순으로 정렬하여 문헌에 우선순위를 부여해서 디스플레이 하면, 검색자는 가장 연관성이 높은 문헌부터 순차적으로 브라우징하여 적합한 문헌을 선택할 수 있다.

하이브리드 불리언 시스템은 불리언 논리에 의해 질의용어와 대응되는 문헌집합만을 선택하고 이들에 대해서만 유사도를 측정하고, 용어 가중치나 유사도에 따른 순위 정렬이 이루어지므로 응답시간이 빠르다.

3.2 이론적 모델들

3.2.1 퍼지검색 모델

불리안 검색에서 색인어는 존재 유무에 따라 1과 0의 값이 각각 부여함으로서 검색된 문헌간의 차별화가 이루어지지 않는데, 퍼지검색에서는 색인어와 질의어에 그 중요도 또는 문헌주제에 대한 소속정도에 따라 1에서 0사이의 값을 부여한다. 이 값을 용어 가중치로 하여 불리언 논리 검색을 확장할 수 있다.

퍼지 모델에서 질의와 문헌간의 유사도는 다음과 같이 계산한다.

$sim(Q_{OR},D) = C_{OR1} \times min(d_{A1}, d_{A2}, \ldots d_{AN}) + C_{OR2} \times max(d_{A1},d_{A2}, \ldots, d_{AN})$

$sim(Q_{AND},D) = C_{AND1} \times min(d_{A1}, d_{A2}, \ldots d_{AN}) + C_{AND2} \times max(d_{A1},d_{A2},\ldots, d_{AN})$

C는 융통성 계수이고 d_{A1}은 문헌 D에 있는 색인어 A_1의 용어 가중치이다.

일반적으로 $C_{OR1} \rangle C_{OR2}$, $C_{AND1} \rangle C_{AND2}$ 이다.

퍼지집합 이론을 바탕으로 한 모델이 제시되기는 하였으나 min과 max에 근거한 유사도는 불리언 모델 유사하여 문헌의 차별화 성능이 약하다.(8)

3.2.2 확률모델

순위부여 모델로서 확률모델의 주안점은 질의에 대한 문헌의 적합성은 그 문헌이 사용자에게 적합한 확률로 표현되어야 한다는 것이며, 그 계산은 매칭 함수를 이용하여 문헌과 질의어의 적합성 확율의 크기대로 문헌을 순서화한다. 용어분포에 대한 정보를 이용하여 관련성의 확률을 검색된 각 문헌에 할당할 수가 있으며, 검색된 문헌들은 가능한 관련성의 순서대로 나열된다.

확률 검색모델에 구현되는 순위부여는 유용한 대안이며 확률모델에서의 질의 용어는 다음 계산식에 의해 계산된 가중치를 갖는다. 검색된 문헌은 문헌.질의간의 유사도에 따라 검색순위가 부여된다.(9)

$$w_i = \log \frac{p_i(1-u_i)}{u_i(1-p_i)} \qquad (5)$$

p_i = P(x_i =1 | relevant) :용어 i 를 포함하고 적합할 확률 (= r_i/R)

u_i = p(x_i =1 | nonrelevant):용어 i 를 포함하고 비적합할 확률(= (n_i-r_i)/(N-R))

따라서 질의와 문헌간의 유사도는 다음과 같이 계산된다.

$$SIMILAR(D,Q) = \sum_{i=1}^{t} d_i \cdot \log \frac{p_i(1-u_i)}{u_i(1-p_i)} + C \qquad (6)$$

3.2.3 P-norm 모델

'확장된 불리언 모델'로도 불리는 P-norm 모델에서는 색인어의 가중치뿐 아니라 질의용어에 대한 가중치도 사용 가능하다. P-norm모델에서는 용어 가중치d_{A1}, d_{A2}, ..., d_{An}, 을 갖고 있는 문헌 D를 n차원 공간좌표에 있는 것처럼 처리한다. 일반적인 질의는 $\{(A,w_a)\ AND^p\ (B,w_b)\}$의 형태이며 w_a, w_b 는 질의 용어 A, B의 가중치이다. 상관계수 P는 불리언 연산에 대한 제한정도를 나타낸다. n차원의 벡터 D는 식(7)에서 와 같이 P의 값이 1에 가까울수록 AND, OR 등 불리언 연산자의 성격은 모호해져서 벡터공간모델과 유사해지고, ∞에 가까울수록 불리언 연산자는 엄격히 해석, 적용된다.(10)

$$|D| = |(w_1, w_2, \ldots w_n)| = \left(\sum_{j=1}^{n} w_j^p \right)^{1/p} \qquad (7)$$

1〈=p〈=∞

w_1, w_2, ...: 벡터 d의 요소

많은 실험결과 최상의 p의 값은 2〈=p〈=5인 것으로 나타났다. P-norm모델이 효율적이라는 것이 많은 실험을 통해 인정되었으나 지수계산을 하기 때문에 계산비용이 크다는 단점이 있다.(11)

4. 개선된 불리언 시스템 사례

DIALOG 시스템의 TARGET은 Mead社의 FREESTYLE과 함께 1993년에 발표되어 주목을 받으며 1994년부터 실용화되었다. 이보다 한해 앞선 1992년에는 소개된 West社의 WIN('Westlaw is natural'의 약어)은 적합성 검색과 아울러 자연어 질의 입력, 시소러스 조회, 법률용어 인식 등의 기능을 갖춘 검색엔진으로 LEXIS-NEXIS의 검색에 이용되었다. 이 세가지 검색 기술 개발의 공통된 목표는 기존의 불리언 모델이 문헌의 적합성에 관한 정보를 이용자에게 전혀 제공하지 못하는 단점을 보완하기 위하여 문헌과 질의사이의 적합성, 또는 유사도를 검색 결과에 반영하는 기능을 추가하고자 하는 것이었다.[12] 이들의 구현에는 앞서 고찰한 통계, 확률, 연관 검색 기법 등, 이론적 모델이 적용되었다.

DIALOG의 TARGET은 통계 모델을 채택한 것으로서 하이브리드 불리안 시스템이다. TARGET은 입력된 질의용어를 갖고 있는 모든 문헌을 추출하고, 용어의 적합성을 계산한다. 계산은 식 (2)에 따라 역문헌빈도(idf)를 문헌의 길이로 정규화한 것을 사용한다. 검색결과는 50개까지 디스플레이되는데 여기에는 완전매치뿐 아니라 부분매치되는 문헌도 다수 포함된다. 따라서 텍스트 검색에 가장 적합하다.

TARGET을 이용한 검색화면은 다음 Fig.2과 같다.

```
?target

Input search terms separated by spaces (e.g., DOG CAT FOOD). You can
enhance your TARGET search with the following options:
   - PHRASES are enclosed in single quotes
        (e.g., 'DOG FOOD')
   - SYNONYMS are enclosed in parentheses
        (e.g., (DOG CANINE))
   - SPELLING variations are indicated with a ?
        (e.g., DOG?  to search DOG, DOGS)
   - Terms that MUST be present are flagged with an asterisk
        (e.g., DOG *FOOD)
Q = QUIT   H = HELP

?'community college?' computer?

Your TARGET search request will retrieve up to 50 of the statistically most
relevant records.
Searching ALL records
...Processing Complete
     Your search retrieved 50 records.
Press ENTER to browse results  C = Customize display  Q = QUIT  H = HELP
```

Fig.2 Screeen display of TARGET menu mode

기본적인 디스플레이 포맷은 Fig.3와 같이 제목, 출처, 출판연도만이 포함되어 있으나 디스플레이에 사용자 정의를 택하면 Fig.4과 같이 각 레코드에서 용어의 원빈도수 등 적합도를 계산하는데 사용된 통계치를 볼 수 있다.

DIALOG-TARGET RESULTS (arranged by percent RELEVANCE)
------- Item: 1 ----------
DIALOG(R)File 201:(c) Format only 1998 The Dialog Corporation, plc. All rts. reserv.

Post Secondary Programs for the Deaf: IV. Empirical Data Analysis. Research Report No. 75.
------- Item: 2 ----------
DIALOG(R)File 201:(c) Format only 1998 The Dialog Corporation, plc. All rts. reserv.

Follow-Up Study: Early Intervention Program.
------- Item: 3 ----------
DIALOG(R)File 201:(c) Format only 1998 The Dialog Corporation, plc. All rts. reserv.

Manoa's Community College Transfers, Fall 1970-Fall 1974.

Fig.3 Results of TARGET search

Press ENTER to continue browsing or enter item number(s) to see full record
M = Modify search T = New TARGET C = Customize display Q = QUIT H = HELP
?c
BROWSE output includes: TI,JN,PD
Term frequency/relevance: off
Continuous display for COMPLETE TEXT: off
Custom display options:
1 Change BROWSE output to Title Only (usually free)
2 Customize BROWSE output with your own choice of display codes
3 Reset BROWSE output to the default (i.e., title, journal, date)
4 Change COMPLETE TEXT output to continuous display
5 Show term frequencies and statistical relevance (%) for each item
Press ENTER for NO CHANGE, or enter option number(s)
(e.g., 1,5) to customize the display. Q = QUIT H = HELP
?5

DIALOG-TARGET RESULTS (arranged by percent RELEVANCE)
------- Item: 1 ----------
DIALOG(R)File 213:(c) 1989 Institution of Electrical Engineers. All rts. reserv.
Title: The merger of man and machine (*computer *aided *design)
Journal: Office Equipment News
- Statistical Relevance: 99%
- Term Frequency: COMPUTER AIDED DESIGN - 8 ; CAD - 7

------- Item: 2 ---------------------------------
DIALOG(R)File 213:(c) 1989 Institution of Electrical Engineers. All rts. reserv.

  Title: The portability of *computer -*aided *design applications on the Graphics Kernel System graphics standard
  Journal: International Journal of Computer Applications in Technology
 - Statistical Relevance: 99%
 - Term Frequency: COMPUTER AIDED DESIGN - 8 ; CAD - 5
------- Item: 3 ---------------------------------
DIALOG(R)File 213:(c) 1989 Institution of Electrical Engineers. All rts. reserv.

  Title: *Computer -*aided *design system requirements for surface mount technology design
  Book Title: Advancing surface mount technology. An IFS executive briefing
 - Statistical Relevance: 99%
 - Term Frequency: COMPUTER AIDED DESIGN - 8 ; CAD - 5

Fig.4 Custom display options

이에 비하여 재래의 불리언 검색 결과는 Fig.5에서와 같이 적합도나 용어의 빈도수와 같은 정보는 제공하지 않으며, 단순한 용어 매칭의 결과로 최신성에 따라 정렬되어 나타난다.

ss community(w)college? and computer?
     S1   3650 COMMUNITY (A SOCIAL GROUP LINKED BY COMMON INTERESTS TH...)
     S2   5436 COLLEGE?
     S3   969 COMMUNITY(W)COLLEGE?
     S4   1361 COMPUTER?
     S5   40 COMMUNITY(W)COLLEGE? AND COMPUTER?
?t s5/3/all

 5/3/1
DIALOG(R)File 201:ONTAP(R) ERIC
(c) Format only 1998 The Dialog Corporation, plc. All rts. reserv.

EJ117996 JC501000
  They Do Come Back Another View of Student Attrition
  Lightfield, E. Timothy
  Community College Frontiers; 3; 3; 45-49Spr 1975

 5/3/2
DIALOG(R)File 201:ONTAP(R) ERIC
(c) Format only 1998 The Dialog Corporation, plc. All rts. reserv.
EJ115406 CE502729
  Training Computer Technicians
  Tontsch, John
  School Shop; 34; 8; 50-2Apr 1975

5/3/3
DIALOG(R)File 201:ONTAP(R) ERIC
(c) Format only 1998 The Dialog Corporation, plc. All rts. reserv.

EJ114646 JC500969
Using Technology to Serve Learning Needs of the Community
Bolvin, Boyd M.
New Directions for Community Colleges; 3; 1; 33-38Spr 1975

Fig.5 Results of the conventional Boolean search operation

재래의 불리언 검색과 적합성 검색방법을 비교하기 위하여 다음과 같이 두 개의 질의를 구성하여 각각의 방식으로 검색, 다운로드한 다음 두 그룹의 정확도, 두 그룹에 동시에 나타나는 문헌의 수를 조사하였다.

1. 검색 File: ERIC(ONTAP)
   TARGET search: 'community college?' *computer?
   Boolean search: ss community(w)college? and computer?
2. 검색 File: INSPEC(ONTAP)
   TARGET search: 'computer aided design' CAD
   Boolean search: ss computer()aided()design OR cad

불리언 탐색 결과 1번은 40건, 2번은 634건이 검색되었는데, 이들을 TARGET 검색결과 얻어진 각 50개의 문헌과 비교하였다.(Fig. 6) 정확도는 검색된 문헌 가운데 정보요구에 적합한 문헌의 비율인데 IRS에서의 적합성이란 문헌의 건텐트를 검토하고 내리는 주관적인 판단일 수 밖에 없다. 본 연구에서는 메타 데이타를 검토하여 중복되는 문헌 수가 적을수록 두 검색 시스템간의 효율성 편차는 큰 것으로 판단하였다. 용어의 매칭 결과로 얻어진 Boolean 검색과 적합성 순위에 따라 검색된 50개의 문헌 가운데 1번 검색에서는 적합도 80% 이상인 문헌과 중복되는 문헌은 10개로 나타났고, 2번의 검색에서는 중복문헌이 전혀 나타나지 않았다. 규모가 제한된 테스트 파일에서 중복 문헌, 즉 불리언 탐색 결과에서 적합문헌으로 판단되는 부분이 0%~25% 정도로 나타나는 것은 데이터베이스의 규모가 클수록 불리언 검색의 정확율이 저하될 수 있음을 뜻한다. 또한 메타 데이터로 판단할 때 TARGET 검색과 불리언 검색의 효율성의 편차는 상당히 클 것으로 보인다.

Fig.6 Comparison of Boolean and TARGET searches

	(1) 불리언 검색 결과	(2) 적합율 80% 이상 TARGET 결과	(1)과 (2)의 중복 건수	불리언 검색결과 상위 50건 정확율
ERIC	40 건	16 건	10 건	25%
INSPEC	634 건	30 건	0 건	0%

TARGET 검색의 경우 매우 정규화된 검색어를 사용하며, 질의 구조에 익숙해야하는 등,

기존의 불리언 검색방식을 이용하는에 필요한 숙련도가 요구되는 등 큰 차이가 없으나 연산자를 직접 입력하여 정교한 검색식을 세우지 않는 것은 이용자의 입장에서는 진일보한 기능이다.

5. 결론

불리언 모델의 대안으로 벡터모델 등 순위부여모델들이 많은 연구자에 의해 제시되었는데, 개선된 불리언 모델은 대부분 불리언 모델과 순위부여모델의 결합을 통하여 개선책을 제시하고 있다. 이들이 공통적으로 목표하는 것은 질의와 문헌과의 연관성을 나타내는 정보와 이에 따른 문헌의 우선순위를 보여주고자 하는 것이다. P-norm 모델 등은 질의용어에도 가중치를 부여해서 중요한 용어와 문헌연관성을 높이고자 하였다.

DIALOG의 하이브리드 불리언 검색기능인 TARGET 명령어를 이용, 적합성 검색을 실시한 결과 정확률 면에서, 그리고 이용자 편의를 위한 기능면에서 향상된 성능을 보였다. 재현율의 경우는 기본적으로 50개의 검색결과물을 보여주므로 전통적 불리언 모델에서와 같은 결과물의 과다, 과소 문제는 발생할 소지가 없다.

참고문헌

(1) Cooper, W.S., 1988, "Getting beyond Boole", *Information Processing & Management*, Vol.25, pp. 243 ~ 248.

(2) Bookstein, Abraham,1985, "Probability and fuzzy-set applications to information retrieval", *Annual Review of Information Science and Technology(ARIST)*, Martha E. Williams, ed. Vol. 20, pp.117 ~ 151.

(3) Frakes, William B. and Baeza-Yates, R., 1992, 정보검색, 류근호, 김진호 공역. 원제: *Information Retrieval: Data Structures & Algorithms.* Englewood Cliffs, N. J., Prentice-Hall, Inc. pp. 509 ~ 582.

(4) Salton, Gerard, Fox, Edward A., Wu, Harry, 1983, "Extended Boolean information retrieval", *Communications of the ACM,* Vol.26, No.12, pp. 1022 ~1036.

(5) Savoy, Jacques, 1997, "Ranking schemes in hybrid Boolean systems: a new approach", *Journal of the American Society for Infromation Science*, Vol. 48, No. 3, pp. 235 ~ 253.

(6) Salton, G. and McGill, Michael J., 1983, *Introduction to Modern Information Retrieval.* New York, McGraw-Hill Book Co. pp. 52 ~ 117.

(7) op. cit., pp. 199 ~ 255.

(8) ibid, Savoy, 1997, pp. 239.

(9) Croft, W. B., Harper, D. I., 1979, "Using probabilitistic models of document retrieval without relevance information", *Journal of Documentation,* Vol. 35, pp. 285 ~ 295.

(10) ibid., Salton, Fox, Wu, 1983, p. 1025.

(11) ibid., Savoy, 1997, pp. 241.

(12) Tenopir, Carol, and Cahn, Pamela, 1994, "TARGET & FREESTYLE: DIALOG and Mead join the relevance ranks", *Online*, V. 18, No. 3, pp. 31 ~ 47.