韓國經營科學會誌
第23卷 第4號
1998年 12月

불리언 질의 재구성에서 의사결정나무의 학습 성능 감도 분석

윤정미* · 김남호** · 권영식*

Sensitivity Analysis of Decision Tree's Learning Effectiveness in Boolean Query Reformulation

Jeong-Mi Yoon* · Nam-Ho Kim** · Young S. Kwon*

■ Abstract ■

One of the difficulties in using the current Boolean-based information retrieval systems is that it is hard for a user, especially a novice, to formulate an effective Boolean query. One solution to this problem is to let the system formulate a query for a user from his relevance feedback documents.

In this research, an intelligent query reformulation mechanism based on ID3 is proposed and the sensitivity of its retrieval effectiveness, i.e., recall, precision, and E-measure, to various input settings is analyzed. The parameters in the input settings is the number of relevant documents.

Experiments conducted on the test set of Medlars revealed that the effectiveness of the proposed system is in fact sensitive to the number of the initial relevant documents. The case with two or more initial relevant documents outperformed the case with one initial relevant document with statistical significances. It is our conclusion that formulation of an effective query in the proposed system requires at least two relevant documents in its initial input set.

1. 서 론

대규모 데이터베이스, 초고속통신망, 전자도서관, WWW(World Wide Web) 등과 같은 새로운 정보 기술의 출현으로 저렴하고 효율적인 저장 장치를 이 용한 정보 수집과 저장은 쉬워지고 정보량 또

* 동국대학교 산업공학과
** 선문대학교 산업공학과

한 기하급수적으로 증가하고 있 다. 이런 정보검색 환경의 변화로 많은 양의 정보 취득이 용이해짐에 따라 정 보탐색자가 원하는 정보를 효과적으로 검색하는 것이 중요한 문제로 제기되고 있다. 특히, 이제는 정보검색을 온라인(on-line)으로 집에서도 할 수 있어 과거에서처럼 전문가(사서)의 도움을 받기가 어려워졌다.

정보검색에서 가장 어려운 부분중의 하나가 정보탐색자가 원하는 정보만을 데이터베이스에서 검색하기 위한 정확한 질의를 구성하는 것이다. 정보탐색자가 단 한번으로 탐색에 성공하는 것은 어렵기 때문에 탐색을 반복적으로 수행하고 전에 검색된 문헌의 평가에 기초하여 질의를 재구성해야 한다. 지금까지 이런 질의 재구성 과정은 정보탐색자가 시스템의 피드백을 받아서 수동으로 다시 구성하는 것이 대부분이었다. 이처럼, 정보검색시스템을 사용하는 정보탐색자는 불명확한 정보요구만을 갖고 여러 번의 시행착오를 거쳐 질의를 만들어 나간다. 그러나 이러한 모든 과정에는 정보탐색자가 질의를 직접 재구성해야하는 어려움이 존재한다. 더우기, 정보탐색자들은 시스템이나 데이터베이스에 대하여 갖고 있는 지식이 불충분한 경우가 많고 또한 탐색자와 시스템이 사용하는 용어가 때로는 일치하지 않기 때문에 질의를 구성하고 또 재구성하기가 쉽지 않아 검색이 용이하지 않은 경우도 많다[1,4]. 이런 문제점들은 초보자가 적절한 질의를 구성하는 것을 더욱 어렵게 하고 있다.

본 연구에서는 이런 문제점을 해결하기 위해서 정보탐색자가 탐색어와 불리언 연산자(AND, OR, NOT)를 이용해 자신의 질의를 직접 작성하는 대신에 정보탐색자에게는 자신의 주관적 판단하에 질의에 적합한 문헌들만을 선택하게 한다. 이런 정보탐색자의 질의과정을 학습 알고리즘인 ID3(Iterative Dichotomizer 3)을 이용해서 학습하여 정보탐색자에게 편리하고 효율적인 지능적 질의 재구성 기법을 제시한다. 또한, 정보탐색자의 주관적 판단인 정보(문헌)들의 적합한 혹은 부적합한 특성을 학습하여 학습 성능 감도를 실험을 통해 분석하고 검색의 성능(효과성)을 향상시킬 수 있는 최적의 환경요소(초기 입력 문헌내의 적합문헌의 수)를 설정하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 질의 재구성 기법으로 이용된 ID3 알고리즘의 이론적 배경을 살펴보았고 3장에서는 새로운 정보검색시스템을 제안했다. 4장에서는 제안된 정보검색시스템으로 모의실험을 하고 그 결과를 분석했으며 5장에서는 본 논문의 요약과 앞으로의 연구과제에 대해 서술했다.

2. 이론적 배경

최근에는 질의 재구성 과정을 학습 과정으로 보고 그 과정을 자동으로 하는 연구가 시도되고 있다[6,7,8]. 정보검색문제에서는 원하는 정보를 단 한번에 검색하기 어렵기 때문에 여러 번에 걸쳐 정보를 검색하게 된다. 이 경우 탐색자의 질의 재구성 과정을 학습하는 것은 초보자가 일반적으로 부딪치는 난관인 질의의 재구성을 용이하게 한다.

ID3은 기계학습 분야에서 의사결정나무(decision tree)를 사용하는 TDIDT(Top Down Induction of Decision Trees) 집단에 속하는 것으로 Quinlan이 개발했다[2,11,12,13]. 이 것은 주로 분류 문제에 사용되어져 왔는데, 정보검색 또한 광범위한 정보로부터 정보탐색자가 원하는 정보만을 탐색자의 질의에 따라 검색하는 분류문제로 볼 수 있다.

ID3은 주어진 사례(문헌)들로부터 바람직한 사례는 모두 포함하면서 바람직하지 않은 사례는 하나도 포함하지 않는 결정나무 중에서 가장 간단한 의사결정나무(적은 수의 테스트만으로도 주어진 사례들을 올바르게 분류할 수 있는 나무)를 생성하는 것을 목적으로 한다. 결정나무는 노드와 가지로 구성된 일반적인 나무로서, 각각의 노드는 사례들이 표현된 속성(탐색어, 색인어)중의 어느 하나가 된다. 또한 가지는 그 속성을 테스트하였을 속성이 취할 수 있는 값을 의미하게 된다. ID3은 정보이론(information theory)에 따른 엔트로피(entropy) 개

념을 이용하여 가장 간단한 결정나무를 구하고 있다[5]. 일반적으로 엔트로피는 무질서도를 나타내는 정량적인 수치이다.

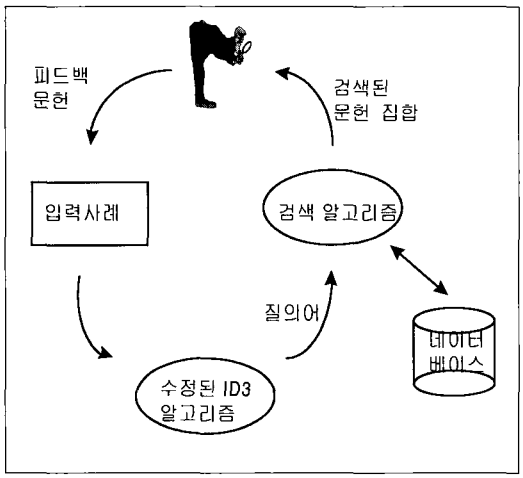
ID3에 입력된 사례 집합은 적합한 사례와 적합하지 않은 사례가 혼합되어 있으므로 어떤 속성이 어떻게 사례를 분류하는지에 대한 정보가 없기 때문에 엔트로피가 아주 높은 상태이다. 그러나, 일단 결정나무가 학습된 상태에서는 각각의 말단 노드가 하나의 등급으로만 결정이 되므로 무질서도는 0이 된다.

그러므로, ID3이 사례를 모두 분류한 상태에서는 엔트로피가 0이 되는 상태를 의미하므로 학습 사례를 가장 잘 분류하는 속성은 정보량의 감소량을 최대화하는 것이다. 따라서, 각 속성들이 분류 기준으로 선택될 경우의 정보량의 감소량을 계산한다. 다음으로, 가장 식별력이 높은 속성이 선택되면, 주어진 속성의 값의 종류수 만큼 가지를 만든다. 각각의 가지에 해당하는 값에 따라 사례들을 분할하고, 각각의 가지에서 지금까지의 과정을 반복한다. 더 이상의 정보량의 감소가 없다면 분할을 멈춘다.

3 제안된 정보검색시스템

3.1 ID3 정보검색시스템

본 연구에서는 [그림 3-1]과 같은 정보검색시스템 모형을 제시하고자 한다. 정보탐색자가 질의를 직접 구성하는 대신에 정보탐색자의 주관적 판정으로 이루어진 입력사례(정보탐색자가 자신의 질의에 적합하다고 판단하는 사례)를 수정된 ID3 알고리즘이 그 특성을 학습하여 분류한다. 분류의 결과 생성된 질의에 적합한 문헌을 데이터베이스에서 검색해서 정보 탐색자에게 피드백해 준다. 정보탐색자는 피드백된 문헌들을 보고 그 검색결과가 충분하면 검색을 중지하고, 그렇지 않으면 질의에 적합한 문헌들을 더 선택해 입력사례에 추가한다.



[그림 3-1] 제안된 정보검색시스템

3.2 개선된 ID3 알고리즘의 개발

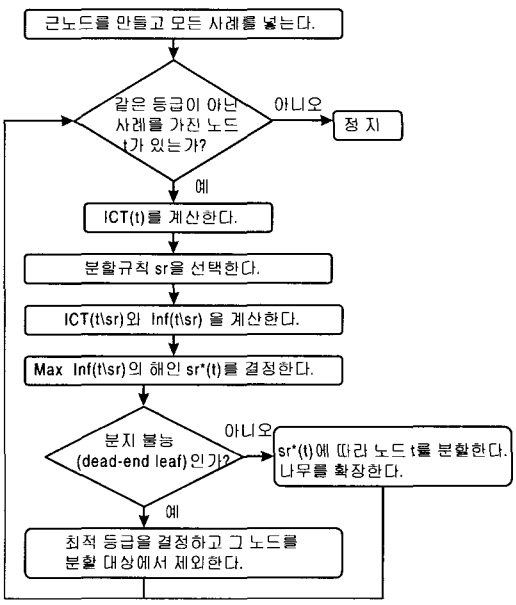
ID3 알고리즘은 정보검색 문제에 바로 적용하기에는 부적합한 몇 가지 문제점이 존재한다. 우선, 특정 속성의 선택 문제이다. ID3 알고리즘에 이용되는 사례 집합을 표현하는 속성들은 그 사례 집합을 가장 잘 표현하는 속성들로 구성되어야 한다. 그러나, 정보검색 분야에서 그 속성들은 색인어로 표현되는데 개개 문헌을 가장 잘 표현하는 것이므로 특별히 사례 집합을 가장 잘 표현하는 속성들로 다시 구성할 필요가 없다. 때문에 여기서는 사례 집합을 대표하는 색인어 전부를 실험 대상으로 설정했다. 그 다음으로, 분지 불능(dead-end leaf)인 경우이다. 이것은 모든 속성의 값이 같은데도 불구하고 다르게 분류하였을 때 일어나는 문제점으로서 어떠한 분할 규칙으로도 두 개의 사례를 분류할 수 없다.

(키, 머리색, 눈빛) -> 판정
사례1 (170, 검정, 갈색) -> 적합
사례2 (170, 검정, 갈색) -> 부적합

이를 수정된 ID3 알고리즘에 적용하기 위해 정보량의 변화가 없는 경우 더이상 노드를 분할하지 않

고 그 등급을 결정하는 알고리즘을 추가하였다. 마지막으로, 알려지지 않은 값이 존재하는 경우인데 이것은 만약 어느 한 사례의 속성값중 하나가 없다면 어떻게 처리할 것인가의 문제이다. 정보검색 문제에서 속성값 중 하나가 없다는 것은 그 속성이 해당 문헌을 표현하지 못하는 것으로 볼 수 있기 때문에 해당 문헌과 속성과의 관련성은 0으로 준다.

위와 같은 사항을 고려한 개선된 ID3 알고리즘은 아래의 [그림 3-2]와 같다. 입력사례의 각 속성에 대해 기대 정보량을 계산하여 정보의 변화량이 가장 큰 속성을 분류 속성으로 한다. 그리고 분류 속성에 대해 사례들이 부노드로 분지가 불가능한지를 판단하여 분지가 불가능할 경우에는 그 노드에 포함된 사례의 등급이 더 많은 등급을 그 노드의 최적 등급으로 결정하고 그 노드는 분할 대상에서 제외한다. 분지 불능이 아닐 경우에는 분류 속성에 따라 부노드로 분지 한다. 부노드로의 분지는 각 부노드에 포함된 사례가 동일할 등급이 될 때까지 반복된다.



[그림 3-2] 수정된 ID3 알고리즘

▶ **표기**

사례(case) : 각각의 속성에 해당하는 값과 속하는 등급(class)값을 가진 하나의 객체

$n(t)$: 노드 t에 포함되어 있는 사례의 개수

$n_i(t)$: 노드 t에서 등급(class) i인 사례의 개수 (i = 0, 1)

$n(t_j)$: 부노드 t_j에 포함되어 있는 사례의 개수

$sr(t)$: 노드 t에서의 분할 규칙(속성번호와 해당 값)

$sr^*(t)$: 노드 t에서의 최적 분할 규칙

TN : 말단 노드들의 집합

$P(i|t) = n_i(t) / n(t)$: 노드 t에 등급(class) i가 있을 조건부 확률

$P(t_j|t, sr) = n(t_j) / n(t)$: 노드 t가 분할 규칙 sr에 의하여 나누어질 때 노드 t에서의 사례가 부노드 t_j로 들어갈 확률

▷ 노드 t에서의 분류에 대한 정보량

$$ICT(t) = -\sum_i P(i|t)\log_2(P(i|t))$$

▷ 분할 규칙 sr에 따라 노드 t를 분할 할 때, 노드 t의 부노드 t_j에서 나무에 의해 전달되는 기대된 정보량

$$ICT(t|sr) = \sum_j ICT(t_j)P(t_j|t, sr)$$

▷ 분할 규칙 sr에 따라 노드 t를 분할하는 변화된 정보량

$$\Delta\text{Inf}(t|sr) = ICT(t) - ICT(t|sr)$$

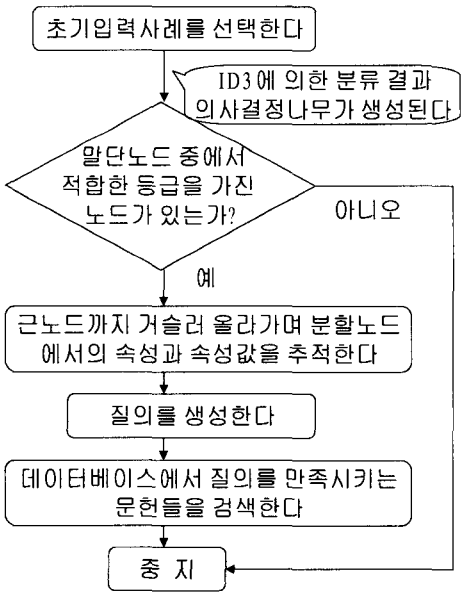
▷ 노드 t에서의 최적 분할 규칙

$$sr^*(t) = Max\,\Delta\text{Inf}(t|sr)$$

3.3 검색 알고리즘

[그림 3-3]에서 보듯이 수정된 ID3에 의한 분류 결과 말단 노드와 분할 노드(의사결정 노드)를 갖

는 결정나무가 생성되면 검색 알고리즘은 ID3에 의한 분류 결과 말단 노드중에서 적합한 문헌들만을 갖는 노드가 존재하면 근노드까지 거슬러 올라가며 각 노드의 속성과 속성값을 추적하여 AND 연산을 이용한 1차 질의로 변환시킨다. 이런 과정은 말단노드에 적합한 문헌들만을 갖는 노드가 없을때까지 계속한다. 그 결과로 만들어진 질의는 OR 연산을 이용하여 2차 질의로 변환된다. 이렇게 의사결정나무를 질의로 변환시킨 후 질의를 만족시키는 문헌을 데이터베이스에서 찾아 정보탐색자에게 그 결과를 알려 준다.



[그림 3-3] 검색 알고리즘

4. 실험 및 결과 분석

4.1 실험의 목적

현재 상용되고 있는 일반적인 정보검색시스템에서 정보탐색자가 이용하고 있는 방법은 탐색어와 불리언 연산자를 조합하여 질의를 만들어 내는 방법이다. 질의를 구성하는 탐색어의 선택은 정보탐색자의 주관적 판단에 근거한다. 따라서, 검색 성능의 효율은 정보탐색자가 만드는 질의의 형태에 따라 다양하게 나타날 수 있다.

이와 마찬가지로, 본 연구에서 제시한 방법인, 정보탐색자가 질의를 직접 구성하는 대신에(즉, 정보탐색자의 주관적인 판단하에 탐색어를 선택하는 것 대신에) 정보탐색자의 주관적 판단에 근거한 초기 입력사례의 특성을 학습하여 질의를 자동으로 구성하는 것 또한 초기 입력사례를 선택하는 방법에 따라 그 성능에 차이가 있을 것으로 추측된다.

따라서, 본 실험에서는 초기 입력사례를 총 5건으로 선택하는데(정보 검색 시작단계에서 정보탐색자가 갖고 있는 정보는 미약하기 때문에 5건으로 제한함.), 적합문헌이 각각 1건, 2건, 3건, 4건일 경우 학습 성능의 감도를 비교·분석하고 검색의 성능을 향상시킬 수 있는 초기 입력 문헌내 최적의 환경요소를 설정하고자 한다.

4.2 실험 환경

실험을 위한 기본 가정은 다음과 같다. 첫째, 초기 입력사례의 선택은 정보탐색자의 주관적 판단에 따라 변화될 수 있으므로 여기서는, 적합문헌 1건과 부적합문헌 4건인 총 5건, 적합문헌 2건과 부적합문헌 3건인 총 5건, 적합문헌 3건과 부적합문헌 2건인 총 5건, 적합문헌 4건과 부적합문헌 1건인 총 5건으로 제한하고 각 각의 초기 입력사례는 랜덤하게 선택한다. 둘째, 시스템이 검색해 준 사례로부터 다음 단계 입력사례를 선택하는 횟수는 검색 후의 결과로 검색한 문헌이 모두 적합문헌일 경우 ID3을 이용한 학습은 불가능하기 때문에 최대 3번으로 제한하고 다음 단계 입력사례의 선택을 위한 문헌의 수는 5건으로 제한한다. (다음 단계 입력사례의 선택을 위한 문헌수가 10건인 경우는 참고문헌[3]을 참고할 것). 셋째, 실험은 30문항의 질의를 각 조건에 대해 행한다.

여기에서 사용되는 실험 사례는 버지니아 CD-ROM 시리즈 중 Medlars 테스트 셋(test sets)이다[9]. MEDLARS (MEDical Literature Analysis &

Retrieval System)는 미국 국립의학도서관(U.S. National Library of Medicine)이 개발한 전산화된 종합 의학문헌 검색도구이다. Medlars 테스트 셋은 검색 효율성을 평가하기 위해 MEDLARS에서 일부분을 추출한 테스트 셋으로 제목, 저자, 요약문 등 서지사항이 들어 있는 문헌들과 질의와 그의 적합성 판정으로 이루어져 있다. 문헌은 총 1033건, 그에 따른 색인어는 총 7276개가 수록되어 있으며 질의는 총 30개로 구성되어 있다. 각 질의에 대한 적합문헌 사례도 수록되어 있다. 색인화일은 서지사항이 들어 있는 문헌화일로부터 실험자가 색인어의 유·무에 따라 만들었다.

검색효율에 대해 일반적으로 사용되는 평가 기준은 재현율(조회율, recall)과 정확률(precision)이 있다[14,15]. 재현율은 질의가 주어졌을 때 데이터베이스내의 적합문헌 수에 대한 검색된 적합문헌의 비율을 말한다. 재현율이 높다는 것은 데이터베이스내의 적합문헌을 많이 검색했다고 볼 수 있다. 정확률은 검색된 문헌에 대한 적합문헌 수의 비율을 말한다. 정확률이 높을수록 검색된 문헌 중에 부적합문헌이 들어있을 확률은 적어져 정밀도가 높아진다.

그러나, 재현율과 정확률의 한쪽 측면에서 정보검색의 성능을 비교하는 것은 편향적인 성격을 갖고 있기 때문에, 두 가지 모두를 동시에 평가하는 방법 이 개발되어 왔다. 한 가지 방법은 재현율-정확률 그래프를 사용하는 것으로, 재현율과 정확률은 0과 1 사이의 값을 갖게 된다. 일반적으로 데이터베이스내의 적합문헌을 가능한 한 많이 검색하기 위해서는 재현율을 높이게 되는데 이때에는 부적합문헌도 많이 검색되기 때문에 정확률은 낮아지는 현상을 보이게 된다.

재현율과 정확률은 역관계를 가지고 있으므로 시스템을 비교하는데 어려움이 있어 재현율과 정확률의 결합측정치 인 E가 Rijserbergen에 의해 다음과 같이 정의되었다[17].

$$E = 1 - \frac{(1 + \alpha^2) * PR}{\alpha^2 * P + R}$$

여기에서 P=정확률, R=재현율, α는 재현율과 정확률을 사용하는 탐색자에게 있어서의 상대적 중요치를 나타낸다. α가 0.5이면 재현율에 정확률의 0.5의 중요도를 부여하는 경우이며 α가 1일 때는 동일한 중요도, α가 2이면 재현율에 정확률의 2배의 중요도를 부여한다. E 측정치가 낮을수록 검색효율은 높다.

본 실험에서는 평가 기준으로 재현율과 정확률 및 E 측정치를 사용하여 초기 입력사례의 변화가 검색성능에 영향을 미치는지 그 평균을 비교하여 알아본다. 실험 결과의 통계적 분석을 위해 분산분석을 통계 분석 소프트웨어인 SPSS를 사용한다[10,16]. 분산분석(analysis of variance)은 두 표본 이상의 평균치 차이를 검정하는 통계적 기법이다. 분산분석을 이용하기 위해서는 정규성의 가정을 필요로 하는데 본 실험의 경우 표본이 30개로 표본크기가 크기 때문에 정규성의 가정을 필요로 하지 않는다. 실험을 위한 ID3의 알고리즘은 Turbo C++로 구현되었으며, IBM PC 586 호환 기종에서 실험하였다.

4.3 실험 방법 및 결과 분석

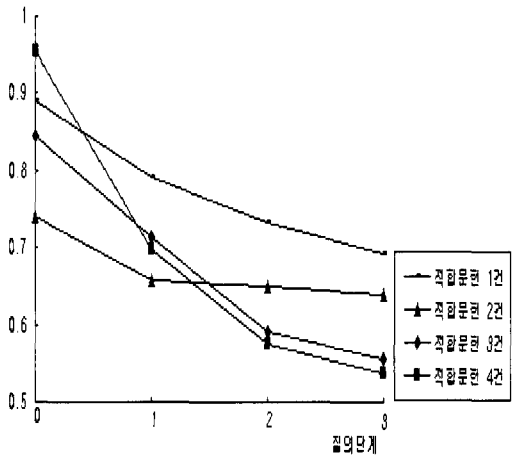
실험은 시스템에 피드백되는 문헌의 수가 최대 5건인 경우 총 5건의 초기 입력사례중 적합문헌을 1건, 2건, 3건, 4건으로 다르게 주었을 경우 적합문헌 수의 변화에 따른 검색 성능을 비교해 행해진다.

초기 입력사례는 적합문헌 1건과 부적합문헌 4건, 적합문헌 2건과 부적합문헌 3건, 적합문헌 3건과 부적합문헌 2건, 적합문헌 4건과 부적합문헌 1건인 총 5건을 질의마다 랜덤하게 선택한다. 각 질의 단계에서 정확률이 1인 경우에는 질의를 재구성하지 않고 검색을 끝낸다. 실험 결과를 나타내면 아래의 <표 4-1>과 같다. 평균 재현율과 평균 정확률, 평균 E 측정치는 질의 30 문항에 대한 각 단계별 평균 재현율과 평균 정확률, 그리고 평균 E 측정치를 평균한 것이다. 질의 재구성 각 단계별 평균 E 측정치, 평균 재현율-평균 정확률의 변화

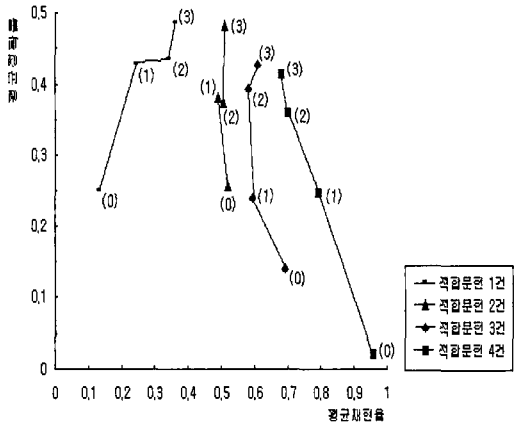
를 그래프로 나타내면 [그림 4-1]과 [그림 4-2]와 같다.

〈표 4-1〉 피드백 문헌이 5건인 경우
(각 단계별 평균 재현율과 정확률 및 평균 E 측정치)

		평균 재현율	평균 정확률	평균 E 측정치
적합문헌 1건	제 0 단계	0.1255	0.250	0.8915
	제 1 단계	0.2403	0.4281	0.7914
	제 2 단계	0.3359	0.4351	0.7317
	제 3 단계	0.3561	0.4868	0.6923
적합 문헌 2건	제 0 단계	0.5170	0.2531	0.7416
	제 1 단계	0.4885	0.3774	0.6578
	제 2 단계	0.4989	0.3702	0.6494
	제 3 단계	0.5054	0.4767	0.6385
적합 문헌 3건	제 0 단계	0.6876	0.1396	0.8449
	제 1 단계	0.5926	0.2401	0.7138
	제 2 단계	0.5760	0.3931	0.5918
	제 3 단계	0.6078	0.4257	0.5567
적합 문헌 4건	제 0 단계	0.9491	0.0234	0.9544
	제 1 단계	0.7855	0.2473	0.6984
	제 2 단계	0.6937	0.3587	0.5739
	제 3 단계	0.6775	0.4126	0.5377



[그림 4-1] 피드백 문헌이 5건인 경우의 평균 E측정치 비교



[그림 4-2] 피드백 문헌이 5건인 경우의 평균 재현율-평균 정확률 비교

<표 4-1>과 [그림 4-1]에서 볼 수 있듯이 4가지 경우(적합문헌이 1건, 2건, 3건, 4건인 경우) 모두 질의 재구성이 진행됨에 따라 평균 E 측정치가 낮아져 검색 성능이 나아짐을 알 수 있다. 실제로, 질의 재구성을 하지 않은 초기 입력사례의 E 측정치를 보면 적합문헌이 2건, 3건인 입력사례가 각각 0.74 0.84로 더 낮음을 알 수 있다. 그러나 질의 재구성이 진행되면서 시스템은 피드백되는 적합문헌과 부적합문헌들의 특성을 학습하기 때문에, 전 단계보다 더 나은 검색 결과를 보여 질의 재구성 마지막 단계인 제 3 단계를 지나면 적합문헌이 3건(0.55), 적합문헌이 4건(0.53)인 입력사례가 적합문헌이 1건(0.69), 적합문헌이 2건(0.64)인 사례보다 평균 E 측정치가 더 낮아 전반적으로적합문헌의 수가 많은 입력사례의 검색 성능이 더 나음을 알 수 있다.

평균 재현율과 평균 정확률의 경우 <표 4-1>과 [그림 4-2]에서 보듯이 초기 입력사례내의 적합문헌의 수가 많으면 데이터베이스내 적합문헌을 훨씬 더 많이 검색하지만(적합문헌 3건: 약 68%, 적합문헌 4건: 약 94%), 검색된 문헌들 중에 실제로 적합한 문헌은 많지 않아(적합문헌 3건: 약 14%, 적합문헌 4건: 약 2%) 검색 성능은 좋지 않다. 그러나, 질의 재구성이 진행됨에 따라 적합문헌의 수

가 적은 입력사례는 전 단계보다 더 나은 재현율을 보이고 적합문헌 의 수가 많은 입력 사례의 경우에는 다소 재현율이 낮아짐을 볼 수 있다. 그러나 여전히 초기 입력사례내의 적합문헌의 수가 3건, 4건인 경우 질의 재구성 제 3 단계 후의 평균 재현율이 각각 60%, 67%로 적합문헌 1건(35%), 적합문헌 2건(50%)인 입력사례보다 더 우수하다. 그리고 평균 정확률은 질의 재구성이 진행됨에 따라 점점 더 높아져 질의 재구성 제 3 단계 후의 평균 정확률은 적합문헌 1건, 2건(48%, 47%)인 입력사례가 적합문헌 3건, 4건(42%, 41%)인 입력사례보다 더 높지만 그 차이가 크지는 않다. 전반적으로, 적합문헌이 2건 이상인 입력사례가 적합문헌이 1건인 입력사례보다 그래프 상에서 더 우측상향에 위치하여 그 검색 성능이 더 우수함을 알 수 있다.

앞서 나온 결과가 통계적으로 유의한지 알아보기 위해 유의수준 5%로 분산분석을 하였다. [그림 4-3]은 질의 재구성 제 1 단계의 통계분석이고, [그림 4-4]는 질의 재구성 제 2 단계의 통계분석이며, [그림 4-5]는 질의 재구성 제 3 단계의 통계분석에 대한 결과이다. 평균 E 측정치와 평균 정확률은 질의 재구성 각 단계마다 유의수준이 0.05 보다 커서 초기 입력사례내의 적합문헌의 수의 변화(1건, 2건, 3건, 4건)는 통계적으로 별다른 차이가 없음을 알 수 있다.

평균 재현율은 질의 재구성 제 1 단계의 경우 적합문헌이 많을수록 통계적으로 유의하다. 질의 재구성 제 2 단계와 제 3 단계에서는, 적합문헌이 1건일 때보다는 그 이상일 경우 통계적으로 유의하고, 적합문헌이 2건 이상인 경우, 서로간에는 통계적으로 큰 차이가 없음을 알 수 있다. 즉, 전반적으로 초기 입력사례내의 적합문헌을 1건으로 주는 것보다는 적합문헌을 2건 이상으로 줄 때 데이터베이스내의 적합문헌을 더 많이 검색하고 또한 적합문헌이 2건 이상일 경우 평균 재현율에 대한 통계적인 차이는 적합문헌이 2건일 때보다 4건일 때 발생한다. 그러나 일반적으로 초기 입력사례내 적합문헌의 수가 많을 수는 없으므로, 적합문헌의 수가 2건 이상인 입력사례가 적합문헌이 1건인 입력사례보다 데이터베이스에서 적합문헌을 더 많이 검색한다고 할 수 있다.

5. 결론 및 제언

본 연구의 결과, 초기 입력사례의 적합문헌 수를 1건, 2건, 3건, 4건으로 제한한 상황하에서 평균 E 측정치를 비교해 보면 질의 재구성이 진행됨에 따라 평균 E 측정치는 낮아져 질의 재구성 제 3 단계 후의 결과를 보면 전반적으로 적합문헌 수가 많은 입력사례의 성능이 더 우수하지만 통계적으로는 별 차이가 없다.

평균 재현율-평균 정확률 측면에서 보면, 적합문헌수에 따른 평균 정확률은 질의 재구성 제 3 단계 후의 결과를 보면 전반적으로 적합문헌의 수가 적은 입력사례가 더 우수하나 그 차이는 미세하고, 통계적으로도 큰 차이가 없다.평균 재현율은 질의 재구성 제 3 단계 후의 결과를 보면 적합문헌의 수가 많을수록 검색 성능이 우수하고, 통계분석 결과 적합문헌의 수에 따른 통계적 차이가 있어 적합문헌의 수가 2건 이상인 입력사례가 적합문헌이 1건인 입력사례보다 데이터베이스에서 적합한 문헌을 더 많이 검색하는 것을 알 수 있다. 그러나 적합문헌의 수가 2건 이상인 입력사례의 경우 적합문헌수의 변화(2건, 3건, 4건)에 따른 통계적인 차이가 거의 없어 초기 입력사례내의 적합문헌을 1건, 2건, 3건, 4건으로 주는 실험환경에서는 정보이용자가 정보검색 초기 단계에 일반적으로 적합 정보를 많이 알고 있지 않기 때문에 적합문헌을 2건으로 주는 경우가 가장 적절하다.

본 연구는 정보탐색자가 원하는 정보를 찾기 위해 탐색어와 불리언 연산자를 이용해 질의를 직접 구성하는 방법대신에, 각 질의단계에서 관련사례(문헌)를 선택하는 정보탐색자의 질의과정을 지능적으로 학습하여 검색하는 질의 재구성 기법을 제시함으로써 초보자라도 정보검색을 쉽게 하는데 도움을 줄 것이다. 이런 식의 질의 방법은 온라인

상에서 고객이 얻고자 하는 기사를 선택하면 다음에는 그와 관련되는 기사만을 선정해서 자동으로 고객에게 보내주는 소위 맞춤신문(customized news delivery service)과 같은 것에 응용할 수 있다.

앞으로 본 연구의 방향은 첫째, 여러 가지 테스트 셋을 이용한 실험을 행하여 검색 성능을 일반화할 수 있을 것이다. 둘째, 정보탐색자가 다음 질의 재구성에 이용되는 입력사례를 위해 피드백하는 문헌의 수를 10건 이상으로 확대하여 검색 성능의 변화를 알아볼 수 있을 것이다. 셋째, 본 실험에서는 초기 입력자료는 총 5건, 적합문헌의 수를 1건, 2건, 3건, 4건으로 제한하여 실험했지만 초기 입력자료의 총 건수와 적합문헌의 수를 더 다양하게 하여 검색 성능을 살펴볼 수 있을 것이다.

참 고 문 헌

[1] 문성빈, "적합성 피드백을 이용한 전문검색시스템의 검색효율성 증진을 위한 연구," 「정보관리학회지」, 제10권, 2호(1993).

[2] 박영택, 이강로, "ID3 계열의 귀납 적기계학습," 「한국정보과학회지」, 제13권, 제5호(1995), pp.7-19.

[3] 윤정미, "불리언 질의어 재구성에서 초기입력자료의 변화에 따른 의사결정 나무의 학습성능 감도분석", 동국대학교, 1996.

[4] 정영미, 「정보검색론」, 구미무역 (주)출판부, 1993.

[5] 조동성 역, Gordon B. Davis, Margrethe H. Olson 원저, 「경영 정보시스템」, 석정 출판사, 1987.

[6] Chen H. and J. Kim, "GANNET: A Machine Learning Approach to Document Retrieval," *Journal of Management Information Systems*, Vol.11, No.3(Winter 1994/95), pp.9-43.

[7] Chen H., "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms," *Journal of the American Society for Information Science*," Vol.46, No.3(1995), pp.194-216.

[8] Croft W. Bruce, "Machine Learning and Informatin Retrieval," COLT '95 Conference, 1995.

[9] Fox, E. A. and S. G. Winett, "Using Vector and Extended Boolean Matching in an Expert System for Selecting Foster Homes," *J. of Amer. Soc. Infor. Sci.*, Vol.41, No.1 (1990), pp.10-26.

[10] Mongomery, D. D., *Design and analysis of experiments*, John Wiley&Sons, New York, 1976.

[11] Quinlan, J. R., "Learning efficient classification procedures and their application to chess end games," in *Machine Learning : An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell(Eds.), Tioga, Palo Alto, CA, 1983.

[12] Quinlan, J. R., *"Induction of decision trees,"* Machine Learning 1(1986), pp.81-106.

[13] Quinlan, J.R., "Simplifing decision trees," *International Journal of Man-Machine Studies*, 27, 1987.

[14] Salton, G., and M. McGill. *An Instruction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[15] Spark-Jones, K. *Information Retrieval Experiment*, Butterworths.,London, 1981.

[16] *SPSS for Windows Base System Users Guide*, SPSS Inc., 1993.

[17] Van Rijsbergen C. J. *Information Retrieval*, Butterworths., London, 1979.