완전매치와 부분매치검색기법에 관한 연구

(A Study of an Exact Match and a Partial Match as an Information Retrieval Technique)

김 영 귀

□ 목 차 □

초 록

본 연구는 그동안 연구되고 개발된 여러 검색기법을 검색된 문헌집합의 특성과 사용된 표현에 의해서 완전매치검색과 부분매치검색으로 구분하였다. 완전매치는 부울논리가 그 대표적이며 현행 대부분의 정보검색시스템에서 사용하고 있는 검색기법이다. 부분매치는 부울논리가 가지고 있는 문제점과 한계점을 극복하기 위한 대안으로서 많은 연구가 있었으나 그 본질은 부울논리 구조안에서 검색을 향상시킨다는 점에서 한계를 가질 수 밖에 없다 하겠다. 대표적인 예로 확률검색, 벡터공간모델, 그리고 퍼지집합을 대상으로 두 검색기법을 비교하고 앞으로의 검색기법이 나아가야 할 방향을 제시하였다.

ABSTRACT

A retrieval technique was defined as a technique for comparing the document representations. So this study classified retrieval technique in terms of the charactristics of the retrieved set of documents and the representations that are used. The distinction is whether the set of retrieved documents contains only documents whose representations are an exact match with the query, or a partial match with query. For a partial match, the set of retrieved document will include also those that are an exact match with the query.

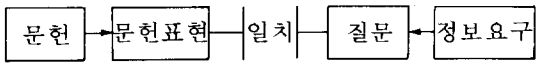
Boolean-logic as one of the exact match retrieval techniques is in current in most of the large operational information retrieval systems despite of its problems and limitations.

Partial match as an alternative technique has also various problems. Existing information retrieval systems are successful in assisting the user whose needs are well-defined(e.g. Boolean-logic), to retrieve relevant documents but it should be successful in providing retrieval assistance to the browser whose information requirements is ill-defined.

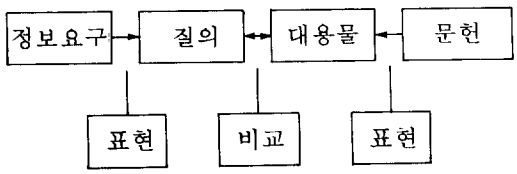
* 부산여자대학 강사

Ⅰ. 서 론

정보검색이라 함은 크게 네가지로 정보분석, 정보축적, 탐색수행, 검색처리로 구분되며, 검색은 질의에 속하는 용어집합과 문헌에 속하는 용어집합을 각각 비교함으로써 실행된다. 따라서 검색기법은 질의어표현과 본문표현을 비교하는 문제를 말한다. 그 과정은 그림으로 나타내면 다음과 같다.[1)2)]



〈그림 1〉 질문과 문헌의 대조(고전적정보검색)[1)]

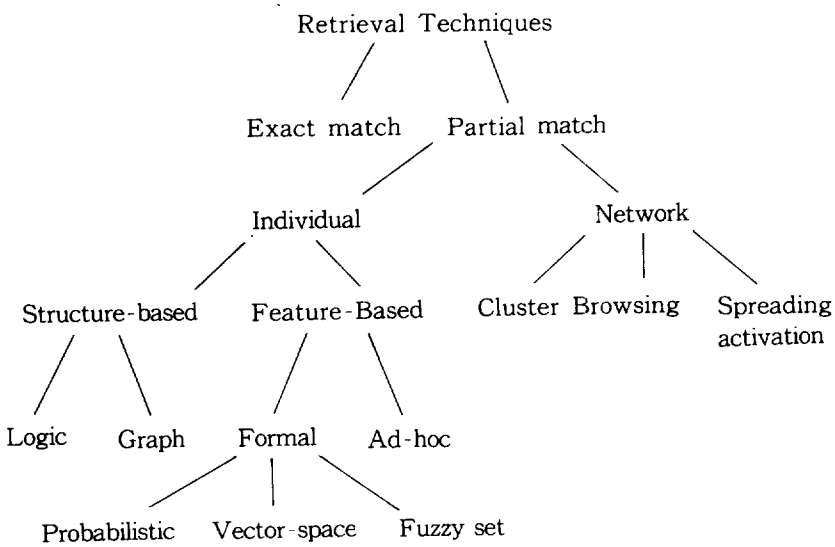


〈그림 2〉 검색기법처리[2)]

따라서 검색기법은 이용자질의에 대한 응답으로, 또는 이용자의 질의어표현과 시스템의 본문표현을 비교하여 적합한 문헌을 검색해 내는 것이며 표현이 다르면 기법도 달라지게 된다. 그러나 궁극적으로는 검색기법은 이용자의 정보요구를 만족시켜 주어야 하며 그러한 만족도를 높이기 위하여 여러 검색기법이 적용되어 왔으며, 또 연구 개발되고 있다. 따라서 본 연구는 현행정보검색시스템에서 사용하고 있는 검색기법은 어떤 것이 있으며, 그 장점과 문제점은 무엇이며, 그 문제점에 대한 대안기법은 어떤 것이 있으며 앞으로의 정보검색 시스템이 나아가야 할 방향을 제시하고자 한다.

Ⅱ. 현행정보검색기법의 개관

검색기법을 질의표현과 문헌표현을 비교하는 기법으로 정의하였을 때 검색된 문헌집합의 특성과 사용된 표현에 의해 검색기법을 분류할 수 있다.



〈그림 3〉 검색기법의 분류[3)]

첫번째 구분은 검색된 문헌들의 집합이 질문과 정확히 일치되느냐 아니면 부분적으로 일치되느냐에 따라 완전매치(exact match)와 부분매치(partial match)로 분류하며 쇼발(Shoval)은 완전매치('Perfect' match)와 부분매치('Fuzzy' match)로 표현하기도 한다.[4]

완전매치는 이용자의 요구모델이 문헌표현내에게 정확하게 질문으로 공식화되는 것을 요구하는 기법으로 부울탐색, 전문(Full-text) 탐색 또는 스트링(string) 탐색 등으로 수행되었으며, 현재 대부분의 대규모정보검색시스템에서 사용되고 있는 기법이다. 이런 형태의 기법의 단점을 여러 문헌에 잘 알려져 있으며 적당한 수행을 위해서는 시소러스같은 다양한 보조도구가 필요하다.

간단한 검색의 경우에 완전매치탐색은 질문과 부분적으로 일치하는 적합한 문헌을 놓치고, 검색된 문헌에 순위를 매기지 않으며, 질의나 문헌내의 개념의 상대적인 중요성(Relative importance)을 고려하지 않으며 복잡한 질의논리공식을 요구한다[5]는 점이다. 따라서 완전매치검색의 논리에서 필요한 중요한 노력은 그릇을 덜 일치하게(less exact) 만들고, 상대적인 중요성을 고려하고, 사리에 맞게 순위를 매기는규칙(Ranking rules)을 만드는 일일 것이다.

여러가지 단점에도 불구하고 완전매치탐색이 운용시스템의 패러다임으로 남을 수 있는 것은 현행시스템에 대한 투자가 너무나 크기 때문에 이것을 변경한다는 것은 경제적으로 타당성이 없고, 대안적인 기법들이 대규모환경에서 실험되지 않았으며, 또 대안적인 기법들의 결과가 실험적인 환경에서조차도 어떤 변경을 정당화할 만큼 충분히 나은 점이 없었기 때문이다.[6]

부분매치는 다시 개별문헌대표들과 질의를 비교하는 기법들(Individual, 본 연구에서는 Formal 기법만 다루었다)과 네트웍에서 다른 문헌들과의 연결을 강조하는 문헌표현을 이용하는 기법들(Network)로 구분된다.

Ⅲ. 완전매치 검색기법

1. 부울논리의 특징과 장점

1950년대는 컴퓨터화 된 정보탐색의 가능성에 대해 신중한 사고를 했던 시기로 문헌디스크립터의 부울논리결합으로 탐색요청을 공식화할 수 있다고 제안되었고, 이 제안은 대부분의 수학자, 컴퓨터과학자, 그리고 기술적으로 정보전문가들도 인정하였다. 10년 후 첫 대형서지검색서비스가 설립되었을 때 부울논리 탐색 접근법은 기초검색전략으로 채택되었다.[7]

현행 온라인정보검색시스템에서 가장 많이 채택

1) Martha J. Bates, "An exploratory paradigm for online information retrieral," In: B.C.Bookes, ed. *Intelligent information system for the information society: Proceedings of the 6th International Research Forum in Information Science(IRFIS, 6)* Frascati, Italy, Sept. 16-18, 1985: New York: North-Holland, 1986, p.91.
2) Nicholas J. Belkin, W. Bruce Croft, "Retrieval techniques," In: Martha E. Williams, ed. *Annual Review of Information Science and Technology(ARIST)*, Vol.22, New York: Knowledge Industry Pub., 1987, p.110.
3) *Ibid.*, P.112.
4) Peretz Shoral, "Principles, procedures and rules in an expert system for information retrieval," *Information Processing & Management*, Vol.21, No. 6(1985), p.476.
5) Nicholas J. Belkin, W. Brure Croft, *op. cit.*, p.113.
6) Nichelas J. Belkin, W. Bruce Croft, *op. cit.*, p.114.
7) William S. Cooper, "Getting beyond Boole," *Information Processing & Management*, Vol.24, No.4 (1988), p.243.

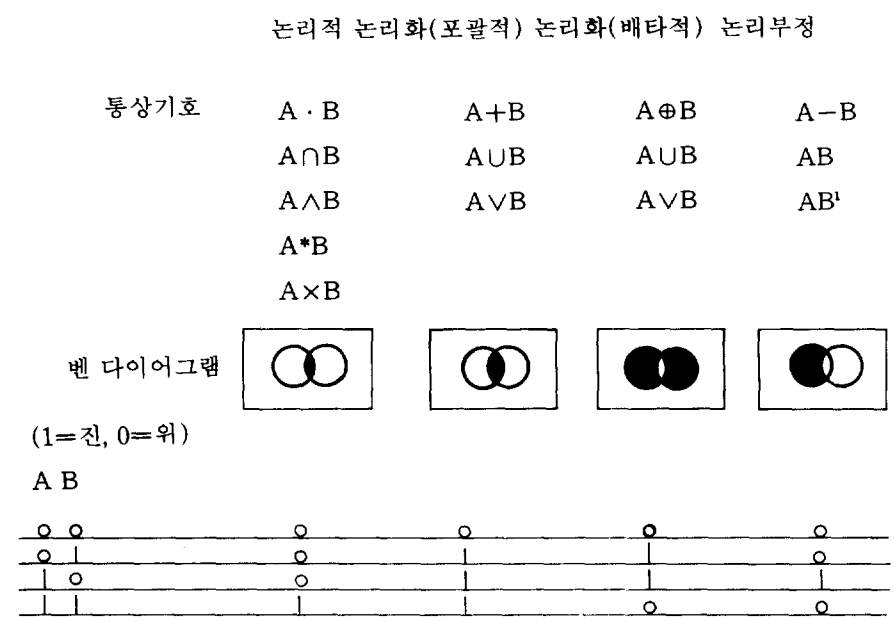
하고 있는 검색기법은 부울논리에 의한 검색이며[8] 검색활동은 부울탐색문과 도치파일 시스템을 사용한다.[9] 부울논리탐색기법을 사용한 검색의 결과는 두개의 뚜렷한 문헌집합을 만들어 낸다. 즉 정해진 질의에 대한 해답으로 검색된 문헌집합과 검색되지 않은 문헌집합이다.[9] 따라서 개별문헌사이의 관계나 유사도 그리고 키워드나 질의용어 사이의 관계도 활용되지 않는다.

부울논리연산의 일반표현형식은 그림 4와 같다.

어쨌든 부울검색언어는 색인에 의해서 기술될 수 있는 어떤 문헌집합은 항상 적절하게 구축된 부울탐색문으로 상술될 수 있다는(적어도 원칙으로는) 의미에서 "표현이 완전한(Expressively complete)"[11] 장점을 가지고 있다. 또 부울시스템이 인기가 있는 이유는[12]

1) 대부분의 이용자들이 부울논리에 익숙해 있고,

2) 부울시스템의 단순성과 자연스러움이다.



〈그림 4〉 부울논리연산의 일반표현형식[10]

8) G. Salton, "Some characteristics of future information systems," *SiGiR Forum*, Vol.18(1985), p.28.

9) H.S. Heaps, *Information retrieval: computational and theoretical aspects*, New York: Academic Press, 1978, p.107.

9) G. Salton, Michael J. McGill, *Introduction to modern information retrieval*, New York: McGraw-Hill, 1983, p.200.

10) Charles H. Davis and James E. Rush, 平山健三, 田淵利明 譯, 情報檢索の原理と實際, 東京: 丸善株式會社, 1977, p.76.

11) William S. Cooper, "Exploiting the maximum entropy principle to increase retrieval effectiveness," *Jouronal of the American Society for Information Science*, Vol.34, No.1(1983), pp.31-32.

12) A. Bookstein, "A comparison of two systems of weighted Boolean retrieval," *Journal of the American Society for Information Science*, Vol.32, No. 4(1981), p.275.

3) 이용자가 자신의 언어로 정보요구를 표현할 때 일반적으로 요구유형의 공식화로 이용자와 시스템 양쪽이 다같이 배우기 쉽다.

4) 부울논리에 의한 요구는 처리하기 쉽다.

5) 부울검색의 논리때문에 체계적인 방법으로 임의의 요구를 편리한 형태로 변형하는 것이 가능하여 같은 문헌의 집합을 검색한다.

2. 부울논리의 문제점

부울시스템에 내재하고 있는 가설은[13] 이용자가 자신의 정보요구를 완전히 그리고 정확하게 시스템의 색인과 양립할 수 있는 논리식으로 맞할 수 있을 것이라는 점이다. 따라서 정보요구에 관련된개념들 사이의 논리적인 관계를 정확하게 표현하고, 질의를 다른 더 편리한 형태로 변형시킬 수 있는 능력을 가지고 있다고 생각한다. 이용자의 관점에서는[14] 입력요건이 매우 간단하여 질의를 한번만 부울형태로 표현하면 되거나 시스템에서 질의를 부울 형태로 변경할 수 있다.

이러한 강점들이 부울검색의 결점들을 간과하게 만든다. 가장 심각한 약점은[15] 고객이나 색인자가 문헌이 어떤 주제에 관해 쓸 수 있는 다양한 수준의 질의에서나 문헌의 색인레코드에 관련된 주제의 "상대적인 중요성"의 표현을 허용하지 않는다는 점이다.

여러 연구논문에서 지적된 부울탐색의 단점 및 문제점들을 조사해 보면 다음과 같다.

살톤(G.Saton)은 도치파일조직을 가진 부울탐색은 도치파일시스템이 가지는 수많은 단점과 제한점을 가지고 있기 때문에 정해진 검색수행을 하기 전에 고려해야 할 사항으로[16]

1) 부울탐색공식은 이용자의 요구에 따라 검색한 자료의 과다-탐색깊이를 다양하게 하는 것이 어렵다.

2) 부울탐색환경에서 문헌과 탐색요구사이의 계속적인 접근의 스펙트럼을 제공하는 벡터매칭함수의 효과를 얻기는 어렵다.

3) 이런 "부분매치" 능력의 결여는 불리안탐색 출력의 질(質)이 "전부가 아니면 전무(All or nothing)" 가 되게 한다. 복잡한 질문식의 요구는 평균이용자를 만들기 어렵게 한다.

4) 출력문헌의 집합을 질의-문헌 유사도의 내림차순으로 우선 순위가 매겨져 이용자에게 제출될 수 없다고 지적한다.

5) 질의와 문헌에 부여된 색인정보는 도치파일 조직에서는 쉽게 변경될 수 없다.

힐드레드(R.Hildreth)는 ① 탐색질문공식의 이용이 초보탐색자에겐 자연스럽지 못하다. ② 그것은 시스템자원면과 탐색자가 부적합한 검색의 조사에 보낸 시간면에서 비싸다. ③ 부울탐색은 정확한 탐색이 아니며, 잘못된 용어의 결합이 보통이다. ④ 키워드 / 부울탐색만으로는 MARC 레코드의 서명과 주제분야의 어휘에 대한 재현을 향상시킬 수 없을 것이다. 왜냐하면 같은 또는 관련주제표목에 해당하는 문헌이 빠질 수 있기 때문이다[16] 라고 비판하였다.

---

13) A. Bookstein, "On the perils of merging Boolean and Weighted retrieval systems," *Journal of the American Society for Information Science*, Vol.29, No. 3(1978), p.156.

14) A. Bookstein, "Fuzzy requests: an approach to weighted Boolean searches," *Journal of the American Society for Information Science*, Vol.31, No.4.(1980), p.246.

15) A. Bookstein,"A comparison of two systems of weighted Boolean retrieval," *op. cit.*, p.275.

16) G. Salton, *Dynamic information and library processing*, Englewood Cliffs: Prentice-Hall, 1975, pp.120-121.

16) R. Hildreth, "To Boolean or not to Boolean?" *Information Technology & Libraries*, Vol.2(1983), pp. 235-237.

쿠퍼(William S.Cooper)는 부울탐색문의 과다출력과 출력제로, 출력순위의 결여외에 부울언어의 표현력 제한과 초보자에게 주는 혼란을 강조한다. 특히 부울기호 OR와 AND이다. A AND B는 A혼자보다 더 많은 문헌을 검색하는 것으로 기대하고 그것은 OR에 대해서도 마찬가지이다. 더 일반적으로 대부분의 훈련받은 사람들은 벤다이어그램(Venn diagram)에 의해서 복잡한 부울수식을 일관성있게 그리고 정확하게 해석할 수 있게 되기란 거의 불가능하다[17]고 까지 얘기한다.

라데키(T.Radecki)는 현행 검색시스템들이 부울구조에 묻혀있고, 시스템들의 기초가 되는 이론적인 모델은 문헌과 이용자 정보요구는 색인용어집합과 부울탐색요구공식으로 각각 정확히 그리고 완전히특징지울 수 있다는 불분명한 가설에 근거하고 있다. 그러나 이 가설에서와 같이 문헌검색과정에 있어서 불확실성이란 아주 본질적인 것이므로 그 정확성이 결여되어 있다. 검색결정의 본래의 오류성은 효과적으로 처리하기 위한 표준부울모델의 무능력[18]이 오늘날 운용. 검색시스템이 보여주는 수많은 심각한 결함들의 주요 이유가 된다고 지적한다.

또 쿠퍼(william Cooper)는 문제점과 함께 해결방안으로 ① 부울공식의 불친절함(해결방안: 자유-심볼파세트-질문) ② 과다출력과 무출력(해결방안: 조합수준에 의한 순위부여) ③ 획일적인 파세트(해결방안: 질의용어에 가중치를 부여) ④ 가중치를 해석하는 것(해결방안: 확률적인 해석) ⑤ 용어의존성(해결방안: 향상된 통계기법)[19]을 제시하고 있다.

이상으로 여러 연구자들의 부울검색에 관해 제기된 문제점들을 요약해 보면 다음과 같다.

1) 탐색어로 표현되는 각 개념의 상대적인 중요성을 나타내지 못한다.

2) 과다출력과 무출력으로 인하여 부적합한 문헌의 검색 또는 적합한 문헌의 누락이 있게 된다.

3) 이용자가 중요하다고 생각되는 순서대로 출력되는 순위가 없다.

4) 문헌과 질문의 유사도의 크기순으로 검색문헌을 출력할 수 없다.

5) 탐색초보자들에게 부울언어는 혼란을 준다.

6) 탐색문과 완전히 일치되는 문헌만 검색이 되므로 부분적으로 일치되는 문헌을 검색할 수 없다.

7) 전통적인 부울처리기술은 가중치가 부여된 용어의 사용을 허용하지 않는다.

8) 검색된 모든 항목은 이용자의 정보요구를 만족시키는데 동등하게 유용성이 있다고 인식하여 수많은 문헌을 임의의 순서대로 제공하고 그 적합성의 범위에 대한 표시는 제공하지 않는다.

이러한 부울논리의 단점을 극복하기 위해 제시된 방안으로는 가중치와 부울논리를 결합시킨 검색기법이 있다. 이 검색의 목적은 크게 두 가지이다.[20] 첫째, 특정한 색인어의 가중치나 또는 가중치의 합이 일정치 이상인 문헌을 검색함으로써 질문과 보다 관련성이 높은 문헌을 검색하여 결과적으로 정확률을 높이는 것이다. 둘째, 검색문헌을

17) William S. Cooper, "Exploiting the maximum, entropy principle to increase retrieval effectiveness," *op. cit.*, pp.32-33.

18) T. Radecki, "Trends in research on information retrieval-the potential for improvements in conventional Boolean retrieval systems," *Information Processing & Management*, Vol.24, No.3(1988), pp.219-227.

19) William Cooper, "Getting beyond Boole," *Information Processing & Management*, Vol.24, No.3(1988), pp.243-247.

20) 정영미, 정보검색론, 서울: 정음사, 1987, p.263.

훑어보아야 한 이용자의 노력을 감소시킨다. 또한 적합성이 높은 문헌들을 먼저 출력함으로써 적합성판정에 의한 피드백을 이용하여 질문을 용이하게 수정할 수 있다. 또 훈련받지 않은 이용자가 효과적인 부울탐색을 만들기 어렵기 때문에 질문자가 제시한 자연어진술문에서 자동으로 부울탐색공식을 가능케하는 연구도 있다.[21][22]

Ⅳ. 부분매치검색기법

1. 확률검색

확률검색은 일반적인 확률접근법을 문헌검색에 적용하는 것으로, 그 개념은[23] 특정한 질문에 대한 각 문헌이 적합한 확률(probability of Relevance)과 부적합할 확률(probability of non-Relevance)을 산출하여 적합확률이 부적합활률보다 큰 문헌을 검색하는 것이다.

정보검색이론의 출발점은 탐색질의에 응답하여 생산된 시스템출력의 항목이 대체로 탐색자에게 유용한 문헌의 우선순위가 매겨져야 한다는 인식에서 나온 것이 확률순위원칙(probability Ranking principle,PRP)[24]이다.

확률순위원칙이란 만약 각 요구에 대한 참조검색시스템의 응답이 요구를 한 이용자에게 유용성이 있는 확률의 내림차순으로된 문헌순위라면, 확률은 가능한한 정확하게 이러한 목적달성을 위해 계산되며,[25] 이용자에 대한 시스템의 전반적인 효과는 그 데이터를 기초로하여 최상의 결과를 얻을 수 있을 것이다.

기본확률모델은[26] 각 문헌 x가 다음과 같이 용어벡터형태로 표현되었으면 xi는 i번째 용어의 유무를 나타내는 것으로 $xi=1$ 또는 $xi=0$의 값을 갖게 된다.

$$X=(x_1, x_2, \cdots\cdots, x_N)$$

이때 문헌 X는 특정한 질문에 대해 적합하거나 또는 부적합한 두가지 경우가운데 하나에 해당된다고 보고 각 경우를 W_1(=적합한 경우)과 W_2(=부적합한 경우)로 나타내면 이 문헌이 적합할 확률은 $P(W_1/x)$, 부적합할 확률은 $P(W_2/x)$가 된다. 따라서 $P(W_1/x) > P(W_2/x)$가 되는 문헌이 검색되는 것이다.

확률모델에 근거한 검색기법은[27] 벡터공간 모델에서 개발된 기법과 매우 비슷하며 기본목적은 질의에 대해 적합확률의 순서대로 문헌을 검색하는 것이다. 만약 문헌용어가중치가 1이거나 0이고 용어들이 서로 독립적이라 한다면 다음과 같이 문헌에 순위를 매길 수 있다.

$$\Sigma d_i q_i$$

qi는 $\log Pri(1-p_nr_i) / Pnri(1-Pr_i)$와 같은 가중치로 Pr_i는 용어i가 적합한 문헌집단에서 발생하는 확률이고, P_nr_i는 용어i가 부적합한 문헌집단에서 발생하는 확률이다. 이런 순위함수를 적용하는 문제는 질의용어가중치에서 확률을 평가하는 것이다.

---

21) G.Salton, C. Buckley and E.A. Fox, "Automatic query formulations in information retrieval," *Journal of the American Society for Information Science*, Vol.34, No.4(1983), pp.262-280.

22) 신성철, "불리언논리 탐색문의 자동생성을 위한 전문가 중개시스팀," 미간행석사학위논문, 경북대학교 대학원, 1988.

23) C.J. Van Rijsbergen, *Information Retrieval*, 2nd ed. London: Butterworths, 1979, pp.111-120.

24) William S. Cooper, "Getting beyond Boole," *Information Processing & Management*, op. cit., p.246.

25) S.E. Robertson, "The Probability ranking principle in IR," *Journal of Documentation*, Vol.33, No. 4(1977), p.295.

26) C.J. Van Rijsbergen, *op. cit.*, pp.115-117.

27) Nicholas J. Relkin, W. Bruce Croft, *op. cit.*, pp. 117-118.

확률모델은 문헌내빈도정보를 이용하기 위해서 확대될 수 있다.[28] 이 경우 위의 공식 $\Sigma d_i q_i$에서 사용된 용어가중치는 ts · idf이다. ts는 특정문헌에 대한 용어의 중요도를 측정하는 용어중요도(term significance) 가중치이다. ts가중치는 표준문헌내 빈도로서 최고로 평가되었다. 이런 형태의 순위함수는 실질적으로 벡터공간모델에서 개발된 것과 같다.

초기의 확률기법은 원래 문헌과 이용자질의 모두가 색인용어의 집합으로 간단히 표현된 탐색환경에 사용되도록 설계되었다. 전통적인 부울검색시스템을 향상하기 위해서 제안된 확률방법론은 기반이 되는 이런 시스템의 근본원리를 바꾸지 않고 쉽게 수행될 수 있다는 것이다.[29] 그래서 확대된 PRP-기반검색기법 역시 순수한 부울탐색의 실제적인 향상을 위해 생존가능한 기반을 제공하는 잠재력을 가지고 있다 하겠다.

전통적인 부울시스템에 본래의 확률검색기법을 확대하려는 이론적인 근거는 잘 알려진 부울대수의 결과에 근거하고 있으며 그러한 대수의 각 요소는 이접적표준형(disjunctive normal form, DNF)로 언급된 특수한 형태로 표현될 수 있다고 말한다.[30]

확대PRP-기반모델의 주요 특징들은 다음과 같다.[31]

1) 상업검색시스템에서 사용된 전통적인 부울탐색문공식을 이런 시스템들이 근거로 하고 있는 근본적인 원리를 변경하지 않고 조정하므로 비용면에서 고려할 가치가 있다.

2) 확률의 내림차순으로 문헌검색을 제공한다. 검색된 문헌집합의 크기에 대한 조정을 준비함으로써 순수 부울검색에서의 출력의 과다가 경감되거나 완전히 극복될 수 있을 것이다.

3) 정보요구를 상술하는데 있어서 광범위한 이용자 능력을 허용한다.

4) 시스템-이용자 적응장치를 적합성피드백정보를 이용하여 용이하게 한다. 질의용어에 대한 적합성가중치를 결정하고 제안된 적합성피드백 전략은 전통적인 도치파일기술과 양립할 수 있다.

현재의 정보검색의 확률이론의 확대는 모든 부울탐색공식은 발전된 이접표준형과 앞서 언급한 이접표준형 두가지로 표현할 수 있다는 사실에 근거하고 있다.[32] 예를 들면 정보검색시스템에 그 색인용가 T_1, T_2, T_3, T_4 네 용어로만 구성되어 있다고 정하면, 문헌집단{ d_1, d_2, ⋯, d_8}의 문헌표현 d_i{i=1, 2, ⋯, 8}로 다음과 같다.

$d_1=\{T_2, T_3, T_4\}$ $d_2=\{T_2, T_4\}$

$d_3=\{T_3, T_4\}$ $d_4=\{T_1, T_2, T_4\}$

$d_5=\{T_3, T_4\}$ $d_6=\{T_2, T_3, T_4\}$

$d_7=\{T_1, T_2, T_3\}$ $d_8=\{T_1, T_3, T_4\}$

이 시스템에 제시된 질의 q를 나타내는 부울탐색공식 q는 다음과 같다.

$q=NOT\ T_1\ AND(T_2\ OR\ T_4)$

q의 발전된 이접표준형은 다음과 같이 표현될 수 있다.

$q=(NOT\ T_1\ AND\ NOT\ T_2\ AND\ NOT\ T_3\ AND T_4)$
$OR\ (NOT\ T_1\ AND\ NOT\ T_2\ AND\ T_3\ AND\ T_4)$
$OR\ (NOT\ T_1\ AND\ T_2\ AND\ NOT\ T_3\ AND\ NOT T_4)$

---

28) *Ibid.*, p.118.

29) T. Radecki, "Trends in research on information retrieval-the potential for improvement in conventional Boolean retrieval systems," *op. cit.*, p. 222.

30) *Ibid.*, p.223.

31) *Ibid.*, pp.223-224.

32) T. Radecki, "A probabilistic approach to information retrieval in systems with Boolean search request formulations," *Journal of the American Society for Information Science*, Vol.32, No.6(1982), p.366.

OR (NOT T_1 AND T_2 AND NOT T_3 AND T_4)
OR (NOT T_1 AND T_2 AND T_3 AND NOT T_4)
OR (NOT T_1 AND T_2 AND T_3 AND T_4)

발전된 이접표준형이 아닌 이접표쥐형으로 표현하면 다음과 같다.

q=(NOT T_1 AND NOT T_2 AND T_4)
OR (NOT T_1AND T_2 AND NOT T_4)
OR (NOT T_1 AND T_2 AND T_4)

이 간단한 예의 결과, 정보검색에 확대통계접근법에 통합된 순위장치는 이접표준형이 발전된 이접표준형보다 덜 효과적이라는 것을 나타내지만 전자는 훨씬 계산이 적으므로 더 실용적이라 할 수 있다.

부울검색질의를 확률검색모델에 통합하는 다른 연구는 이접의 결합인 접속표준형(conjunctive normal form, CNF)에 부울데이터베이스질의를 배치하는 것[33]이 있다.

부울탐색에 기반한 현행운용검색시스템은 출력문헌순위에 가중치를 통합함으로써 급진적으로 향상될 수 있다. 그러나 이 노선에 따라 전통적인 부울검색시스템을 개선하기 위한 수많은 이전의 시도들이 성공하지 못한 이유는[34] 시스템 본래의 모순과 애매성 때문이다. 따라서 탐색 공식뿐 아니라 문헌표현이 단순히 색인용어의 집합인 검색시스템에 적용될 수 있는 것은 잘 알려진 확률출력순위방법을 확대하는 것으로 최근의 실험들은 이런 방법들을 검증하여 적합성피드백의 체계적인 통계이용의 가치를 증명했다. 따라서 전통적인 부울검색시스템에 확대확률문헌순위방법론의 적용이 유용한 것이라고 기대하는 것이다.

이러한 확률검색은 이론적으로는 매력있는 검색모델이지만 실제상에 있어서 적합한 정보의 획득문제로 인해 사용상의 어려움이 뒤따른다. 이러한 본질적인 문제외에 확률검색의 단점으로는[35] ① 확률이론이 처음 제시되었을 때 불리안관계를 고려하지 못했다. ② 축적정보량의 증가와 더불어 계산해야 할 자료의 양도 늘어난다. ③ 일반적인 접근법에 대해 합일점은 있으나 최선의 모델과 기술적인 의문점을 해결할 방법에는 의견일치가 없다. ④ 수학에서 억지이론을 뒤따르는 논증이 요구되고 이론의 접근을 제한하게 된다는 점등이다.

2. 벡터공간모델

전통적인 부울탐색의 대안으로서 어떤 부울연산자없이 간단한 탐색용어의 집합으로 구성되는 "벡터" 질의를 사용하는 것이 가능하다.[36] 그 경우에 혼성유사도측정은 문헌들에 부가된 용어집합과 질의용어집합 사이에 계산되어 검색된 문헌들은 질의-용어유사도의 내림차순으로 이용자에게 제공될 수 있다. 게다가 새롭게 향상된 벡터질의들은 이용자에게 적합하다고 알려진 사전에 검색된 문헌에 포함된 새로운 용어들을 질의문에 통합하는 초기 탐색활동에 따라 쉽게 구축될 수 있다.

검색기법의 관점에서 이 모델이 운용시스템에 어떻게 적용되어야 하느냐에 대한 권고사항은

33) Robert M. Losee, A.Bookstein, "Intergrating Boolean queries in conjunctive normal form with probabilistic retrieval models," *Information Processing & Management*, Vol.24, No.3(1988), pp.315-321.

34) T. Radecki, "Probabilistic methods for ranking output documents in conventional retrieval systems," *Information Processing & Management*, Vol. 24, No.3(1988), pp.281-302.

35) A. Bookstein, "Probability and Fuzzy-set application to information retrieval," In: Martha E.Williams, ed. *Annual Review of Information Science and Technology*, Vol.20, New York: Knowledge Industry Pub., 1985, pp.127-128.

36) G. Salton, E.A. Fox, E.Voorhees, "Advanced feedback methods in information retrieval," *Journal of the American Society for Information Science*, Vol.36, No.3(1985), p.200.

다음과 같다.[37]

1) 용어가 중치는 표준화된 문헌내빈도(within-document frequency, tf)와 역문헌빈도(inverse document frequency, idf)의 결합을 이용해서 계산된다. 이 tf · idf 가중치는 검색과정의 부분으로서 또는 덜 정확하게 문헌이 색인되었을 때 문헌용어를 위해 계산될 수 있다.

2) 형편없는분리가(discrimination value)를 가진 용어들(문헌들을 구별하는데 유용하지 못한 용어)은 낮은-빈도용어를 위한 시소러스류와 높은-빈도용어를 우한 句(phrase)를 나타내는 용어들로 대체되었다.

3) 문헌들은 코싸인상관계수(벡터공간에서 질의에 가장 가까운 문헌을 직관적으로 검색하는)가 측정한 것처럼 질의에 대한 유사도의 내림차순으로 순위가 매겨졌다.

$$\Sigma\, d_i q_i \,/\, \sqrt{\Sigma\, d_i^2 q_i^2}$$

(d_i는 tf · idf가중치이다)

벡터공간모델에서 순위는 질의-문헌유사도의 내림차순으로, 순위의 장점은[38] ① 이용자는 검색된 항목의 수를 쉽게 통제할 수 있다. ② 이용자는 훈련없이 광범위하고 쉽게 질의를 구축하여 좋은 결과를 산출할 수 있으므로 적합한 항목을 검색할 수 있다. ③ 나머지 검색된 문헌을 무시하는 선택권을 이용자에게 제공하는 것이 항목을 전연 검색하지 않는 것보다 더 안전하다 등이다.

이 모델은 객관적인 호소력을 가지고 있고 스마트(SMART) 시스템을 비롯하여 상당부분 정보검색연구의 기초를 이루었고 벡터공간모델에 기초한 연구는 적합성피드백, 클라스터링, 그리고 문헌공간수정같은 다른 기법들을 유도했다. 스마트시스템은 아마도 표준도치파일기술에 기반하지 않는 가장 잘 알려진 상업시스템이다. 이 스마트시스템은 그 자체가 다른 전통적인 검색시스템과 구별되며 그 차이점은 다음과 같다.[39]

1) 문헌과 탐색문에 내용식별자를 지정하기 위해서 완전히 자동색인방법을 사용한다.

2) 공통주제류의 관련문헌을 수집해서 특수주제영역에 있는 특정한 항목으로 시작할 수 있게 하고 이웃 주제분야에서 관련항목을찾을 수 있도록 한다.

3) 축적된 항목과 들어오는 질의사이에 유사도 계산을 수행함으로써 검색된 문헌을 확인하고 검색된 항목을 질문과 유사도의 내림차순으로 우선순위를 부여함으로써 확인한다.

4) 초기검색운용의 결과로서 획득된 정보에 기반한 향상된 탐색문을 만들기 위해 자동절차를 포함한다.

지난 10년동안 스마트와 사이어(Sire) 시스템에 관한 연구는 부울검색의 효과를 향상시키는데 새로운 기법들을 개척했다.[40] 확대부울논리, 자동부울질의구축, 그리고 부울피드백은 스마트시스템으로 다양한 실험을 함에 따라 중요한 발전을 가져왔다. 사이어시스템은 확대부울질의를 위한 P-표준도표(p-norm scheme)를 포함함으로써 향상되었다.

정확히 Lp백터표준(Lp Vector norm)에 기반한 일반화된 거리함수(distance function)는[41] 부울질의와 내용용어의 집합이 확인한 문헌과 비교하는데 사용되었다. 이 거리함수는 변수 P를 포함하고

37) N.J. Belkin, W.B. Croft, "retrieval techniques" *op. cit.*, p.116.

38) E.A. Fox and M.B. Koll, "Practical enhanced Boolean retrieral: experiences with the SMART and SIRE systems," *Information Processing & Management*, Vol.24, No.3(1988), p.260.

39) G. Salton, M. McGill, *op. cit.*, pp.118-120.

40) E.A. Fox and M.B. Koll, *op. cit.*, p.257.

41) G. Salton, "Some characteristics of future information systems," *SiGiR FORUM*, Vol.18(1985), p.36.

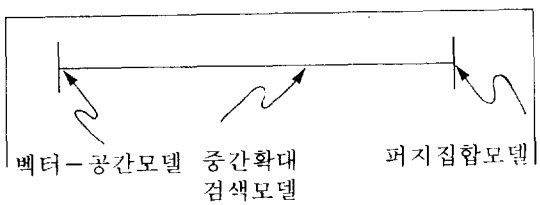
P는 1에서 무한대까지 다양하게 사용할 수 있다 (P, 1≦P≦∞).

1) P=∞일 때 부울 연산자는 전통적인 부울논리에서 처럼 엄격히 취급되었다.

2) P가 무한대에서 내려감에 따라 연산자 AND와 OR은 훨씬 덜 엄격해진다. 그래서 P값이 10에 해당하는 AND 연산자는 문헌에서 모든(All) 질의용어들 보다 대부분(most)의 질의용어가 있는 것이 더 나을 것이다; P-값이 10에 해당하는 OR-연산자는 하나(one)의 질의용어보다 몇개의(Some) 질의용어가 있는 것이 더 나을 것이다.

3) P가 더 낮은 1의 한정값에 이르면 AND와 OR연산자는 그들 사이의 함수가 완전히 없어지는 정도까지 완화되고 질의(A AND B)와 (A OR B)는 벡터질의 (A, B)로 간단히 처리되었다.

P값의 변화를 그림 5에서 알 수 있다.[42]〈그림 5〉



〈그림 5〉 P-값변화의 범위[42]

P-표준방법은[43] 특히 자동피드백과 연결되었을 때 분명 잇점이 있으며, 부울질의는 부울피드백을, P-표준검색은 P-표준피드백을 사용하여 향상될 수 있다. P-표준질의는 질의가 가지고 있는 그 표현력 때문에 표준질의 형태의 하나로서 코더시스템에서 채택되었다. 코더(COmposite Documentation Expert / Extended / Effective Retrieval, CODER) 시스템은[44] 정보검색 시스템의 효과를 증진시키기 위해서 인공지능방법의 적용을 연구하는 실험대이다. 초보자는 P-표준질의구축에 어려움을 느낄수 있으므로 피드백데이터가 있던 없던간에 자동질의공식화를 위한 방법들이 개발되었다.

그러나 벡터공간모형에 근거한 검색기법의 가장 중요한 확대는 확대부울 검색이다. 확대부울 검색시스템은 벡터처리시스템의 장점을 부울처리환경에 적용하여 벡터시스템의 이용자편의성과 부울처리시스템의 장점의 결합을 가능케 한다. 더 신축성있는 확대 부울처리시스템은 부울연산자를 전통시스템에서 보다 덜 엄격하게 처리할 수 있다.

이러한 배경을 가지고 있는 확대부울 검색시스템에서의 기본적인 처리단계는 아래와 같다.[46]

1단계 : 잠재적인 적합한 문헌의 상당부분을 검색하기 충분할만큼 넓은 전통적인 부울검색질의를 구성한다.

2단계 : 전통적인 도치파일시스템을 사용하는 이용가능한 문헌집단에 대해서 이런 질의를 처리하고 그것으로 집단속의 문헌들의 부분집합을 검색한다.

3단계 : 낮은 P-값(1≦P≦3)을 가지는 확대부울시스템에서 완화된 질의를 구축하기 위해 이용자 요구의 최초의 자연어진술문을 쓰거나 1단계에

42) G. Salton, E.A. Fox and H. Wu, "Extended Boolean information retrieval," *Communication of ACM*, Vol.26, No.11(1983), p.1026.

43) E.A. Fox and M.B. Koll, "Practical enhanced Boolean retrieval: experiences with the SMART and SiRE system," *op. cit.*, p.261.

44) E.A. Fox, "Development of the CODER System: a testbed for Artificial Intelligence methods in information retrieval," *Information Processing & Management*, Vol.23, No.4(1987), pp.357-358.

45) G. Salton, E.A. Fox, E. Voorhees, *op. cit.*, p.200.

46) G. Salton, "Some characteristics of future information systems," *op. cit.*, p.37.

서 얻은 최초의 전통적인 부울질의를 사용한다.

4단계 : 2단계에서 얻은 집단에 대해서 3단계에서 얻은 확대부울질의를 처리하고 질의-문헌유사도의 내림차순으로 문헌에 순위를 매긴다.

5단계 : 4단계에서 검색된 항목을 조사하고 향상된 확대부울질의를 구축하기 위해서 적합하다고 확인된 검색된 항목들에 포함된 용어들을 이용한다; 4단계에 되돌아가서 이용자가 출력에 만족할 때까지 반복한다.

이러한 처리단계를 거치는 벡터공간모형에 근거한 확대부울검색시스템의 장점으로는 ① 전통적인 부울검색에서 사용된 표준질의 구조를 수용하나 매치하는 용어의 수와 가중치에 의존하는 질의-문헌유사도를 허락함으로써 부조리한 해석을 피한다. ② 용어와 문헌 둘다를 각각의 구문에서 추정된 중요도에 따라 용어가중치의 혼성을 허용한다. ③ 질의-문헌유사도의 내림차순으로 문헌의 검색을 제공하기 때문에 검색된 문헌집합의 크기에 대한 통제를 제공한다. ④ 강제적인 구와 엄격한 동의어 해석 그리고 임시의 구와 엉성한 동의어관계 사이의 차이를 용이하게 한다. 질의구조를 특징화하는데 있어 전자는 높은 P-값을, 후자는 낮은 P-값을 사용함으로써 가능하다.

이러한 확대부울 검색시스템의 장점을 기반으로 하여 적용된 시스템은 앞서 언급한 코더시스템외에 포시스(Foster Care Expert System, FOCES)[47]가 있다. 포시스는 사회복지기관에서 사회사업가들이 고아에게 입양할 가정을 선정해 주는 전문가시스템으로 벡터와 확대부울매치를 사용하고 있다. 이 시스템은 입양을 필요로 하는 어린이들을 위해 입양가정을 선정하는데 관련된 사회사업가들을 도와줄 한 원형(Prototype)으로서 이다. 이 프로젝트는 인공지능분야의 전문가대체시스템의 활용을 증명하는 것으로 정보검색에 인공지능방법을 이용하는 것이다. 포시스의 설계는 게스(General purpose Expert System Shell, GUESS)를 사용하고 확대부울 매칭과 벡터의 상호관계를 위해서 프로로그루틴(Prolog routine)을 필요로 했다. 게스는 벡터와 질의처리에 대한 자원이 결핍되었으므로 게스표현을 벡터형태로 변환하고 뒤따르는 분류는 순위단계를 수행하기 위해서 프로로그코드(Prolog Code)가 사용되었다.

입양가정과 어린이 기록은 비슷한 방식으로 사전에 처리되고 어린이 기록들은 벡터와 색인과 분류처리의 결과로서 분류표를 생기게 한다. 입양가정과 어린이들 사이의 유사도는 그들의 분류표를 상호관련시켜서 계산되었다. 그래서 정해진 어린이들에 대한 모든 후보가정들이 어린이와 입양가정분류표의 유사도의 내림차순으로 순위가 매겨질 수 있다.

게스데이터구조의 벡터표현은 다음과 같다.[48]

어린이: $V^C x=(Cmf_1, \cdots\cdots, Cmf_m)$

사례 x번의 벡터이다. Cmfi는 어린이개념 m의 i번째 어린이 개념에 대한 실수-값을 가진 구성원함수값이다.

입양가정: $V^H y=(hmf_1, \cdots\cdots, hmf_n)$

사례 y번의 벡터이다. hmfi는 가정개념 n의 j번째 가정개념에 대한 실수-값을 가진 구성원함수값이다.

지금까지의 벡터공간모델에 기반한 현재의 부울처리원형에 대한 확대를 개발하기 위하여 스마트와 사이어에서 상당한 연구가 이루어졌었고 이것은 더 효율적인 검색으로 이끌 것이다.

---

47) E.A. Fox, Sheila G. Winett,"Using vector and extended Boolean matching in an expert system for selecting foster homes," *Journal of the American Society for Information Science*, Vol.41, No.1(1990), pp.10-26.

48) *Ibid.*, p.16.

3. 퍼지 집합

전통적인 부울검색에서 특정한 문헌은 색인어가 주어졌느냐에 따라 해당되는 문헌집합에 속하거나 속하지 않게 된다. 반면 색인어에 가중치를 주는 경우에는 색인어가 주어진 문헌은 해당문헌집합의 부분멤버쉽(partial membership) 이 된다. 이렇게 부분적인 구성원을 허용하는 집합을 퍼지집합(Fuzzy set)이라고 한다.[49]

퍼지집합은 구성원의 경계가 뚜렷하지 않은 대상류로, 다시 말하면 고려된 대상들은 꼭 그러한 류에 속하거나 속하지 않거나 할 필요성이 없고 그 구성원등급은 [0, 1]의 중간에 조정될 수가 있다. 즉, 주어진 개념은 퍼지집합에서 전체집합의 특정요소들의 구성원등급이 특성함수의 일반화인 멤버쉽함수로 결정되는 것으로 보통 집합이론과 대비하여 멤버쉽과 비멤버쉽의 변이가 연속적인 것[50]을 말한다.

퍼지집합의 모든 조작은 이 멤버쉽함수에 의해서 정의[51]될 수 있다.

1) 포함관계(inclusion): 집합 A는 집합B의 부분집합으로 문헌집단 D에서 모든 포인트 x에 대해 $A \subseteq B$, if $f_A(x) \leq f_B(x)$로 표시한다.

2) 합집합(Union): A and B의 합집합, $A \cup B$는 멤버쉽함수 $f_{A \cup B}$를 가지고 이것은 $f_{A \cup B}(x) = \max[f_A(x) \cdot f_B(x)]$로 규정되었다.

3) 적집합(intersection): A AND B의 적집합 $A \cap B$의 구성원함수는 $F_{A \cap B}(x) = \min[f_A(x), f_B(x)]$로 주어졌다.

4) 보집합(Complementation): 퍼지집합의 보집합 A는 A로 표시되고 $f_A(x) = 1 - f_A(x)$로 규정된 멤버쉽함수를 가진다.

멤버시함수는 모든 부울식에서 다음과 같은 규칙에 따라 확대되었다.[52]

F(d, t AND t′)=Min(F (d, t), F(d, t′))

F(d, t OR t′)=Max(F (d, t), F(d, t′))

F(d, NOT t)=1-F(d, t)

(d, t)는 문헌-용어의 쌍

퍼지집합개념은 색인어에 가중치가 할당된 퍼지탐색문의 개념의 기초를 형성하게 된다.[53] 퍼지질의와 색인가중치의 이용은 두가지 목적을 가진다.[54] 첫째는 이용자질의를 얼마나 잘 매치시키는가에 의해서 각 문헌을 평가하기 위해서 가중치를 사용하고 두번째는 출력에 순위를 부여하기 위해서만 가중치를 사용한다. 가중치의 잠재적 이용은 하나는 0과 1의 값을 가지는 불연속의 가중치를 이용하는 것으로 부울검색시스템의 전통적인 접근법이며, 또 하나는 [0, 1] 구간에 계속되는 변수인 퍼지가중치도 사용할 수 있다는 것이다.[55] 이것은 가중치검색시스템을 일반화할 수 있게 허용한다.

전통적인 부울질의처리의 일반화는 단순히 색인을 퍼지화하는 것보다 더 복잡하며 질의표현도 가중치가 주어질 수 있다. 게다가 문헌적합성 평가장치는 검색상황값보다 다른 것을 생성할 수 있다. 이러한 일반화의 네가지 수준은 아래와 같다.[56]

---

49) 정영미, *op. cit.*, p.262.

50) 이순재, "정보검색시스템의 Fuzzy set 이론의 적용," 圖書館·情報學研究, 제 1 집(1989), p.213.

51) A. Bookstein, "Fuzzy requests : An approach to weighted Boolean searches," *op. cit.*, pp.241-242.

52) D.A. Buell and D.H. Kraft, "Threshold value and Boolean retrieval systems," *Information Processing Management*, Vol.17, No.3(1981), p.127.

53) A. Bookstein, "Fuzzy requests : An approach to weighted Boolean searches," *Journal of the American Society for Information Science*, Vol.31, No. 4(1980), p.240.

54) W.G. Waller and Donald H. Kraft, "A mathematical model of a weighted Boolean retrieral system," *Information Processing & Management*, Vol. 15, No.5(1979), pp.238-239.

55) *Ibid.*, p.238.

56) D.H. Kraft and D.A. Bueel, "Fuzzy sets and

1) 부울(0-1) 색인과 부울질의(non- 부울검색 상황값을 가진다)

2) 퍼지(non 0-1) 색인과 부울질의(퍼지부집합 규칙을 사용해서 계산한 검색상황값을 가진다)

3) 부울색인과 퍼지질의(전통적인 부울처리규칙과 동등한 것으로 알려진 검색상황 값이나 전통적인 처리출력의 출력을 순위매기는데 사용된 검색상황값을 가진다)

4) 퍼지색인과 퍼지질의(어떤 일반적인 함수로 계산되는 검색상황값을 가진다)

문헌검색시스템에 퍼지집합이론을 적용하는 것은 문헌과 질의의 일반화이며, 이것은 문헌에 non- 부울색인가중치를 부여하고 질의표현의 개별용어에 non-부울가중치나 기준치를 허용함으로써 수행되었다. 거기엔 질의가 부울논리를 포함할 때 부울구조의 보존과 관련된 문제가 있고 게다가 퍼지부집합을 포함하지 않는 검색모델이 있다.[57] 퍼지모델은 이런 다른 것들과 관련되었을 때 진정한 일반화의 시작이라 할수 있을 것이다.

부울질의처리방법을 일반화하고 향상시키려는 시도에 대해 상당한 연구가 있었으며 그러한 예가 퍼지집합모델이다. 이 모델은 표준부울논리와 양립할 수 있고 어떤 집단의 문헌(질의가 아닌)에 지정된 용어에 가중치를 부여하는 것이 가능하다고 제안되었었다. 퍼지집합이론 및 퍼지논리에 근거한 문헌검색시스템의 장점은[57]

① 문헌표현속의 색인용어의 다양한 중요도를 고려한다. ② 문헌은 가중치가 부여된 색인용어로 색인된 문헌검색시스템의 질의대표로서 색인용어들의 부울결합의 이론적으로-정당화된 이용을 허용한다. ③ 집합이론과 2진논리에 기반한 문헌검색방법은 퍼지집합이론과 퍼지논리ㅡ기반방법들의 특수한 경우이므로, 더 일반적인 문헌검색이론을 개발하는 것을 가능케 한다.

검색기법면에서 이 연구의 주요 공헌은 부울질의와 순위기법의 병합이었으나 벡터공간모델이나 확률모델에서의 용어의존성의 이용에 근거한 확대부울검색과 비교했을 때 제한되었다.[57] 또 퍼지집합모델은 단일질의용어에만 의존하는 OR-질의에 응답하여 문헌을 검색하고 모든 질의용어의 완전한 집합에 의존하는 AND-질의에 응답하여 검색한다[58]는 의미에서 보통의 부울처리모델과 꼭 같은 단점을 보여주고 있다.

V. 결 론

문헌검색에 대한 접근법의 결함, 불일치. 그리고 애매성으로 인해서 더 나은 해결방법은 찾는 것은 당연하다 하겠다. 적어도 문헌표현과 질의탐색공식에서 색인용어의 상대적인 중요성, 시스템출력에 대한 문헌의 순위부여, 그리고 정보검색 시스템에 이용자가 정보요구를 제출할 수 있는 일차적인 장치로서 부울질의를 나타내는 가능성을 병합하는 문헌검색방법론이 가능한지 어떤지에 대한 의문 또한 자연스러운 것이라 할 수 있다.

지금까지 문헌검색기법을 검색된 문헌들의 집합이 질문과 정확히 일치하느냐 아니면 부분적으로 일치하느냐에 따라 완전매치기법과 부분매치기법으로 나누어 조사하여 보았다. 완전매치기법은 부울탐색을 대상으로, 부분매치기법은 확률검색,

---

generalized Boolean retrieval systems," *International Journal of Man-Machine Studies*, Vol.19, No.1(1983), pp.48-49.

57) *Ibid.*, p.52.

58) T. Radecki, "Generalized Boolean methods of information retrieval," *International Journal of Man-Machine Studies*, Vol.18, No.5(1983), p.436.

59) N.J. Belkein and W.B. Croft, "Retrieval techniques," *op. cit.*, p.119.

60) G. Salton, "Some characteristics of future information systems," *op. cit.*, p.36.

벡터공간모델, 퍼지집합을 대상으로 하여 각각 그 장·단점을 분석하였다.

부분매치는 순수한 부울구조와 도치파일기술과 양립할 수 있는 여러 기술개혁을 이행함으로써 운용검색시스템에 상당한 향상을 기할 수 있을 것이다. 그러나 부분매치 역시 완전매치의 대안으로서라기 보다 부울구조안에서 검색을 향상시킨다는 점에서 한계를 가질 수 밖에 없다고 하겠다.

그런 관점에서 앞으로의 연구및 방향은 부울논리처럼 명확한 정보요구를 가진 이용자뿐 아니라 명확하지 못한 정보요구를 가진 이용자도 도울 수 있어야 할 것이다.

그러나 상업서비스제공자들은 운용시스템에서 현재 사용하고 있는 주요정보, 특히 문헌색인과 탐색요청공식절차와 양립할 수 없는 한 특별한 기술개혁을 바라지 않은 것이다. 그래서 제시된 개선방안들은 현존검색시스템의 서비스제공자들이 지금까지 기울인 노력을 헛되이 하지 않고 도입할수 있어야 한다. 뿐만 아니라 연구자들도 연구의 잠재력을이용한 필요성을 알게 하고 서비스제공자도 연구가 이론이나 실험으로 그치지 않고 현실적인 이용가능성이 있음에 대한 인식의 제고가 있어야 할 것이다.

참 고 문 헌

신성철, 불리언논리탐색문의 자동생성을 위한 전문가 중개시스템, 미간본석사학위논문, 경북대학교 대학원, 1988.

이순재, "정보검색시스템에 Fuzzy Set이론의 적용," 四書館·情報學研究, 제 1 집(1989), pp. 201~234.

정영미, 정보검색론, 서울: 정음사, 1987.

Belkin, N.J., Croft, W.B., "Retrieval techniques," ARIST, V.22(1987), pp.109~145.

Bookstein, A., "A comparison of two systems of weighted Boolean retrieval," Journal of Asis, V.32, No.4(1981), pp.275~279.

_____, "Fuzzy requests: an approach to weighted Boolean Searches," Journal of Asis, V. 31, No.4(1980), pp.241~247.

_____, "on the perils of merging Boalean and weighted retrieval systems," Journal of Asis, V.29, No.3(1978), pp.156~157.

Buell, D.A., Kraft, D.H., "Threshold values and Boolean retrieval systems," Information Processing & Management, V.17, No.3(1981), pp.127~130.

Cooper, W.S., "Exploiting the maximum entropy principle to increase retrieval effectiveness," Journal of Asis, V.34, No.1(1983), pp.31~39

_____, "Getting beyond Boole, "Information Processing & Management V.24, No.3(1988), pp.243~248.

Croft, W.B., "Boolean queries and term dependencies in probabilistic retrieval models, "Jounal of Asis, V.37, No.2(1986), pp.71~77.

Croft, W.B., Harper, D.J., "Using probabilistic modele of document retrieval without relevance information," Journal of Documentation, V.35, No.4(1979), pp.285~295.

Fox, E.A., Koll, M.B., "Practical enhanced Boolean retrieval: experiences with the SMART and SiRE systems," Information Processing & Management, V.24, No.3(1988), pp.257~267.

FOX, E.A., Winett, S.G., "Using vector and extended Boolean matching in an expert

system for selecting foster homes," Journal of Asis, V.41, No.1(1990), pp.10~26.

Gordon, M.C., "The necessity for adaptation in modified Boolean document retrieval systems," Information Processing & Management, V. 24, No.3(1988), pp.339~347.

Hildreth, C., "To Boolean or not to Boolean?," Information Technoology & Libraries, V.2(1 983), pp.235~237.

Koll, M.B., Noreault, T., McGill, J., "Enhanced retrieval techniques on a microcomputer," Proceedings of the National online Meeting, (1984), pp.165~170.

Kraft, D.H., Buell, D.A., "Fuzzy-sets and generalized Boolean retrieval systems," International Journal of Man-Machine Studies, V.19, No. 1(1983), pp.45~56.

Losee, R.M., Bookstein A., "Integrating Boolean queries in conjunctive normal form with probabilistic retrieval models," Information Processing & Management, V.24, No.(1988), pp. 315~321.

McCall, Fiona. M., Willett, P., "Citeria for the selection of search strategies in best-match document-retrieval systems," International Journal of Man-Machine Studies, V.25(198 6), pp.317~326.

Noreault, T., Koll, M., McGill, M.J., "Automatic ranked output from Boolean searches in SiRE," Journal of ASIS, V.28, No.6(1977), pp.333~3 39.

Radecki, R., "Generalized Boolean methods of information retrieval," International Journal of Man-Machine Studies, V.18, No.5(1983), pp.407-439.

____, "A probabilistic approach to information retrieval in systems with Boolean search request formulations," Journal of ASIS, V.33, No. 6(1982), pp.365~370.

____, "probabilistic methods for ranking output documents in conventional Boolean retrieval systems, "Information Processing & Management, V.24, No.3(1988), pp.281~302.

____, "Reducing the perils of merging Boolean and weighted retrieval systems," Jounal of Documentation, V.38, No.3(1982), pp.207~2 11.

____, "Trends in research on information retrieval-the potential for improvements in conventional Boolean retrieval systems," Information Processing & Management, V.24, No.3(1988), pp.219~227.

Robertson, S.E., "The probability ranking principle in IR," Journal of ASIS, V.2730(1976), pp.129~146.

Salton, G., Introduction to modern information retrieval, New York: McGraw-Hill, 1983.

____, Dynamic information and library processing, Englewood cliffs: Prentice-Hall, 197 5.

____, "Some characteristics of future information systems", SiGiR Forum, V.18(1985), pp. 28~39.

Salton, G., Voorhees, E., Fox, E.A., "A comparison of two methods for Boolean query relevancy feedfack, "Information Processing & Management, V.20, No.5/6(1984), pp.637~6 51.

Salton, G., Buckley, C., Fox, E.A., "Automatic queuy formulatione in information retrieval," Journal of ASIS, V.34, No.4(1983), pp.262~280.

Salton, G., Fox, E.A. Wu, H., "Extended Boalean information retrieval", Communictions of the ACM, V.26, No.11(1983), pp.1022~1036.

Salton, G, Bucklay, C., "Term-weighting approaches in automatic text retrieval," Information processing & Management, V.24, No.5(1988), pp.513~523.

Shoval, P., "Principles, procedures and rules in an expert system for information retrieval, "Information Processing & Management, V. 21, No.6(1985), pp.475~487.

Smith, L.C., "Artificial intelligence in information retrieval," ARIST, V.12(1976), pp.189~222.