적합 판단 영향 요인에 관한 이론적 고찰*

A Theoretical Review of Relevance Judgments

유재옥(Jae-Ok Yoo)**

목 차

초 록

정보검색시스템평가에서 적합 판단(relevance judgments)은 중요한 역할을 수행한다. 그 이유는 적합 판단의 결과에 따라 탐색성능이 결정되기 때문이다. 본 논문은 적합 판단과 관련하여 적합개념의 이론적 배경을 역사적으로 고찰하고 적합개념이 적합 판단에 미친 영향을 살펴보고자 한다. 또한 적합 판단과정에서 판단에 영향을 주는 요인이 있는지를 조사하고자 한다. 선행연구논문들에 관한 문헌조사연구 결과는 적합 판단에 영향을 미치는 변인으로 다음의 네 종류의 변인이 파악되었다. 즉, 적합 판단근거로 제시되는 문헌대용물의 종류, 제시순서, 적합측정도구, 판단자 변인의 네 변인들은 적합 판단에 영향을 미치는 것으로 확인되었다.

ABSTRACT

Relevance judgments play a very important role in evaluation of information systems since the degree of success of the information retrieval depends on the relevance judgments. This article reviews the theoretical background of the concept of 'relevance' associated with information retrieval evaluation and tries to identify whether there is any factor that affects relevance judgments. By reviewing previous researches done in the information retrieval evaluation field, four variables have been identified as impacting factors, such as document surrogates presented to judges, the order of presentation, measuring devices of relevance judgments and judges.

* 본 연구는 덕성여자대학교 1995년도 연구비 지원에 의함
** 덕성여자대학교 도서관학과 부교수
■ 논문접수일 : 1996년 11월 6일

1. 서 론

정보검색시스템의 성능을 평가한 최초의 실험연구에 1966년 크랜필드(Cleverdon 1966) 연구를 들 수 있다. 크랜필드 연구방법은 과거 30여 년간 정보검색분야 연구의 실험모델이 되고 있는데 다음과 같다. 먼저 실험용 데이터베이스를 구축하고, 실제의 탐색용 질문을 수집한 후, 각 질문에 해당하는 적합문헌을 대학원생들로 하여금 판단하도록 하였다. 즉, 실험이 시작되기 전에 이미 탐색문제와 적합문헌의 관계가 결정된 것이다.

따라서 정보검색시스템의 성능은 적합자료와 부적합자료의 탐색 비율로 측정할 수 있다고 보는 것이다. 즉, 특정 질문에 해당하는 적합자료가 얼마나 탐색되었나 하는 재현율(recall)과 탐색된 자료 중에서 적합자료의 비율이 얼마나 되는지를 측정하는 정확률(precision)이 그 예이다.

이와 같이 정보검색 분야의 연구에서 정보검색시스템의 성능을 평가할 때 지금까지 관행적으로 사용하는 공통의 평가지수인 정확률과 재현율을 사용하기 위해서는 탐색된 자료가 과연 탐색질문을 만족시키는 적합자료인가를 판단해야 하는 문제와 요청한 탐색질문에 해당하는 적합자료는 데이터베이스 내에 어느 정도 존재하는가 하는 적합자료의 양을 측정해야 하는 업무가 선행되어야 한다. 이 두 작업은 공히 특정 문헌은 특정 탐색 질문에 적합한 자료인가 아닌가 하는 판단을 요구하고 있다. 이와 같이 정보검색시스템의 성능평가에 검색된 문헌에 대한 적합 판단이 필요한 이유는 정보검색시스템의 궁극적인 목적이 이용자에게 적합문헌을 제공하는 것이라면 탐색된 문헌은 실제로 탐색질문을 만족시키는 자료이어야 한다는 것이다. 따라서 문헌에 대한 적합 판단은 정보검색시스템 평가에서 중요한 역할을 한다. 그 이유는 탐색성능 평가는 이 적합 판단에 의해 영향을 받기 때문이다.

이와 같이 정보검색시스템의 성능을 평가하기 위해서는 적합 판단은 불가피한 실정이다.

1.1 연구 목적

정보검색실험에서 적합 판단이 수행하는 기능의 중요성에 대해서는 더 이상 강조할 필요조차 없을 정도로 중요한 역할을 하고 있는 것이 사실이다. 그럼에도 불구하고 적합 판단에 관한 비판적 고찰과 의문은 여러 저자들에 의해 끊임없이 제기되고 있다(Doyle 1963 ; Cooper 1973 ; Ellis 1984 ; Meadow 1985). 이러한 의문점은 실증적으로 조사되기도 했는데 Swanson(1971)은 크랜필드 연구에서의 판단오류를 지적한 바 있으며 Harter(1971)는 이를 실제로 증명하기도 하였다.

크랜필드 연구에서 문제점으로 지적된 바 있었던 적합 판단 오류와 관련된 이러한 문제점은 최근의 연구에서도 그대로 노정되고 있다.

Fenichel(1979)과 Bellardo(1984)는 ERIC ONTAP화일을 이용하여 실험연구를 수행하였는데 두 연구가 우연하게 동일한 탐색문제를 사용하였다. 그러나 해당 문제의 적합문헌수에 대한 견해는 연구 초기에 서로 일치하지 않았다. 이는 동일한 문헌일지라도 적합문헌인지 아닌지에 관한 견해가 연구에 따

라 달라질 수 있다는 개연성을 드러내고 있다고 하겠다. 따라서 어떤 기준하에 적합/부적합 판단을 내리며 누가 문헌의 적합 여부를 판단할 것인지 등에 관한 문제는 많은 논란의 대상이 되고 있다. 더욱이 적합 여부라는 측정도구의 타당성과 신뢰성에 관해서도 의문이 제기되고 있는 실정이다(Ellis 1984).

이와 같은 맥락에서 본 논문은 정보검색분야의 연구에서 중요한 역할을 담당하고 있는 적합 판단(relevance judgment)에 관한 이론적 배경을 고찰하고자 한다. 연구자들이 적합 판단을 실험연구에 기용할 때 파생될 수도 있는 문제점을 인식하고 과학적인 적합 판단 방법을 연구에 도입함으로써 검색성능평가를 보다 과학적으로 수행할 수 있도록 관련 연구들을 개관하여 이론적 배경을 소개하고자 하는 것이 본 논문의 목적이다.

1.2 연구문제 및 연구방법

본 연구는 다음과 같은 연구문제를 살펴보고자 한다.

1) 적합 판단의 기초가 되는 적합개념에 관한 이론적 틀을 역사적으로 살펴보고 각 개념의 논점을 고찰한다.

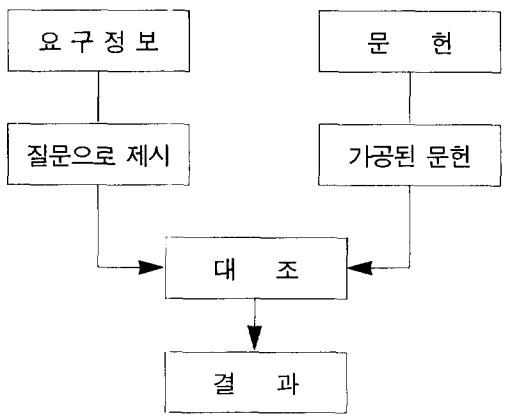
2) 적합 판단에 영향을 미치는 변인이 있는지를 살펴보고 있다면 영향의 정도를 살펴본다.

3) 적합 판단차이가 검색성능 결과에 미치는 영향 정도를 조사한다.

이와 같은 연구문제를 살펴보기 위해 본 논문은 적합개념 및 적합 판단에 관해 수행된 이론연구, 조사연구, 실험연구 등의 선행연구들을 수집하여 고찰하는 문헌연구방법으로 관련 문제를 파악하고자 한다.

2. 적합개념

적합(relevance)이란 무엇인가? 탐색질문에 해당하는 적합자료의 기준은 무엇인가? 탐색질문과 탐색된 문헌의 관련 여부를 규정함에 있어 적합 판단이라는 인간의 주관적인 판단이 개입되는 경우 측정의 신뢰성을 보장할 수 없다는 지적은 주목할 만하다. 정보검색시스템은 간단한 기계처럼 작동하도록 설계된 대로 움직일 뿐인데 여기에 왜 적합 판단이라는 인간의 개입이 필요한지에 대한 의문인 것이다. 시스템의 입장에서 본다면 탐색되지 않은 문헌은 이용자의 탐색질문에 해당되지 않기 때문에 탐색되지 않았으며 따라서 부적합 문헌으로 간주되는 것은 당연하다는 견해이다(Saracevic 1975).



〈그림 1〉 적합개념
(Cuadra and Katter 1967b)

Cuadra(1967b)가 제시하는 적합개념은 〈그림 1〉에서 보는 바와 같다. 이용자의 요구정보는 탐색질문으로 표출되고 문헌은 가공되면서 색인어가 부여된다. 적합이라는 개념은 질문으로 제시된 용어와 문헌이 가공되면서 부여된 색인어 간의 대조를 통해 적합과 부적합이 결정된다. 그렇다면 탐색질문과 탐색된 문헌이 적합한지 혹은 그렇지 않은지의 여부는 시스템의 대조기능에 의해 가장 현저하게 영향을 받는다는 의미가 되지 않겠는가? 따라서 검색시스템의 대조기능을 극대화할 수 있도록 색인작업, 색인어, 화일조직 등 문헌의 가공작업을 보다 효율화시킴으로써 부적합 문헌의 탐색을 최소화하여 검색시스템의 본래의 목적을 달성해야 한다는 주장이다. 바꾸어 말한다면 검색시스템이 수행하는 소정의 적합판단을 인간이 개입하여 내리는 경우 타당성이 결여될 수밖에 없다고 판단하는 입장이다.

이러한 비판 가운데는 '적합' 개념에 대한 정의가 이미 전제되어 있다고 하겠다. 바꾸어 말한다면 탐색질문의 용어와 문헌에 나타난 용어의 대조(match)가 적합이라는 의미로 사용되고 있음을 알 수 있다. 탐색된 문헌이 탐색질문에 적합한 문헌인지 혹은 부적합한 문헌인지의 여부가 판별되는 '적합'의 의미와 관련하여 적합개념에 대한 정의를 먼저 살펴보기로 한다.

2.1 객관적 적합

문헌의 적합에 관한 정의 중에서 일반적으로 가장 널리 알려진 정의는 크게 두 종류로 나눌 수 있겠는데, 객관적 내지는 시스템 중심의 적합개념과 주관적 내지는 이용자 중심의 적합개념으로 대별된다. 객관적 적합이란 일반적으로 특정문헌과 특정질문이 주제나 분야에 관한 한, 서로 '관련 있음'을 의미한다. 다시 말하면 탐색된 문헌의 주제와 탐색질문이 서로 주제적으로 부합할 경우 탐색질문과 탐색된 문헌이 서로 적합하다고 간주하는 것이다(Eisenberg and Schamber 1988). Eisenberg and Schamber는 이를 주제적 적합(topical relevance)이라 부르는데 탐색된 문헌의 주제와 요청한 질문의 주제가 서로 부합하는가를 측정하고자 하는 측면이다. 크랜필드 연구에서 사용된 적합은 '주제적으로 적합하다는' 의미로 사용되었음을 알 수 있다.

특정문헌이 특정탐색질문에 대해 객관적으로 적합하다고 보는 것은 문헌이 탐색질문의 주제를 다루고 있기 때문이라고 보는 것이다. 따라서 특정질문에 적합한 문헌인지 아닌지에 관한 판단은 전문가에 의해 행해질 수 있다고 보는 입장이다. 객관적 적합을 측정함에 있어 주제는 가장 명확하고 보편적인 문헌의 적합성을 대변한다고 가정하는 것이다.

요약하면 객관적인 적합이란 탐색질문을 객관화시킨 명문화된 질문과 탐색된 문헌과의 상호 주제관련성을 의미한다고 상정하는 것이다(Foskett 1972 ; Swanson 1986).

2.2 주관적 적합

반면에 주관적 적합(subjective relevance)은 이용자 관점에서 판단하는 문헌의 적합이다(Cooper 1971 ; Swanson 1986).

주관적 적합은 이용자의 입장을 고려하는 적합개념으로(Swanson 1986) 여러 용어로 설명되기도 하는데 대표적인 것으로는 적절성(Foskett 1972 ; kemp 1974), 상황적 적합성(Wilson 1973), 유용성(Cooper 1973), 정보제공성(Boyce 1982), 유익성(Buckland 1983) 등의 의미를 담고 있다.

적절성(pertinence)이란 이용자가 유용한 것으로 판단하는 특별한 상황에 맞는 적합한 문헌이라는 정의이다(Kemp 1974). 개인의 상황에 따라 정보를 평가하는 관점이 다를 수 있으므로 개인의 필요를 충족시킬 경우 그 정보는 주관적으로 적합하다고 판단하는 것이다(Wilson 1973). 주관적 적합의 범주에서는 이용자의 주관적 판단이 개입된다. 따라서 주관적인 적합은 이용자가 제시한 탐색질문에 대한 탐색결과가 이용자 자신의 필요에 도움이 되는지에 관한 필요개념으로 간주할 수 있다(Foskett 1972).

주관적 적합을 지지하는 연구자들은 이용자에게 적합문헌을 제공하는 것이 정보검색시스템의 목적이라면 검색성능평가는 최종이용자에 의한 적합 판단으로 결정지어져야 한다(Barhydt 1967, 149)고 주장한다. 또한 주제전문가는 적합 판단시 탐색질문과 문헌이 주제적으로 서로 적합한지의 관점에서만 보기 때문에 주제정보를 얻을 수 있는 색인이나 초록 등을 보고 평가하는 경향이 있으며 따라서 정보를 필요로 하는 최종 이용자가 내리는 적합 판단과는 상당한 거리가 있으므로 판단은 최종이용자에 의해 이루어져야 한다는 주장이다(Meadow 1985).

요약하면 객관적 적합이 탐색질문과 연관시켜 판단할 문제라고 한다면 주관적 적합은 정보요구를 가지고 있는 최종이용자에 의해 판단되어야 할 문제라고 보는 견해이다.

2.3 심리적 적합

주관적 적합개념에 대한 신뢰도에 의문을 던지는 측면에서는 주관적 적합개념은 이용자 측면에서 탐색문헌의 적합여부를 검토하는 것임에는 분명하지만 이를 뒷받침할 수 있는 이론적인 배경이 부족하다고 비판하고 있다. 이용자가 적합하다고 선언하면 적합한 문헌이 된다는 것은 주관적 적합개념의 모호성을 단적으로 표현한다는 지적이다. 어떤 개인이 특정 정보를 적합하다고 지각하는 범위는 그 개인의 인식이지 그 문헌이 대변하는 정보는 아니라고 간주하기 때문이다(Schamber, Eisenberg and Nilan 1990). 왜냐하면 이용자들은 현재 자신이 가지고 있는 문제해결에 관한 자료를 원하기 때문에 즉, 심리적으로 적합한 자료를 찾고 있기 때문에 이용자들의 현재의 인지상태(cognitive state)에 영향을 미칠 수 있는 자료만이 이용자의 요구를 만족시킨다는 것이다(Harter 1992).

이와 같은 맥락에서 심리적 적합(psychological relevance)으로 적합개념을 정의하고자 하는 시도(Sperber and Wilson 1986 ; Harter 1992)에 유의할 필요가 있다.

이용자의 정보요구는 정적이고 변하지 않는 고정된 상태로 이해되고 있는 것이 일반적인 경향이지만 Schamber, Eisenberg and Nilan(1990)과 Katzer and Snyder(1990)는 개인이 가지고 있는 정보요구를 일종의 정신

적인 상태(mental state)로 규정하고 이 정신적인 상태는 탐색된 적합문헌을 접하면서 역동적으로 변할 수 있다고 보는 것이다. 따라서 이용자의 정보요구를 정적인 상태로 보기보다는 역동적인 상태로 간주하는 것이 특징이다.

또한 탐색자의 인지 상황이 적합문헌을 발견하는 과정에서 변할 수 있다고 보기 때문에 이용자의 정보요구는 절대적이지 않고 따라서 탐색목적도 각기 다를 것이기 때문에 질문과 문헌의 주제관련성을 측정하는 객관적 적합은 심리적 적합의 입장에서 보면 탐색자의 인지활동에서 발생하는 변화에 역동적으로 대응하지 못하므로 오히려 적합개념이 약하다는 지적이다(Harter 1992).

심리적으로 적합하다고 간주되는 문헌은 보다 극단적으로 표현하면 탐색문제와 주제적으로는 관련있는 문헌일지라도 정보요구를 가진 이용자에게 새로운 지적 자극을 줄 수 없거나 또다른 인지활동을 촉진시킬 수 없다면 이용자에게는 심리적으로 적합한 문헌이 아니라는 것이다.

심리적인 적합개념을 상정한 Harter는 고정적이고 변하지 않는 적합 판단에 의존하여 '적합'을 탐색성능 평가에 사용하는 것은 지양되어야 한다고 주장하고 있다. 적합 판단은 문헌이 읽히는 순간에 개인의 심리적인 상태를 표현하는 기능을 한다고 보기 때문에 지금까지 사용된 정확률 및 재현율 사용이 지양되고 탐색평가 방안이 개선되어야 한다고 제안하고 있다(Harter 1992, 612).

3. 적합 판단에 영향을 미치는 요인

자료의 적합여부를 누가 판단하는지에 따라 동일한 문헌일지라도 적합 판단여부가 달라질 수 있다는 연구결과가 정보검색 관련연구자들의 관심을 끌고 있다. 또한 같은 자료인 경우에도 판단 시점에 따라, 혹은 제시되는 문헌의 종류에 따라 문헌의 적합 판단 여부가 상이하게 나올 수 있다는 실험연구 결과들을 살펴보면 문헌의 적합 판단에 영향을 주는 요인들을 다음과 같이 네 가지로 요약할 수 있다. 판단근거로 제시되는 문헌대용물의 종류, 제시순서, 적합 측정도구, 판단자 변인으로 각각의 변인들이 적합 판단에 미치는 영향 및 그 정도를 관련 논문들의 연구결과를 요약하면 다음과 같다.

3.1 문헌대용물(document surrogates)의 종류

판단자들은 적합 판단시 주어진 문헌대용물의 어떤 부분을 가장 많이 참고하여 적합 판단을 내리는가, 참고한 문헌대용물의 종류가 다를 때 적합 판단에 차이를 보일 것인가, 또한 어떤 부분이 가장 정확한 적합 판단을 이끌어 낼 것인가 하는 등의 문제는 이 분야의 관심의 대상이 되어 왔다.

일반적으로 판단자들은 표제(title)를 가장 많이 참고하는 것으로 나타났으나(Saracevic 1969 ; Shaw 1995) 표제에 의한 적합 판단의 정확성은 낮은 것으로 나타나 표제에 의한 판단은 주의가 요망되는 것으로 관찰되었다(Marcus et al. 1978).

한편 판단자들은 적합 판단의 근거인 문헌대용물의 어떤 부분을 보고 적합 판단을 내리는가에 따라 판단 차이가 심하게 나타난다는 사실이 발견되었다(Saracevic 1969 ; Janes 1991 ; Janes and McKinney 1992).

적합 판단에 가장 많은 영향을 주는 문헌대용물의 형태는 초록으로 지적되었는데 초록은 적합 판단에 대한 예측력도 가장 높은 것으로 나타나 초록을 보고 내린 적합 판단이 가장 정확한 것으로 여러 연구에서 공통적으로 조사된 바 있다(Marcus 1978 ; Burgin 1992).

선호하는 문헌대용물의 종류

Saracevic(1969)은 22명의 이용자들을 대상으로 실제의 질문 99개에 대한 적합 판단을 내릴 때 판단자들이 주로 참고하는 문헌대용물의 종류를 살펴보았다. 표제, 초록, 전문(全文)을 제시하고 적합 판단을 내리도록 요청한 후 그 결과를 분석하였다. 판단자들은 모두 표제를 먼저 본 후, 초록을 보고 그 후 전문을 보는 것으로 나타났다. 판단자들은 문헌의 부적합 판단을 짧은 길이의 표제필드를 보고 신속히 결정하는 경향인 것으로 나타나 표제가 판단의 근거로 선호되고 있는 것으로 알려져 있다.

Shaw(1995)는 언어학과 문학을 전공하는 대학원생의 CD-ROM 탐색결과에 대한 적합 판단 과정을 살펴보았다. 판단자들은 서지 레코드에서 표제와 색인어에 기준한 적합 판단을 가장 많이 내리는 것으로 관찰되었다.

그러나 표제만으로 적합 판단을 내리는 경우에는 조심스러운 접근이 필요하다는 지적이다. 각 필드별로 적합 판단의 정확성을 조사한 Marcuse(1978)에 의하면 표제필드는 0.637, 주제필드는 0.672, 초록필드는 0.73으로 표제가 가장 정확성이 낮은 필드로 확인되었다.

판단 갈등

한편 판단자들이 문헌대용물의 어떤 부분－표제, 초록, 전문－을 보고 적합 판단을 내렸는가에 따라 같은 문헌이라도 극심한 판단 갈등을 노정한다는 사실이 발견되었다.

Saracevic(1969)에 의하면 표제, 초록, 전문을 주고 적합 판단을 내리도록 하였는데 표제를 보고 판단한 판단자의 15%가 전문을 본 후 적합 판단을 바꾸었으며 초록을 보고 판단한 판단자의 10%가 전문을 보고 판단결과를 바꾸었다. 전체적으로 22%의 판단이 바뀐 사실을 미루어 볼 때 적합 판단의 근거가 되는 문헌대용물의 종류에 따라 적합 판단이 크게 영향을 받는 것으로 보고하고 있다. 판단자들이 문헌대용물의 어떤 부분을 보고 적합 판단을 내렸는지에 따라 판단 차이가 나타날 수 있음을 보여주는 연구이다.

Janes(1991)와 Janes and McKinney (1992) 등은 제시되는 문헌의 대용물이 주는 정보에 따라 적합 판단은 바뀐다는 흥미있는 사실을 발견하였다.

Janes는 미시간 대학의 교수와 박사과정 학생들이 제기한 39종의 실제 탐색질문을 탐색한 결과를 탐색문제를 제기한 미시간 대학의 교수와 박사과정 학생들에게 다음과 같은 순서로 제시하고 적합 판단을 요청하였다. 처음에는 표제만, 다음에는 표제와 초록을, 다음에는 표제, 초록, 색인어가 있는 세 부분을

제시하고 적합 판단이 문헌의 어떤 부분에 의해 가장 많은 영향을 받는지를 고찰하였다.

적합 판단은 초록에 의해 가장 많은 영향을 받는 것으로 나타났다. 681건의 문헌 중에서 표제에만 의존하여 내린 적합 판단이 추가정보에 의해서도 변하지 않은 문헌은 165건으로 24.2%에 불과했다. 적합 판단의 28.3%는 색인어에 의해 판단이 바뀌었으며, 29.0%는 서지사항이 추가됨에 따라 판단이 바뀌었으며, 68.9%는 초록을 참고한 후에 적합판정이 바뀜으로써 초록이 가장 영향력있는 적합 판단의 기준이 됨을 시사하였다.

Janes and McKinney는 1991년에 수행된 Janes의 연구에서 일부자료를 다시 추출하여 1991년의 연구방법과 동일한 방법으로 분석하였는데 초록에 의한 적합 판단의 변화율은 68.9%(1991)보다 높은 77.4%로 나타나 초록이 적합 판단에 미치는 영향력을 말해주고 있다. 표제에만 의존한 적합 판단 중 변하지 않은 판단은 13.5%로 1991년의 24.2%보다 낮아짐으로써 표제에 의한 적합 판단의 신뢰도가 더욱 낮아졌음을 볼 수 있다.

예측력

적합 판단에 가장 영향력을 행사하는 문헌대용물의 종류는 초록으로 나타났으며 초록은 가장 정확한 적합 판단을 이끌어 내고 있음이 여러 연구에 의해 확인되었다. Marcus et al.(1978)은 판단자들이 적합 판단을 내릴 때 표제, 초록, 색인어, 키워드필드만을 참고로 한 경우와 전문을 보고 적합 판단을 내린 경우와 비교하여 각 필드는 어느 정도의 예측력을 갖는가를 조사하였다. 표제필드는 0.637, 주제필드 0.672, 초록은 0.73으로 표제필드가 가장 정확성이 낮은 것으로 나타났으며 초록 필드가 가장 높은 예측력을 갖는 것으로 나타나 초록에 의한 판단이 가장 정확한 것으로 나타났다.

또한 문헌대용물을 복합적으로 제시하는 것은 적합 판단을 보다 효과적으로 이끌어 내는데 기여하는 것으로 조사되었다. Burgin(1992)은 적합 판단시 문헌의 대용물이 복합적으로 제시될 때 판단의 정확성에 차이가 있는지를 조사하였다. 색인어, 표제, 초록, 표제+초록, 표제+초록+색인어별로 제시했을 때 적합 판단의 정확성은 .17, .26, .32, .33, .36으로 문헌에 관한 정보가 추가되면서 적합 판단의 정확성이 증진된 것으로 나타났다.

3.2 제시 순서

문헌이 판단자에게 제시되는 순서에 의해서도 적합판정은 영향을 받는 것으로 밝혀졌다.

Eisenberg and Barry(1988)는 Syracuse 대학의 학부학생과 대학원생을 판단자로 기용하여 탐색문제와 함께 탐색된 15개의 문헌에 관한 정보(document description)를 제시한 후 적합판정을 요청하였다. 연구자들은 15개의 문헌을 적합문헌의 정도에 따라 적합도가 낮은 문헌으로부터 높은 문헌의 순서로 배열한 것과 적합도가 높은 문헌에서 낮은 문헌의 순서로 배열한 두 종류로 나누어 판단자에게 제시하였다. 한 집단의 판단자는 문헌이 제시되는 순서대로 한 문헌씩 검토하여 적합정도를 표시하되 등급별(1-7)로 평가점수를 주고 또 한 집단은 크기예측(magnitude estima-

tion) 점수로 적합 정도를 평가하도록 설계한 실험이었다.

문헌이 제시되는 순서에 따라 적합 판단은 심하게 영향을 받는 것으로 보고되고 있다. 적합도가 높은 문헌에서 낮은 문헌의 순서로 문헌이 판단자에게 제시될 때 판단자들은 적합도가 높은 문헌일지라도 낮게 평가하는 경향을 드러내는 반면에 적합도가 낮은 문헌에서 높은 문헌의 순서로 제시될 때 판단자들은 중간 정도의 적합문헌을 높게 평가하는 경향을 보였다. 이러한 현상은 특히 등급척도 평가에서 현저하게 드러났는데 판단자들은 첫번째로 평가하는 문헌을 아주 높게 혹은 아주 낮게 적합 정도를 주는 것을 꺼리는 것으로 보인다고 연구자들은 설명하고 있다. 판단자들은 첫번째 평가대상 문헌에 2~3점을 줌으로써 다음의 평가대상 문헌을 이에 기준하여 높거나 낮은 점수를 주는 경향이었다고 보고하고 있다.

적합도가 낮은 문헌에서 높은 문헌의 순서로 판단자에게 제시될 때 판단자는 첫번째의 평가대상 문헌을 기준점으로 간주하며 적합도가 높아짐에 따라 점수가 계속 높게 부과된 것으로 나타났다. 반면에 적합도가 높은 문헌에서 낮은 문헌의 순서로 제시될 때는 적합도가 높은 문헌이 낮게 평가되는 경향을 연구자들은 확인하였다. 판단자들은 적합도의 정도가 낮아짐에도 불구하고 이를 점수에 반영하지 않는 것으로 나타났다. 즉 적합도가 낮은 문헌이 높게 평가되는 현상이 관찰되었다. 판단자들이 혼란에 빠졌거나 일관성 있는 점수 부여가 불가능한 것으로 해석되고 있다. 그러나 이와 같은 문헌제시 순서에 따른 적합 판단의 차이는 등급척도 평가방법으로 측정했을 때만 의미있는 차이로 나타났으며 크기예측 점수로 측정했을 때는 차이가 관찰되지 않았다. 이 부분에 대한 설명은 3.3 측정 도구에서 다루고 있다.

이와 같이 문헌이 제시되는 순서에 의한 영향은 있는 것으로 관찰되었으나 문헌의 수가 15개보다 작을 경우는 문헌제시 순서에 영향을 받지 않는 것으로 나타나 이 문제에 관한 한 상반되는 연구결과가 주목의 대상이 되고 있다(Purgailis Parker and Johnson 1990).

3.3 측정 도구

정보검색분야에서 사용한 적합성 측정도구는 등급척도(category rating)로서 일반적으로 두 종류를 사용해 왔다. 매우 적합, 적합, 보통, 부적합, 매우 부적합의 5등급 척도와 적합, 부적합의 2등급 척도이다.

이 등급 척도는 측정한 적합데이터가 등간척도로 변환될 뿐만 아니라 상당한 문맥편견효과(contextual biasing effect)가 발생한다는 점이 단점으로 지적된 바 있다(Rath, Resnick and Savage 1961). 더욱이 적합, 부적합의 2등급으로 측정하는 방법은 더욱 부적절하다는 지적이다(Eisenberg 1988, 374). Gluck(1996)에 의하면 5등급의 값 혹은 2등급으로 측정한 적합 판단 측정치에 따라 적합 판단은 상당한 영향을 받는 것으로 나타났다(gamma 0.74-0.89). Cuadra and Katter (1967)도 등급척도에서 등급의 수는 적합평가에 영향을 미친다고 보고한 바 있다.

한편 크기예측 측정 도구는 Stevens(1966)에 의해 고안된 방법으로 측정자가 직접 주어진 자극에 대해 예측한 값을 임의의 숫자로 부여하는 방법이다. 영(zero)이나 마이너스(-) 숫자만을 제외한 어떠한 숫자도 임의로 사용할 수 있다. 주어진 자극의 정도를 숫자로 표현하도록 하는 것이다. 각자가 임의로 값을 부여하기 때문에 측정자에 따른 차이는 있지만 일반적으로 측정자들은 일관성 있는 값을 부여하는 것으로 관찰되었다(Zwislocki 1983). 크기예측 측정 도구는 Rees and Schultz(1967)와 Cuadra and Katter (1967)에 의해 적합 측정 도구의 대안으로 제안된 바 있다.

한편 적합 판단 도구로 관행적으로 사용되어 온 바 있는 등급척도와 크기예측 측정 도구를 비교한 실험연구 결과가 Eisenberg and Barry(1988)와 Eisenberg(1988)에 의해 보고된 바 있다. Eisenberg and Barry는 15개의 문헌을 적합 정도에 따라 적합도가 낮은 문헌으로부터 높은 순서로, 높은 문헌에서 낮은 순서로 배열한 두 종류를 제시한 후 등급별(1-7) 점수와 크기예측 점수로 적합 정도를 평가한 결과를 비교하였다.

문헌이 제시되는 순서에 따라 적합 판단은 영향을 받는 것으로 보고되었다. 그러나 이러한 현상은 특히 등급척도에서만 현저하게 드러났는데 문헌제시 순서에 따른 적합 판단의 차이는 등급척도로 측정했을 때는 차이를 보였으나 크기 예측 점수로 측정했을 때는 차이가 관찰되지 않았다. 연구자는 크기 예측 점수가 문헌의 제시 순서에 의한 판단 차이의 영향력(편견)에서 자유롭다는 증거로서 등급척도보다 신뢰성있는 측정 도구라고 주장하고 있다.

또한 Eisenberg는 크기예측 측정방법이 문헌의 적합 여부를 측정하는 적절한 측정 도구가 될 수 있을 것인가를 등급척도 방법과 비교하였으며 두 측정방법이 실제로 문맥효과로부터 자유로운지를 조사하고자 하였다. 80명의 피험자를 선택하여 한 개의 특정 탐색질문과 탐색된 문헌 15건을 피험자에게 제시하고 각 문헌에 대한 적합 판단을 하되 7등급평가와 크기예측 방법으로 적합 정도를 표시하도록 요청한 후 두 측정방법을 비교하였다.

피험자들은 등급평가 측정방법을 선호하는 경향을 보였으나 크기예측 측정방법에 쉽게 적응하는 것을 관찰할 수 있었다고 연구자는 보고하고 있다. 등급평가로 측정된 적합도 수치는 문헌이 제시되는 순서나 측정도구에서 등급정도가 배열되는 순서에 따라 영향을 받는 것으로 나타났으나(alpha = .05) 크기예측 측정점수는 척도배열 순서나 문헌 제시 순서 등의 영향으로부터 자유로운 것으로 나타나 크기예측 측정 도구가 등급평가 측정 도구보다 신뢰성 있는 측정 도구라고 연구자는 주장하고 있다. 한편 어떠한 적합 판단 측정치를 사용하든지 간에 특정문헌에 대한 적합 판단은 다른 문헌에 의한 문맥효과의 영향을 받는 것으로 나타났다.

3.4 판단자

적합 판단에 가장 중요한 영향을 주는 요인은 판단자변인으로 알려져 있다. 이러한 판단자들이 적합 판단을 내릴 때 과연 얼마나 판

단의 일관성을 유지할 수 있을 것인지가 가장 우선적으로 관심의 초점이 되고 있다.

판단의 일관성

Resnick and Savage(1964)는 판단자들이 적합 판단시 일관성을 보이는지를 파악하기 위해 46명의 IBM 종사자들에게 IBM의 내부 자료들을 서지사항만, 서지사항과 초록만, 혹은 서지사항과 색인어만을 제시하고 적합 판단을 내리도록 요청하였다. 한 달 후 동일한 조사를 재수행했을 때 10%의 판단만이 바뀐 것으로 나타나 판단자의 적합 판단은 상당히 일관성이 있는 것으로 보고되었다.

판단자 간의 불일치

판단자의 판단의 일관성은 유지되는 것으로 보인다. 그러나 그렇다 하더라도 동일문헌에 대한 적합 판단이 판단하는 사람에 따라 달라지는 적합 판단의 불일치는 문제점으로 남는다.

Barhydt and Gordon(1967)은 최종이용자의 적합 판단과 제삼자 간의 판단의 차이를 감수성, 특정성, 효율성의 수치로 비교하여 판단 일치율을 계산하였다. 교육분야에 종사하는 연구자를 최종이용자로 간주하고 주제전문가와 시스템전문가를 제삼자의 판단자로 간주하였다. 적합 판단을 탐색된 문헌의 초록을 보고 적합, 부적합의 2등급으로 나누어 판단한 결과를 분석하여 감수성(sensitivity), 특정성(specificity) 그리고 효율성(effectiveness) 점수로 계산하여 판단자 간의 적합 판단 일치율을 비교하였다.

감수성은 이용자와 제삼자의 판단자의 적합 판단이 일치하는 문헌의 수를 이용자가 적합하다고 판단한 문헌의 수로 나눈 값이다. 특정성은 이용자와 제삼자의 판단이 일치하는 부적합문헌의 수를 이용자가 판단하는 부적합 문헌수로 나눈 값이다. 효율성은 감수성의 값과 특정성의 값을 합한 수치에서 1을 뺀 값으로서 이용자와 제삼의 판단자의 적합 판단이 완전일치할 경우에는 감수성=1, 특정성=1, 효율성=+1의 값이 된다. 반면에 이용자와 제삼자의 판단이 완전 불일치할 경우에는 감수성=0, 특정성=0, 효율성=-1의 값이 된다. 감수성, 특정성, 효율성의 세 값은 제삼자인 판단자의 적합 판단이 얼마나 최종이용자의 판단과 일치하는가를 나타내는 수치라고 연구자들은 정의하고 있다.

Barhydt and Gordon에 의하면 주제전문가의 효율성 값은 평균 .34(-.74 ~ .98)이며 시스템전문가는 평균 .35(-.20 ~ .96)를 기록하여 적합 판단의 일치율이 매우 낮음을 보고하면서 정보검색시스템이 이용자에게 적합 문헌을 제공하는 것이 목적이라면 검색성능 평가는 최종이용자에 의해 문헌의 적합 판단이 내려져야 한다고 주장하고 있다(p. 149).

Lesk and Salton(1968)도 탐색질문을 제기한 최종 이용자와 제삼자가 내린 적합 판단에서 30%의 저조한 판단일치율을 일찍이 고찰한 바 있다. 그 후의 연구들이 연구 결과마다 차이는 있지만 판단자 간에 일치율이 최고 85%까지 이르는 등 편차가 심한 것으로 보고되고 있다.

Janes and McKinney(1992)는 1991년의 Janes의 연구에서 수행한 최종이용자의 적합 판단 데이터와 1992년에 수행한 제삼자인 판

단자의 적합 판단 데이터를 사용하여 양자 간의 적합 판단 행태를 비교하였으며 동시에 적합 판단 일치율에 차이가 있는지를 조사하였다.

실제로 정보를 필요로 하는 이용자들은 판단시 보수적인 태도를 취하였다는 사실이 발견되었는데 표제를 보고 내린 적합 판단이 추가정보에 의해서도 변하지 않은 비율이 최종이용자는 24.2%로서 제삼자인 판단자의 13.5%에 비하면 상당히 높은 것으로 보고되고 있다. 이와 함께 문헌 한 건당 적합판정을 내리는 비율이 제삼자인 판단자가 17.8%임에 비해 이용자는 16.1%로 낮게 나타나 보다 엄격한 적합 판단을 내리는 것으로 관찰되었다.

연구자들에 의하면 적합 판단을 위한 정보가 추가됨에 따라 적합 판단에 도움이 된다고 응답한 이용자는 71.8%임에 비해 제삼자인 판단자는 73.7%로서 제삼자인 판단자는 추가정보를 보다 유용하게 간주하는 것으로 나타났다. 적합 판단시 기준이 되는 추가정보의 종류가 이용자와 제삼의 판단자에 따라 다름을 발견하였는데 이용자는 제삼자보다 초록, 표제, 기타 서지정보 순으로 중요하게 생각하는 반면에, 제삼자는 색인정보를 이용자보다 더 중요하게 생각하는 경향을 밝혔다. 제삼자인 판단자의 적합 판단과 이용자의 적합 판단의 일치율은 67.3%로서 양자간의 판단일치율은 비교적 높은 편으로 보고되었다.

그러나 SMART시스템을 이용하여 Medline 데이터베이스를 실험대상으로 한 Burgin(1992)의 연구에서도 Lesk and Salton (1968)의 결과처럼 최종이용자와 제삼자 간의 판단 일치율은 36.0%로서 역시 저조하였다. 이러한 판단 일치율의 차이는 판단자 집단 사이에 나타날 뿐만 아니라 동일집단 내에서도 판단자집단의 성격에 따라 판단 차이가 있다는 사실이 관련 연구들에 의해 확인되었다. 판단자 집단간의 판단일치율을 보면 주제전문가들 사이의 판단일치율은 평균 .34(Barhydt 1967) 혹은 .55~.75(Saracevic 1975)로 알려진 바 있으며 최종이용자의 판단일치율은 .45~.60(Saracevic 1975)으로 보고된 바 있다. 최종이용자도 아니고 주제전문가도 아닌 제삼의 판단자 간의 일치율은 평균 .35 (Barhydt 1967) .30(Lesk and Salton 1968). .67(Janes and McKinney 1992)로 판단자 간의 일치율은 상당한 편차를 보이는 것으로 알려져 있다. 그러나 일반적으로 탐색문제에 관한 전문지식이 많은 집단 내의 판단일치(.55~.75)는 최종이용자 판단자(.45~.60)들 간의 일치율에 비해 높은 것으로 나타나 주제지식이 많은 판단자들 사이에는 판단에 대한 이견이 적은 것으로 인식되고 있다 (Saracevic 1975).

이와 같이 판단자들의 판단 일치율은 일반적으로 저조한 편이다. 이러한 판단자 간의 판단 차이에 영향을 주는 요인은 무엇인가? 동일문헌에 대한 적합 판단이 판단자 간에 다를 수 있다면 판단자의 어떤 요인이 판단 차이를 이끌어 내는지를 살펴볼 필요가 있다.

판단자 변인

판단자의 다양한 특성이 적합 판단에 영향을 줄 것이라는 가정을 조사한 연구결과를 살펴보면 판단자가 연구자 혹은 학생인가에 따라서 또는 같은 연구자나 학생일지라도 학년

이 높거나 낮은, 연구에 종사한 연도에 따라서도 달라질 수 있으며 같은 문헌에 대한 적합 판단일지라도 판단자의 전공에 따라서도 그 판단은 달라질 수 있다는 가설이 Regazzi (1988)에 의해 제시된 바 있다.

또한 판단자의 주제 배경에 대한 정규교육이나 실제경험이 적합 판단의 차이를 낳는다는 연구결과가 Rees and Schultz(1967)에 의해 보고되고 있다. 또한 탐색질문에 대한 주제지식이 많은 판단자들 사이에는 적합 판단여부에 대한 이견이 적으며 탐색질문에 대한 주제지식이 적은 판단자일수록 적합판정을 많이 내리는 것으로 알려져 있다(Cuadra and Katter 1967 ; Saracevic 1970). 주제전문가들 사이의 판단의 일치율은 .55～.75로 나타난 반면에 최종이용자의 판단일치율은 .45～.60으로 주제전문가에 비해 비교적 낮다는 사실은 주제지식과 적합 판단의 관련성을 지지하고 있는 연구결과라고 하겠다(Saracevic 1975, 341-2). 이 외에도 O'Connor (1969)는 판단자들이 판단 갈등에 관해 의견을 나눌 경우 이미 내린 적합 판단이 변하기도 한다는 것을 고찰하였다.

4. 적합 판단의 차이

선행연구들이 제기한 적합 판단과 관련한 문제점을 고찰할 때 다음의 두 가지 요인으로 요약할 수 있다. 첫째는 연구자들이 추종하는 적합개념은 정보검색 실험에 판단자변인으로 반영됨으로써 적합 판단에 영향을 미치며, 둘째는 실제적인 적합 판단 과정은 다양한 변인에 의해 영향을 받는다는 사실이다. 따라서 적합 판단이 보다 타당성 있고 신뢰성 있게 수행되기 위해서는 과학적인 접근방법이 요청된다고 하겠다.

4.1 판단자 선정의 문제

적합개념은 역사적으로 다양하게 정의되어 왔다. 크게는 객관적 적합과 주관적 적합으로 나뉘어 상정되어 왔으며 최근에는 심리적 적합개념이 대두되고 있다. 연구자들이 어떠한 적합개념을 지지하는가에 따라 연구설계시 기용하는 판단자 타입이 다르게 나타난다. 예를 들면 객관적 적합을 지지하는 연구는 탐색문제와 탐색결과가 주제적으로 일치하는지의 여부가 적합 판단의 관건이 됨으로 판단수행자로 주제전문가를 기용하는 경향이다. 반면에 주관적 적합의 타당성을 따르는 연구는 탐색문제를 제기한 최종이용자를 판단자로 사용하고자 한다.

이 외에도 심리적 적합개념이 최근에 소개된 바 있으나 실제 연구에 적용한 구체적인 연구는 아직까지 나타나지 않고 있다.

연구자들이 지지하는 적합개념에 따라 기용하는 판단자가 달라지고 판단자가 다를 때 나타나는 문제점은 무엇인가? 비록 동일한 정보검색시스템을 평가한다 하더라도 적합 판단을 누가 했는가에 따라 시스템의 검색성능의 평가지수인 정확률과 재현율이 달라질 수 있다는 의미이다. 판단자 간의 판단일치율을 보면 주제전문가들 사이의 판단일치율은 평균 .34 (Barhydt 1967)혹은 .55～.75(Saracevic 1975)로 알려진 바 있으며, 최종이용자의 판

단일치율은 .45～.60(Saracevic 1975)으로 보고된 바 있다. 최종이용자도 아니고 주제전문가도 아닌 제삼의 판단자 간의 일치율은 평균 .35(Barhydt 1967) .30(Lesk and Salton 1968), .67(Janes and McKinney 1992)로 판단자 간의 일치율은 상당한 편차를 보이는 것으로 나타났다. 그러나 일반적으로 탐색문제에 관한 전문지식이 많은 집단 내의 판단일치(.55～.75)는 최종이용자 판단자(.45～.60)들 간의 일치율에 비해 높은 것으로 나타나 주제지식이 많은 판단자들 사이에는 판단에 대한 이견이 적은 것으로 인식되고 있다(Saracevic 1975). 한편 주제지식이 많은 판단자보다는 주제지식이 적은 판단자가 적합판정을 더 많이 내리는 것으로 알려지고 있다(Cuadra and Katter 1967 ; Saracevic 1970).

그렇다면 동일한 정보검색시스템에 관해 수행하는 연구라 하더라도 적합 판단자로 누구를 기용하는가에 따라 탐색성능은 상당한 차이를 보일 수 있다. 더욱이 서로 다른 시스템의 성능을 상호비교할 경우에는 적합 판단자 선정은 성능평가에 민감한 영향 요인이 될 소지가 충분하다. 지금까지의 정보검색에 관한 연구들은 주제전문가, 탐색문제를 제기한 이용자 혹은 주제전문가도 아니고 최종이용자도 아닌 제삼자로 하여금 적합 판단을 내리도록 연구설계를 하는 것이 일반적인 경향이었다. 판단자가 적합 판단에 민감한 영향을 줄 수 있다는 사실이 정보검색 실험설계에 반영되어야 할 것이 지적되고 있다.

4.2 적합 판단 방법의 문제

적합 판단을 내리는 방법론상의 문제점이 선행연구에 의해 확인된 바 있으며 이는 적합 판단에 영향을 주는 요인이기도 하다.

첫째, 판단자들이 적합 판단 근거로 제시되는 문헌대용물의 어떤 부분을 보고 판단을 내렸는가에 따라 같은 문헌이라도 판단 갈등이 야기된다는 지적이다. 비록 판단자들은 판단 근거로 제공받는 정보 중에서 표제를 가장 많이 참고하는 것으로 나타났으나 표제가 의외로 판단의 정확도가 떨어지고 초록이 가장 영향력있는 문헌의 대용물로 나타남에 따라 적합 판단 제공시의 문헌대용물의 종류로 초록이 제안되고 있다. 또한 문헌대용물이 복합적으로 제시될 때 판단의 정확성은 증진되는 것으로 알려졌다.

둘째, 적합 판단 측정에 일반적으로 사용하는 측정도구인 등급척도는 부적절하다는 비판이 제기되어 왔으며 특히 적합과 부적합 중의 하나를 선택하는 2분법적 측정은 더욱 부적절하다는 지적이다. 크기예측 측정 방법이 적합 판단 측정 도구로 제안되고 있으나 실제 연구에서는 사용되지 않고 있는 실정이다.

또한 판단근거로 제시되는 문헌대용물의 제시순서도 영향요인이 될 수 있다는 지적이 제시되었다. 판단자들은 첫번째로 제시되는 문헌을 적합의 기준점으로 설정하고자 하는 경향이 발견되었다.

4.3 적합 판단의 차이가 검색성능 결과에 미치는 영향

적합 판단이 판단자에 따라 다르다면 이러한 판단자 사이의 차이가 과연 정보검색 성능

평가에 영향을 미칠 수 있는 것인지 또는 판단자들의 판단차이는 무시할 수 있을 정도의 차이인지를 조사한 연구들을 살펴보고자 한다.

Lesk and Salton(1968)은 문헌정보학 분야의 1,268건의 초록을 사용하여 네 집단의 판단자들에게 적합 판단을 요청한 결과를 비교하였다. 탐색질문을 제시한 최종이용자, 제삼자인 판단자, 적합 판단이 일치한 이용자와 판단자, 적합하다고 판단한 이용자와 적합하다고 판단한 판단자집단으로 나누어 네 집단의 적합 판단의 차이를 분석하였다. 탐색질문을 제시한 이용자와 제삼자인 판단자 간의 적합 판단 일치율은 30%로서 매우 낮은 일치율을 보였다. 그러나 이러한 적합 판단의 불일치로 인한 차이가 탐색성능을 재현율과 정확률로 산정한 평가 결과에는 영향을 주지 않은 것으로 드러났다. 네 집단의 적합 판단 결과에는 판단 차이가 분명히 존재함에도 불구하고 그 차이는 탐색성능 측정치에는 유의미한 영향을 미치지 않은 것으로 보고되었다.

이에 대해 Lees and Salton(1968, 346)은 "대개의 정보검색 실험에서 재현율이나 정확률을 산출할 때 한 질문에 대해 많은 탐색을 수행한 후 그 결과를 평균치로 측정하는 방법을 쓰고 있기 때문에 산출된 평균값들이 적합 판단의 차이에 대해 민감하게 영향을 주지 않는 것으로 보인다"고 설명하고 있다.

한편 Saracevic(1970, 134)은 Lesk and Salton의 연구결과에 대해서는 검증이 필요하므로 보다 깊이있는 연구가 수행되어져야 할 것이라고 제안하고 있다. Lesk and Salton은 판단자의 적합 차이만을 사용했을 뿐이므로 기타 변인, 즉 적합 판단에 영향을 미치는 요인으로 알려진 기타 변인들을 연구에 포함시킬 경우에도 같은 결과가 나올 것인지에 대한 연구가 요청된다고 제안하고 있다. 더욱이 적합 판단 측정을 적합, 부적합으로만 나누는 2등급 측정방법을 지양하여 크기예측 방법을 적용했을 때도 같은 연구결과가 나올 것이지에 관해서도 연구해야 할 것을 제시한 바 있다.

Burgin(1992)은 Lesk and Salton과 유사한 연구를 수행하였다. SMART시스템을 이용하여 낭포성 섬유증으로 색인어가 주어진 1,239개의 논문이 저장된 Medline 데이터베이스를 실험대상으로 하였다. 적합 판단을 내리는 판단자를 네 집단으로 나누었다. 낭포성 섬유증연구에 종사하는 교수나 의사로서 탐색질문을 직접 제공한 최종이용자, 질문을 제공한 최종이용자의 동료, 박사후 과정에 있는 동료, 온라인 탐색경험이 많은 의학분야 서지전문가의 네 집단으로 구성되었다. 각 집단이 내린 적합 판단의 정확성은 39.9% ~ 55.0%로서 차이를 보였다. 판단자 집단 간의 일치율도 36.0%에서 63.6%까지 다양함을 보였는데, 질문을 제공한 최종이용자 집단과 의학분야 서지학자 집단 간의 판단일치율은 36.0%로 나타나 가장 낮았다. 반면에 최종이용자 집단과 박사후 과정의 동료집단 간의 일치율은 63.6%로 가장 높아서 주제전문 지식이 높은 판단자들 간에 일치율이 높다는 Cuadra and Katter(1967)와 Rees and Schultz(1967), Saracevic(1970)의 연구결과를 뒷받침하는 것으로 나타났다.

Burgin의 연구도 Lesk and Salton의 연

구결과와 마찬가지로 실험에 사용된 각각의 탐색질문에 대한 평균 정확률 및 재현율에는 차이가 있음에도 불구하고 탐색질문 전체에 대한 평균 정확률이나 평균 재현율에는 네 집단 간 차이가 별로 크지 않은 것으로 드러났다. 따라서 관련 연구들은 판단자 간의 차이는 엄존함에도 불구하고 검색성능평가에 미치는 영향은 통계적으로 의미있는 차이가 되기에는 미미한 차이라고 결론짓고 있다.

5. 결 론

탐색성능평가에서 적합 판단은 불가피한 과정인 것은 주지의 사실이다. 이러한 적합 판단의 결과에 따라 검색시스템의 성능이 평가되므로 정보검색실험에서 적합 판단에 관한 타당성있고 신뢰성있는 연구방법이 요청되고 있다. 적합 판단이 다양한 변인에 의해 영향을 받을 수 있다는 개연성은 정보검색분야의 연구자들로 하여금 연구설계시 적합 판단 영향 요인을 가능한 한 배제할 수 있는 실험환경을 구성할 것을 촉구한다고 볼 수 있다.

먼저 적합 판단자로 기용하는 판단자 집단에 대한 신중한 선택이 요망된다. 판단자들의 판단의 일관성은 신뢰할 수 있으나 판단자의 특성에 따라 판단차이가 존재하므로 판단자 선정에 대한 고려가 연구설계에 반영되어야 할 것이다. 최종이용자들은 판단에 보수적이기는 하나 판단 일치율은 낮은 반면에 주제전문가는 판단 일치율은 높은 편이나 탐색문제를 제기한 본인이 아니라는 점에서 적합 판단에 부적절하다는 비판이 꾸준이 제기되어 왔다. 더욱이 최종이용자도 아니고 주제전문가도 아닌 제삼자를 판단자로 기용하는 것은 더욱 신중해야 할 것이다.

둘째, 적합 판단에 영향을 미칠 수 있는 요인으로 확인된 변인에 대한 과학적인 통제가 연구설계에 반영되어야 할 것이다.

판단자들은 일반적으로 표제를 주로 참고하는 것으로 나타났으나 표제가 적합 판단에 기여하는 정확도는 떨어지는 것으로 확인된 바 있다. 적합 판단에 가장 신뢰성있는 정확한 예측을 하는 것으로 확인된 문헌 대용물의 종류는 표제보다는 초록으로 드러난 만큼 적합 판단시 판단자에게 제시하는 문헌으로 초록이 추천되고 있다.

적합 판단시 일반적으로 사용하는 측정도구인 등급척도는 부적합한 측정 도구인 것으로 나타났다. 특히 적합, 부적합의 이등급으로 나누는 측정은 지양되어야 한다는 지적이다. 크기예측 측정 도구가 제안되고 있으나 판단 측정 도구에 관한 보다 깊은 연구가 요청되는 바이다.

이상의 논의에 유의해 볼 때 탐색된 문헌이 탐색질문에 부합하는 적합문헌인지를 결정짓는 과정에는 많은 요인들이 작용한다고 하겠다. 그러나 이와 같은 적합 판단에 영향을 미치는 변수들을 적절히 통제하여 연구나 실험을 과학적으로 설계한다면 적합판정은 비록 인간의 주관적인 활동에 의해 내려지는 판단이기는 하지만 정보검색의 효과를 평가하는 믿을 만한 측정도구가 될 수 있을 것으로 기대 되는 바이다.

참 고 문 헌

Barhydt, G. C. 1964. "A comparison of relevance assessments by three types of evaluator." Proceedings of the American Documentation Institute. October 5-8 : 383-385.

———. "The effectiveness of non-user relevance assessments." *Jr. of Documentation* 23, 1 : 46-49, 251.

Barry, Carol L. 1993. "A preliminary examination of clues to relevance criteria within document representations." Proceeding of the ASIS Annual Meeting 30 : 81-86.

———. 1994. "User-defined relevance criteria an exploratory study." *Jr. of American Society for Information Science* 45, 3 : 149-159.

Bellardo, Trudi. 1984. "Some Attributes of Online Search Intermediaries That Relate to Search Outcome." Ph.D. dissertation, Drexel University.

Boyce, 1982. "Beyond topicality : A two stage view of relevance and the retrieval process." *Information Processing and Management* 18 : 105-109.

Buckland, M.K. 1983. "Relatedness relevance and responsiveness in retreival systems." *Information Processing and Management* 19 : 237-241.

Burgin, R. 1991. "The effect of indexing exhaustivity on retrieval performance." *Information Processing and Management* 27 : 623-628.

Burgin, Robert. 1992. "Variations in relevance judgments and the evaluation of retrieval performance." *Information Processing and Management* 28, 5 : 619-627.

Cleverdon, Cyril., Mills, J. and Keen, M. 1966. ASLIB *Cranfield Research Project : Factors Determining the Performance of Indexing Systems*. 2 Vols. Cranfield, Bedfordshire : College of Aeronautics.

Cooper, W. S. 1971. "A definition of relevance for information retrieval." *Information Storage and Retrieval* 7 : 19-37.

———. 1973. "On selecting a measure of retrieval effectiveness." *Jr. of American Society for Information Science* 24 : 87-100.

Cuadra, Carlos A. and Katter, Robert V. 1967a. *Experimental Studies of Relevance Judgments : Final Report, Vol. 2. Description of Individual Studies.* Santa Monica, Calif. : System Development Corp.

———. 1967b. "Opening the black box of relevance." *Jr. of Documentation* 23, 4 : 291-303.

Doyle, L.B. 1963. "Is Relevance an Adequate Criterion for Retrieval System Evaluation?" In H.P. Luhn(Ed). Automation and Sciencetific Communication, Short papers, part 2(pp. 199-200). Washington, DC : American Documentation Institute.

Eisenberg, M. B. 1986. "Magnitude Estimation and the Measurement of Relevance." Ph. D. dissertation, Syracuse University, Syracuse, NY.

Eisenberg, Michael. 1988. "Measuring relevance judgments." *Information Processing and Management* 24, 4 : 373-389.

Eisenberg, Michael and Barry, Carol. 1988. "Order effects : A study of the possible influence of presentation order on user judgments of document relevance." *Jr. of American Society for Information Science* 39 : 293-300.

Eisenberg, Michael and Schamber, Linda. 1988. "Relevance : The search for a definition." *ASIS Mid-Year Proceedings* : 164-68.

Ellis, David. 1984. "Theory and explanation in information retrieval research." *Jr. of Information Science* 8 : 25-38.

Fenichel, Carol H. 1979. "Online Information Retrieval : Identification of Measures That Discriminate Among Users With Different Levels of Types of Experience." Ph.D dissertation, Drexel University.

Foskett, D.J. 1972. "A note on the concept of relevance." *Information Storage and Retrieval* 8, 2 : 72-78.

Gescheider, G. 1985. Psychophysics : Method, theory and application, 2nd ed. Hillsdale, N.J. Lawrence Erlbaum.

Gluck, M. 1996. "The exploring the relationship between user satisfaction and relevance in information systems." *Information Processing and Management* 32, 1 : 89-104.

Gull, C. D. 1956. "Seven years of work on the organization of materials in the special library."

American Documentaion 7 : 320-329.

Harter, Stephen P. 1971. "The Cranfield II relevance assessments : A critical evaluation." *Library Quarterly* 41, 3 : 229-43.

———. 1992. "Psychological relevance and information science." *Jr. of American Society for Information Science* 43, 9 : 602-616.

Janes, J. W. 1989. "Towards a Search Theory of Information." Ph. D dissertation, Syracuse University, Syracuse, NY.

Janes, Joseph W. 1991. "Relevance judgments and the incremental presentation of document representations." *Information Processing and Management* 27, 6 : 629-646.

Janes, Joseph W. and Mckinney, Renee. 1992. "Relevance judgments of actual users and secondary judges : A comparative Study." *Library Quarterly* 62, 2 : 150-68.

Katzer, J., and Snyder, H. 1990. "Toward a More Realistic Assessment of Information Retrieval Performance." Proceedings of the ASIS, Washington, DC. pp. 80-85.

Kemp, D. A. 1974. "Relevance, pertinence and information system development." *Information Storage and Retrieval* 10, 1 : 37-47.

Lancaster, F. W. 1969. "MEDLARS : Report on its operating efficiency." *American Documentation* 20 : 119-142.

Lesk, M. E. and G. Salton. 1968. "Relevance assessment and retrieval system evaluation." *Information Storge and Retrieval* 4, 3 : 343-359.

Marcus, R. S., Kugel, P. and Benenfeld, A. R. 1978. Catalog information and text as indicators of relevance." Jr. of *American Society for Information Science* 29 : 15-30.

Meadow, Charles T. 1985. "Relevance?" *Jr. of the American Society for Information Science* 36 : 345-55.

O'Connor, J. 1967. "Relevance disagreement and unclear request forms." *American Documentation* 18, 3 : 165-177.

———. 1969. "Some independent agreements and resolved disagreements about snswer-providing documents." *American Documentation* 20, 4 : 311-319.

Park, Taemin Kim. 1993. "The nature

of relevance in information retrieval : An empirical Study." *Library Quarterly* 68, 3 : 318-351.

Parker, Lorraine M. Purgailis and Johnson, Robert E. 1990. "Does order of presentation affect users' judgment of document?" *Jr. of American Society for Information Science* 41, 7 : 493-494.

Rath, G.J., A. Resnick, and Savage, T.R. 1961. "Comparisons of four types of lexical indicators of content." *American Documentation* 12, 2 : 126-130.

Rees, A.M. 1967. "Evaluation of information systems and services." *Annual Review of Information Science and Technology* 2 : 63-86.

Rees, A.M and Schultz, D.G. 1967. A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. Final Report. Center for Documentation and Communication Research, School of Library Science, Case Western University. Cleveland, OH.

Rees, Alan M. and T. Saracevic. 1966. "The Measurability of Relevance." *Proceedings of the American Documentation Institute* 3 : 225-234.

Regazzi, J.J. 1988. "Performance measures for information retrieval systems-an experimental approach." *Jr. of American Society for Information Science* 39 : 235-251.

Resnick, A. 1961. "Relative effectiveness of document titles and abstracts for determining the relevance of documents." Science 134(3484) : 1004-1006.

Resnick, A. and Savage, T.R. 1964. "The consistence of human judgments of relevance." American Documentation 15 : 93-95.

Saracevic, Tefko. 1969. "Comparative effects of titles, abstracts, and full texts on relevance judgments." Proceedings of ASIS 6 : 293-299.

———. 1970. "The Concept of" Relevance "in Information Science : A Historical Review." In T.Saracevic ed. *Introduction to information science New York : R. R. Bowker Co.* 111-151.

———. 1970. "Ten Years of Relevance Experimentation : A Summary and Synthesis of Conclusions." *Proceedings of the American*

Society for Information Science 7 : 33-36.

———. 1975. "Relevance : A review of and a framework for the thinking on the notion in information science." *Jr. of American Society for Information Science* 26 : 321-343.

——— et al. 1988. "A study of information seeking and retrieving. I. Background and methodology." *Jr. of American Society for Information Science* 39 : 161-76.

———. 1991. "Individual differences in organizing, searching and retrieving information." Proceeding of the ASIS Annual Meety 28 : 82-86.

Schamber, Linda., Eisenberg, Michael B. and Nilan, Michael S. 1990. "A reexamination of relevance : Toward a dynamic, situational definition." *Information Processing and Management* 26, 6 : 755-76.

Shaw, D. 1995. "Bibliographic database searching by graduate-students in language and literature-search strategies, system interfaces, and relevance judgments." *Library and Information Science Research* 17, 4 : 327-345.

Smithson, Steve. 1994. "Information retrieval evaluation in practice : A case study approach." *Information Processing and Management* 30, 2 : 205-221.

Sperber, D., and Wilson, D. 1986. Relevance : Communication and cognition. Cambridge, MA : Harvard University Press.

Stevens, S.S. 1966. "A metric for the social consensus." Science 151 : 530-41.

Swanson, Don R. 1971. "Some unexplained aspects of the Cranfield Tests of indexing performance factors." *Library Quarterly* 41, 3 : 223-28.

———. 1986. "Subjective versus objective relevance in bibliographic retrieval systems." *Library Quarterly* 56 : 389-98.

Thompson, C.W. N. 1973. "The functions of abstracts in the initial screening of technical documents by the user." *Jr. of American Society for Information Science* 24 : 270-76.

Zwsilocki, J.J. 1983. "Group and individual relations between sensation magnitudes and their nu merical estimation." *Perceptionand Psychophysics* 33, 5 : 460-468.