정보검색 연구의 방법론에 관한 고찰

이 명 희*

목 차

1. 서 론

지난 30년간 정보검색연구에 있어서 유용한 결과를 얻기위해 실험상의 꾸준한 진보가 있었으나 중요한 문제는 방법론에 있었다. 방법론의 잠재적 문제점은 연구자들에 의해 사용된 인위적인 데이타, 작은 표본크기, 외부변인에 대한 통제실패, 결과에 대한 부적절한 해석 등이 있다. 또한 실험실에서의 실험은 때때로 현실세계를 반영하지 못하였으며, 정보검색 연구의 방법론에서 개선이 있어야 한다고 비판자들은 지적하고 있다. 실존하는 데이타의 사용, 충분한 탐색 질문수와 함께 충분히 큰 표본의 크기, 변인에 대한 철저한 통제, 더욱 적절한 검색효율 측정도구의 사용, 주의 깊게 실시된 테스트, 결과에 대한 적절한 해석 등이 이루어져야 한다고 지적되고 있다 (Robertson, 1981). 정보검색의 실험절차를 결정하는데 있어서는 다음의 세가지가 고려되어야 한다.

* 성균관대학교 사서교육원 강사

1) 절차의 신뢰성 : 그 실험이 정말로 원래 그 연구자가 의도한대로 원하는 것을 보여주고 있는가? 만약 어떤 연구가 이용자 만족에 대한 문헌의 범위에 관계되는 것이라면, 범위의 척도로서의 인용문헌수의 사용과 만족의 척도로서 "적합한" 것으로 판정된 참고문헌의 수는 진정으로 이 목적을 만족시킬 수 있는가?

2) 절차의 타당성 : 이 실험이 다른 실험자에 의해 시행될 때 동일한 결과를 가져올 수 있는가? 만약 어떤 사람이 색인자간 일관성의 문제에 대해 연구할 때 같은 잡지로부터 열개의 문헌을 색인하는 두 색인자간의 일관성 테스트는 다른데서 시행되어도 동일한 결과를 낳을 수 있는가? 어떤 테스트는 유효하지 않고도 타당할 수 있는데, 예를 들면 일관된 결과를 낳지만 평가하고자 의도했던 것이 아닌 다른 것을 평가하는 경우가 이에 해당된다. 다시 말해 원래의 의도는 검색의 효율성을 평가하는 것이나 검색의 과정을 평가하고서도 얼마든지 타당한 검사를 할 수 있는 것이다. 그러므로 사실상 타당성보다 신뢰성은 더욱 중요한 요소라 말할 수 있다.

3) 테스트 과정의 효율성: 그 테스트를 위해서 사용되는 시간, 자원, 경비가 적절한가? 예를 들면 각 이용자가 전체 데이타베이스를 고려할 때 탐색의 절대재현율을 평가하는 것이 적절한가 아니면 데이타베이스의 크기에 어떤 제한점을 가지고 있는 것이 나은가?

이 논문은 특히 신뢰성과 타당성을 중심으로 하여 고찰하여 보았다. 먼저 Campbell과 Stanley에 의해 제기된 설계의 타당성을 위협하는 요소들을 나열하여 보고 거기에 준하여 데이타베이스의 비교, 질문, 통제의 실패에 대해 논하였다. 특히 통제의 실패에서는 탐색과정에서의 여러 변인들인 탐색자간의 일관성, 탐색자내의 일관성, 탐색전략의 일관성에 대해 언급하였다. 적합성의 평가에 대한 정의와 적합성의 등급, 문헌의 제시순서 등이 고려되었으며 검정을 위해 사용된 통계방법에 대해 언급하였다. 특히 비모수통계인 sign test와 Wilcoxon test에 대한 자세한 설명이 있었다.

2. 설계의 타당성을 위협하는 요인들.

설계의 타당성을 위협하는 요인은 내적타당성(internal validity)과 외적타당성(external validity)으로 나뉠 수 있다. 내적타당성이란 그 실험의 처치가 특정 실험에서 의도했던 바대로 실제로 어떤 차이를 만들어 내었는가를 따지는 타당성이다. 내적타당성은 근본적으로 그 실험이 이차적 변인을 포함하느냐 여부로 결정된다. 이차적 변인이 들어 있지 않으면 내적으로 타당한 실험이라 말할 수 있다. 다시말해 종속변인 y의 결과는 독립변인 x의 차이에 의한 것인가를 알아보는 것으로서 이것은 주로 이차적 변인을 통제함으로써 가능하다. 내적으로 타당한지 여부는 그 실험이 실험으로서 성립할 수 있느냐 여부를 결정짓는 기준이 된다.

외적타당성은 실험에서 나온 결과를 얼마나 여러 상황에 일반화시키느냐, 즉 적용할수 있느냐 하는 문제이다. 이 실험의 처치효과가 어떤 모집단에, 어떤 사태에, 어떤 처치변인에, 어떤 측정변인에 일반화할 수 있는가를 문제삼는 타당성이며 그 결과의 적용범위가 넓으면 그 실험은 외적타당성이 높다고 말한다. 원칙적으로 내적타당성이 없는 실험에 대해서는 외적타당성은 말할 필요도 없으며 실제에 있어 대체로 내적타당성과 외적타당성과의 관계는 상호반비례적이 되기 쉽다. 즉, 통제를 가하여 내적타당성을 증가시키면 외적타당성이 낮아지고, 그 결과를 다른 유사한 상황에 일반화시키기 위하여 외적타당성을 높이면 내적타당성이 낮아지기 쉽다.

내적타당성을 위협하는 요인들을 Campbell과 Stanley는 6가지로 열거하고 있다 (1963). 첫째 역사(history)인데 이는 실험변인에 첨가되어서 최초의 측정과 두번째 측정사이에서 일어나는 특수한 사건을 말한다. 둘째 성숙(maturation)은 나이를 먹어간다든지, 공복감을 더 느낀다든지, 더 피곤해졌다든지 하는 것과 같은 것을 포함해서 시간의 경과 그 자체로 인해서 생기거나 작용하는 피험자 속에 진행되는 여러 과정들을 말한다. 셋째 검사(test)는 이전에 받은 테스트가 다음의 테스트에 영향을 미치는 것이고, 넷째 도구사용(instrumentation)은 측정도구의 눈금의 변동이나 관찰자나 채점자의 변동이 결과 측정

치에 변화를 가져오게 되는 것이다. 다섯째 통계적 회귀(statistical regression)는 아주 극단적인 점수를 기초로 해서 집단선정을 할 경우 작용하는 통계적 현상을 말한다. 여섯째 피험자의 선정(selection)은 비교집단을 만들기 위해서 피험자를 선정할 경우 생겨나는 집단간의 편향을 말한다.

외적타당성에서 가장 중요한 대표성을 위협하는 요인들을 설명하면 다음 4가지로 정리할 수 있다. 첫째 검사실시의 반동적 영향 또는 상호작용적 영향(the reactive or interaction effect of testing)인데 여기서는 사전검사가 피험자의 실험변인에 대한 감수성이나 반응성에 영향을 줄지도 모른다는 점을 말하고 있다. 그래서 사전검사를 받는 모집단에서 얻은 결과가 실험에 동원된 피험자들이 추출되어 나온 사전검사를 받지않는 모집단에서 얻는 실험변인의 효과를 전체적으로 대표할 수 있는 결과가 되지 못하게 할지도 모른다. 둘째 피험자 선정으로 인한 편향과 실험변인과의 상호작용효과(the interaction effect of selection biases and the experimental variables)이고, 셋째 실험배치의 반동효과는 비실험적 사태하에 피험자를 노출시켰을 때 실험변인이 미칠 영향에 관해서 어떤 일반화를 내리는 것을 방해하는 요소가 된다. 넷째 다중처치간 간섭(multiple-treatment interference)은 여러가지 실험변인의 조작을 같은 피험자에게 적용시켰을 때 일어날지 모르는 효과이다.

3. 데이타베이스의 비교

정보검색실험에서 사용되는 데이타베이스는 세가지로 나누어지는데, 첫째 실험데이타베이스의 구축, 둘째 기존의 실험데이타베이스, 셋째 실제로 운영되는 데이타베이스이다. 실험적인 데이타베이스를 구축하는 일은 비용이 많이 들고 시간이 걸리며, 또한 통계적으로 테스트되기에는 종종 양이 너무 적다. 실험적이거나 실제 운영되는 데이타베이스를 사용할 때 내적타당성을 증가시키기 위해 많은 외부적인 변인을 통제하는 것이 필요하다.

검색효과를 비교하기위해 단일 데이타베이스를 사용하는 것이 바람직하지만 단일 데이타베이스가 없을 때 현재 운영되고 있는 데이타베

이스에 통제를 가하면서 조심스럽게 사용되어야 한다. 예를 들면 수록 연대, 언어, 범위, 매체, 색인언어의 망라성 등이 검색효율 측정에 영향을 미치므로 이러한 특성등을 고려하여 데이타베이스의 선정이 이루어져야 한다. 또한 결과의 해석시에도 이러한 속성이 고려된 제한된 영역에서의 결과만을 기술하여야 외적타당성을 손상하지 않을 수 있다. 그러나 어떤 연구자들은 이러한 데이타베이스들간의 특성을 고려하지 않고 단순비교함으로써 바이어스를 초래하게 되었다 (Katzer et al., 1982; White and Griffith, 1987; McCain et al., 1987; McCain, 1989).

검색연구 결과가 신뢰성있는 것인가하는 질문에 대한 대답은 그 연구결과가 현재 운영중인 시스템에 응용되었던 결과인가 하는데 있다. 실험적인 데이타베이스로부터의 결과는 통제가 용이하다는 장점이 있음에도 불구하고, 좋은 연구결과는 비록 그것이 긍정적이든 부정적이든간에 실지 운영되는 시스템으로부터의 결과에 근거하여야 외적타당성을 증가시켜서 일반화시킬 수 있는 것이다 (Spark Jones, 1981). 그런 의미에서 Lancaster에 의해 행해진 MEDLARS 평가는 방법론상의 문제점에도 불구하고 고무적이라 할 수 있다. 실험실에서 실험적인 장치를 통해 이루어진 결과가 어떻게 한 모집단에 대한 표본이라고 정당화할 수 있는가가 실험실 테스트의 문제점인데, 우리가 그렇게 할 수 있는 유일한 방법은 그들 질문이 선택되어진 기준을 고려하고 이러한 기준을 만족시키는 질문이나 문헌의 모집단을 정의하는 것이다.

4. 인위적인 질문

질문이 정보요구자로부터 적절하게 선택되면 그 질문을 통하여 질문의 특성에 대한 합리적인 추측을 할 수 있게 된다. 그러나, 인위적인 질문은 샘플로서 모집단의 실질질인 표집이 아니라는 점이다. 이러한 인위적인 질문을 얻는 방법은 실제적이거나 잠재적인 이용자에게 최근 수년동안의 어떤 시스템에서의 가능한 질문을 요구하는것, 정보전문가에게 특정 주제분야의 질문의 샘플을 요구하는것, 색인용어의 무작위 추출에 의해 실제질문과 유사한 질문을 구성하는 것 등

이다. 비록 인위적인 질문이 실제의 질문을 모방하려해도 문제는 재창조하려고 하는 실제질문의 중요한 특징을 우리가 잘 모르고 있다는 사실이다. 어떤 연구에서 연구자들은 전문가들에게 샘플질문을 요구하고 또한 거기에 따른 검색결과가 가장 적절할 것이라고 생각되는 답을 결정하도록 요구하였다 (Tenopir, 1984; Keen and Digger, 1972; Hersey et al., 1970). 이것은 정보검색에서의 전형적인 전문가 패널 방법이지만 각 전문가들은 같은 문헌에 대해 적합성의 판정을 각각 다르게 하기 때문에 그 효용성은 점차 회의적이다.

5. 통제의 실패

일반적으로 실험연구에 있어서 종속변인의 효과가 독립변인으로부터 기인한다는 것을 보여주기 위해 변인들의 통제가 필요하나 어떤 연구자들은 이러한 통제에 실패했다. 이러한 경우 종속변인의 결과가 독립변인때문이라고 하는 것은 신뢰성을 주지 못하고 있다. 따라서 비교된 두 그룹은 통계적으로 중요한 차이를 드러내지 못하고 말았다. 탐색자간의 일관성, 탐색전략의 일관성, 탐색자내의 일관성과 같은 탐색전략의 수립에 영향을 미치는 변인들은 실험설계에서 충분히 고려되어야 할 사항들이다.

다른 경험, 교육, 개인적 특성을 가진 탐색자들의 탐색은 탐색결과에 큰 차이를 보여준다는 것이 관찰되었으며 (Katzer, 1973; Martin, 1973; Katzer et al., 1982; Saracevic and Kantor, 1988), 탐색자간 일관성(inter-searcher consistency)은 중요한 것으로 밝혀졌다. 특히 탐색전략을 연구한 Fenichel(1981)은 다른 경험을 가진 탐색자들은 같은 문제를 해결하기 위해 다른 탐색전략을 사용한다고 보고하였으며, 특히 전문적인 탐색자들 사이에서도 개인에 따라 탐색전략에 커다란 차이를 가지고 있다고 지적하였다. Fidel(1985)은 2개의 테스트 질문을 탐색하기 위해 10명의 전문적인 탐색자를 대상으로 연구하였는데 탐색자의 차이에 관계되는 두 요소를 파악하였다. 첫째 이용자 인터뷰의 부재시나 이용자의 부재시에 탐색자들은 큰 차이를 드러내었으며, 둘째 전문적인 질문보다 더욱 일반적인 질문에 있어서 탐색자들은 큰 차이를 보여 주였다. 그러나 Fenichel(1979)의 연구에서 전문

적 탐색자로부터 데이타를 수집하는 경우에 있어서 그들의 탐색결과를 우편으로 받았는데, 이 과정에 있어서 탐색전략개발과 탐색에 소요된 시간과 탐색환경, 탐색건수 등에 대한 통제를 가하지 않았다. 또한 탐색자의 경험의 과다에 대한 구분이 없었다. 따라서 관찰된 결과의 차이는 탐색자간의 차이에서 기인하는 것인지 아니면 다른 변인들 (예를 들면 전체 명령어의 수, 사용된 디스크립터의 수, 탐색건수당 소요시간 등)에 의해 야기된 것인지 알 수가 없으며 그러므로 탐색자의 행위가 다른 탐색결과를 낳게하는 요인이라고 하는 것을 확신을 가지고 주장하기는 매우 어렵다. 따라서 일관성있는 탐색자의 사용이 탐색자간의 차이에서 오는 바이어스를 통제할 수 있는 중요한 요소임은 부인할 수 없다. 동일 탐색자의 시간의 추이에 따른 일관성 (intra-searcher consistency)도 고려되어야 할 사항이다. 같은 탐색자라 하더라도 시간이 지나고 환경이 바뀜에 따라 탐색의 차이를 가지게 된다는 것은 알려진 사실이다 (Lancaster, 1979). 따라서 동일한 시간적, 공간적 환경을 조성하기 위해서 연구자는 각별히 신경을 써야 할 것이다.

6. 적합성의 평가

정보시스템에 의해 검색된 문헌은 이용자의 요구에 대한 적합성의 판정에 의해 평가되어졌다. 정보검색에서 적합성이란 단어는 역사적으로 여러가지 명칭으로 사용되었고 그 정의의 애매모호성으로 인해 어려움을 겪게 되었다 (Schamber et al., 1990). 초기에는 적합성의 판정이 시스템적 관점에서 이루어진, 즉 이용자의 질문에 매치하는 문헌을 검색하는 시스템의 능력으로 이해되었으나, 현재는 주제적 적합성(topicality)을 넘어서는 복잡한 개념으로 인식되어졌다. 주제적 적합성도 이용자가 적합성을 결정하는데 있어서 어떤 역할을 담당하기는 하지만 그것은 어디까지나 적합성의 한 부분이라는 것을 여러 문헌들은 밝혀주고 있다. 이후의 많은 저자들은 시스템 중심의 적합성에 대한 정의를 벗어나서 이용자 중심의 정의를 제공하려고 시도하였다. 이러한 정의들은 유효성(pertinence), 유용성(usefulness), 효용성(utility), boutness 또는 만족(satisfaction) 등으로 불리고 있다

(Wilson, 1973; Hillman, 1964; Cooper, 1971; Foskett, 1972; Kemp, 1974). 적합성에 대한 자세한 분석을 시도했던 Saracevic(1975)도 여러 종류의 적합성을 구분하면서 궁극적으로 적합성 판정은 주어진 특수한 상황하에서의 이용자의 판단에 전적으로 의존해야 한다고 제안하였다. 예를 들면, 적합성 판정에 영향을 미치는 논문의 서지사항은 주제를 포함하지 않는 데이타화일을 포함하는데, 그것들은 잡지명, 저자명, 출판날짜, 저자의 소속기관 등이다. 또한 이용자의 주제에 대한 전문지식, 참신성(novelty), 탐색목표, 연구단계, 탐색의 적절성(urgency) 등도 적합성 판정에 영향을 미치는 판정자의 내적, 외적 요소들이다.

적합성은 이용자가 자신이 처해진 특수상황에서 정보를 인식하게 되는 다차원적인 인식개념이다. 적합성은 개념적으로 뿐 아니라 운영적으로도 이용자적 측면에서 접근되어진다면 조직적이고 또한 측정될 수 있는 개념이다. 실험에 있어서, 적합성의 최적의 판단자는 그 정보요구를 가장 잘 알고있는 이용자 자신이다. 그러므로 적합성의 평가에 있어서 첫째 질문은 누가 평가를 하는가인데, 적합성 판정의 평가자는 분명히 요구자여야 한다. 그런데 어떤 연구자들은 적합성 판정을 일련의 전문가 집단에 의존했다 (Keen and Digger, 1972; Hersey et al., 1970; Tenopir, 1984). 그러나 이러한 전문가 패널에 의한 적합성 판정은 각각 전문가들이 같은 문헌에 대해서도 다르게 판정하기 때문에 신뢰성이 없는 것으로 간주되어지고 있다.

두번째의 문제점은 적합성의 등급을 결정하는데 있다. 과거의 적합성 판정은 그 문헌이 정말 적합한가 적합하지 않은가의 이분법적인 기반위에서 행해졌다. 그러나 1958년 '과학정보를 위한 국제회의'의 합의이후 적합성은 정도(degree)를 가지고 있기 때문에 그 판정은 이분법적이 아니라 둘 이상의 카테고리가 이상적이라고 생각되어졌다. 특히 많은 양의 문헌이 데이타베이스로부터 검색될 때 다양한 카테고리를 사용하는 것이 좋다.

여러가지의 척도가 적합성 판정을 위해 제안되었다.

1) 3 value 적합성 : 아주 적합하다, 부분적으로 적합하다, 적합하지 않다.
2) 등급화된 적합성 : 적합성에 대한 순위가 1-5 사이에서 정해졌

다.

3) 가중치 : 이용자에 의해 질문에 대한 적합성에 가중치가 주어졌다.

실질적인 문제점은 적합성의 등급을 결정하는데 있으며 일관성있는 태도로 평가자들이 적합성 판정을 할 수 있도록 가르치는 것이다. 평가자가 이러한 등급을 선택하는데 있어서 타당성에 대한 고려가 있어야 하는데 같은 평가자에 의해 등급화된(평가된) 적합성이 일관성이 있는가? 일관성은 평가자간의 일관성과 한 평가자 내의 일관성 모두를 포함하는데 Lesk와 Salton의 연구에 의하면 적합성의 등급화는 비교적 인정적이라고 한다.

셋째 적합성 판정을 위하여 판정자가 문헌의 어느 부분의 정보까지를 보아야 하는가 하는 것이다. 이론적으로는 그 문헌의 본문까지를 포함한 전체를 다 보아야 하지만 실제로 초록을 포함한 서지사항 등이 사용되어진다. 이것 또한 전체 문헌을 통틀어 서지사항 등이 일관성있게 제공되어야 한다.

넷째 판정자가 이미 판정한 한 문헌의 내용이 다른 문헌의 판정에 영향을 미칠 수가 있으므로 판정자에게 문헌이 제시되는 순서 또한 중요하다. 따라서 문헌이 제시되는 순서가 판정자의 판정에 영향을 미치는 것을 방지하기 위하여 연구자는 최선을 다하여야 한다. 예를 들어, 한 시스템으로부터의 결과를 다 판정한 후에 다른 시스템으로부터의 결과를 판정하는 것이 아니라 두 시스템의 결과를 섞어서 판정케 하는 것이 순서에 의한 바이어스를 줄일 수 있는 한 방법이 된다.

7. 통계

검색시스템의 검정은 필연적으로 어떤 종류의 측정을 포함하고 있는데 검색시스템을 검정하는 이유는 미래의 검색을 위한 어떤 특징을 발견하거나 추론해내기 위한 것이다. 통계적 추론의 일반적 방법은 그 표본이 추출되는 모집단에 대한 전제에 근거하고 있다. 모집단의 정규분포, 동일 분산, 그리고 적어도 등간척도 이상의 데이타 등의 전제를 근거로 한다. 이러한 조건을 만족시킬 때 T-test, 분산분석

(ANOVA), 회귀분석 (Regression) 등이 사용된다. 만일 정확률 스코어가 정규분포를 따른다는 것을 가정할 수 있다면 T-test를 사용할 수 있을 것이다. 셋 이상의 실험처치를 위해서는 분산분석 등이 쓰이는데 이 역시 모집단이 정규분포를 이루는 것을 전제한다. 다른 처치하의 샘플의 분산이 동질이 아닐 때 원자료의 변형(transformation)에 의해 안정화시킬 수 있다. 흔히 쓰이는 변형은 square root, logarithmic, 그리고 arcsin 등이다. 따라서 데이타가 정규분포에 근사할 때 분산을 변형함으로써 정규분포를 만들 수 있다. 실례로 *arcsin* 변형은 재현율과 정확율 등의 비율의 정규분포를 증가시키고 분산을 안정화하는데 유용하며, 낮은 값(value) 쪽으로 기울어진 times는 logarithmic transformation에 의해 분포를 개선시킬 수 있다. 검색된 문헌수가 질문내에서 너무 변동이 심할 때 square root 변형을 씀으로써 데이타를 정규분포에 근접시켜 준다. 모평균이 측정한 값과 같다는 가설을 검정하기 위해 테스트를 사용하려면 모집단의 분포가 적어도 근사적으로 정규분포를 따라야 하나 정보검색에서 검정하기를 원하는 많은 변인들은 정규분포를 벗어난다. 그 좋은 예가 재현율인데 이것은 적합한 전체 문헌 중 검색된 적합문헌의 비율이다. 실제로 사람들은 높은 재현율을 가진 문헌을 검색하기 원하며, 주어진 질문에 대해 아주 적은 문헌만이 적합하며 또한 재현율의 값은 넓게 퍼져 있으므로 질문에 대한 재현율 값의 분포는 정규분포를 이루고 있지 않은 경우가 많다. 또한 질문내에서 검색된 문헌의 양은 강하게 skewed 되고 정규분포를 가지고 있지 않은 경우가 많다. Van Rijsbergen(1979)은 실제로 정보검색에서는 모집단 분포의 형태가 알려져 있지 않기 때문에 어떠한 알려진 통계적 테스트도 적합하지 않다고 주장하고 있다. 이러한 경우에 있어서 적절한 통계를 발견한다는 것은 대단히 어렵기 때문에 비모수통계에 의존해야 한다. 비모수통계는 가정을 거의 요구하지 않을 뿐 아니라 순위와 서열로 구성된 자료를 분석하는데도 이용된다. 특히 약간의 전제만을 요구하고 있는 sign test나 Wilcoxon test와 같은 비모수통계는 어느정도 대안이 될 수 있으나 일반적으로 비모수통계는 정보검색 연구에서 여러 측정치(measures)를 사용하는 복잡한 연구계획을 위해서는 잘 개발되어 있지 않는 실정이다. 이것은 통계학자들과 정보학자들에 의해

더욱 연구 개발되어야 할 분야이다.

Sign test는 대응표본으로 발생하는 자료를 분석하는데 이용되는데 자료는 수치, 순위 또는 선호도로 구성된다. Sign test는 모집단에서 관찰치의 반이 어떤 특징을 가지고 다른 반은 그 특징을 가지지 않는다는 가설을 검정하는데 사용된다. 예를 들면, 검색효율 측정에 있어서 한 집단의 결과와 다른 집단의 결과는 차이가 없다는 귀무가설을 검정하는데 이용된다. 차이가 0보다 크면 (+)로 기록하고 차이가 0보다 작으면 (−)로 기록한다. 만약 차이가 0이라면 표본으로부터 그 관찰치를 없애고 표본의 크기를 하나 줄인다. 여기서 sign test는 (+)의 수와 (−)의 수가 거의 같다는 귀무가설을 검정하는데 이용된다. Sign test는 대응관찰치를 근거로 두 모집단의 평균이나 중앙값이 같은지를 검정하기 위해 사용되나 대응관찰치 사이에 대한 부호만을 이용하고 차이의 크기는 무시한다. 이에 대한 대안이 Wilcoxon 의 순위검정(Mann-Whitney-Wilcoxon test)이다. 이 검정은 대응관찰치 사이의 차이에 대한 부호뿐 아니라 차이의 순위를 지료로 요구한다. 이것이 더 많은 정보를 사용하므로 부호검정보다 우수하다. 반면에 대응관찰치 사이의 차이에 대한 순위가 때때로 알려져 있지 않기 때문에 그것은 부호검정만큼 흔히 적용되지는 않는다. 모집단 1로부터 n개의 관찰치 표본과 모집단 2로부터 n개의 관찰치 표본을 얻었다고 하자. 여기서 표본 1의 각 관찰치는 표본 2의 하나의 관찰치와 관계되어 있다. 귀무가설은 두 모집단이 동일하다는 것이다. 만약 귀무가설이 사실이라면 대응관찰치 사이 중에서 반은 음이고 반은 양으로 기대된다. 각각의 대응관찰치에 대한 절대값을 계산하여 최저값에서 최고값까지 순서를 정하고 양수인 차이에 대한 순위의 합을 얻어 이 합을 T+로 나타낸다. 그리고 나서 음수인 차이에 대한 순위합을 얻어 T−로 나타낸다. 만약 이 두 분포가 동일하다면 두 순위의 합이 대략 같다는 것을 기대할 수 있다. 그러나 T+와 T−가 크게 다르다면 모집단 1과 모집단 2가 다른 분포를 갖는다는 가설이 지지된다.

8. 결론

정보검색연구에서 이루어지고 있는 방법론상의 문제점을 살펴 보았

다. 정보검색의 실험절차는 신뢰성, 타당성, 테스트 과정의 효율성을 고려하면서 결정되어야 한다. 특히 설계의 타당성을 위협하는 내적타당성과 외적타당성을 위협하는 요인들은 세심히 통제되면서 사용되어야 하는데 이들 요인들은 상호 반비례적이 되기 쉽기 때문에 실험설계의 목적 등을 고려하여 결정되어야 한다. 데이타베이스의 사용에 있어서 운영중인 시스템내에서의 단일 데이타베이스를 사용하거나 두 데이타베이스의 특성을 고려하여 사용되어야 하며 결과의 해석시에도 이를 염두에 두어야 한다. 탐색자간의 일관성, 탐색전략의 일관성, 탐색자내의 일관성을 유지하여 탐색전략 수립에서 외부적인 변인의 통제가 이루어져야 한다. 적합성의 평가는 주어진 상황에서 정보요구를 가진 이용자 자신에 의해 이분법 이상의 다양한 등급에 따라 주어져야 한다. 정보검색에서는 데이타가 정규분포를 이루는 경우가 극히 제한되어 있으므로 비모수통계가 흔히 사용되어지는데 특히 sign test와 Wilcoxon test는 좋은 대안이 될 수 있다.

참 고 문 헌

Campbell, D.T. and Stanley, J.C.(1963), *Experimental and Quasi-Experimental Designs for Research.* Chicage : Rand McNally.

Cooper, W.S.(1971), A Definitions of Relevance for Information Retrieval. *Information Storage and Retrieval.* 8(2), 457-471.

Fenichel, C.H.(1979), *Online Information Retrieval: Identification of Measures that Discriminate among Users with Different Levels and Types of Experiences.* Doctoral Dissertation. Philadelphia, PA : Drexel University.

Fenichel, C.H.(1981). Online Searching: Measures that Discriminate among Users with Different Types of Experiences. *Journal of the American Society for Information Science.* 32, 23-32.

Fidel, R. (1985), Individual Variability in Online Searching Behavior. *Proceedings of the 48th ASIS Annual Meeting.* Las Vegas, Nevada. 22, 69-72.

Foskett, D.J.(1972). A Note on the Concept of Relevance. *Information Storage and Retrieval.* 8(2), 77-78.

Hersey, D.F. et al.(1970). Comparison of Online Retrieval Using Free Text Words and Scientist Indexing. *Proceedings of the American Society for Information Science 33rd Annual Meeting.* Philadelphia, PA . 256-268.

Hillman, D.J.(1964). The Notion of Relevance (1). *American Documentation.* 15(1). 26-34.

Katzer, J.(1973), The Cost Performance of an Online Free Text Bibliographic Retrieval System. *Information Storage and Retrieval.* 9, 321-329.

Katzer, J. et al,(1982). *A Study of the Impact of Representations in Information Retrieval Systems.* Syracuse, N.Y.; Syracuse University.

Keen, E.M. and Digger, J.A.(1972). *Report of an Information Science*

Index Languages Tests. Aberystwyth, U.K. : College of Librarianship Wales. 2 vols.

Kemp, D.A.(1974), Relevance, Pertinence and Information System Development. *Information Storage and Retrieval.* 10(2), 37-47.

Lancaster,F.W.(1979), *Information Retrieval Systems; Characteristics, Testing and Evaluation.* 2nd ed. N.Y.: John Wiley & Sons.

Martin, W.A.(1973), A compariative Study of Terminal User Techniques in Four European Countries on a Large Common Online Interactive Information Retrieval System. *First European Congress on Documentation Systems and Networks.* 107-167.

McCain, K.W.(1989), Descriptor and Citation Retrieval in the Medical Behavioral Science Literature; Retrieval Overlaps and Novelty Distribution. *Journal of the American Society for Information Science.* 40(2), 110-114.

McCain, W., White, H.D. and B.C. Griffith (1987), Comparing Retrieval Performance in Online Databases. *Information Processing and Management.* 23(6), 539-553.

Robertson, S.E.(1981), The Methodology of Information Retrieval Experiment. ed by Karen Spark Jones. *Information Retrieval Experiment.* London : Butterworths, pp. 9-13.

Saracevic, T.(1975), Relevance; a Review of and a Framework for the Thinking on the Notion in Information Science. *Journal of the American Society for Information Science.* 26, 321-343.

Saracevic, T. and Kantor, P.(1988), A Study of Information Seeking and Retrieving. 3 parts. *Journal of the American Society for Information Science.* 39, 161-216.

Schamber, L., Eisenberg, M.B. and Nilan, M.S.(1990). A Reexamination of Relevance; Toward a Dynamic, Situational Definition. *Information Processing and Management.* 26, 755-776.

Spark Jones, K.(1981), Retrieval Systems Tests 1958-1978. ed by

Karen Spark Jones. *Information Retrieval Experiment.* London: Butterworths. pp 213-255.

Tenopri, C.(1984), *Retrieval Performance in a Full Text Journal Article Database.* Doctoral Dissertation. Urbana, IL: University of Ilinois.

Van Rijsbergen, C.J. (1979), *Information Retrieval.* 2nd ed. London: Butterworths.

White, H.D. and Griffith, B.C.(1987), Quality Indexing in Online Databases. *Information Processing & Management.* 23(3), 221-224.

White, H.D. et al.(1984), *Evaluation of the National Library of Medicines Programs in the Medical Behavioral Sciences. Quality of Indexing: the Development of and Testing of a Behavioral Science Literature.* Report to the NLM. Study 3. Philadelphia, PA: Drexel University.

ABSTRACT

Methodological Problems in Information Retrieval Research

Lee, Myeong-Hee*

A major problem for information retrieval research in the past three decades has been methodology, even though some progress has been made in obtaining useful results from methodologically sound experiments. Within a methodology, potential problems include artificial data generated by the researcher, small sample size interpretation of findings. Critics have pointed out that some room exists for improving methodology of information retrieval research; using existing data, having big enough sample size, including large numbers of search queries, introducing more control in relation to variables, utilizing more appropriate performance measures, conducting tests carefully and evaluating findings properly.

Relevance judgments depend entirely on the perception of the user and on the situation of the moment. In an experiment, the best judge of relevance is a user with a well defined information need. Normally more than two categories for relevance judgments are desirable becase there are degrees of relevance.

In experimental design, careful control of variables is meeded for internal validity. When no single detabase exists for comparison, existing operational databases should be used cautiously, Careful control for the variations of search queries, inter-searcher sonsistency, intra-searcher consistency and search strategies is necessary.

Parametric statistics requiring rigid assumptions are not appropriate in information retrieval research and non-parametric statistics requiring few assumptions are necessary. Particularly, the sign test and the Wilcoxon test are good alternatives.

* Part-time lecture, Korea School of Library Services