OPAC에서 탐색결과의 클러스터링에 관한 연구*

The Effectiveness of Hierarchic Clustering on Query Results in OPAC

노 정 순(Jung-Soon Ro)**

목 차

초 록

본 연구는 한글 OPAC에서 문헌의 분류와 브라우징에 적합한 정적 계층클러스터링 모형이 서명단어 탐색으로 검색된 탐색결과를 클러스터링하는데도 효과적인지를 규명하기 위해 수행되었다. 서명에 출현하는 단어와 색인자가 부여한 통제어를 통합한 색인어를 이진빈도로 가중치를 주어, 다이스와 자카드 계수, 집단간 평균연결과 완전연결 클러스터링 기법이 테스트되었다. 16개의 서명단어 탐색으로 검색된 문헌을 클러스터링한 결과 최적으로 선택된 클러스터의 정확률은 유사도 계수나 클러스터링 기법에 관계없이 서명단어탐색보다 100%이상 향상되었다. 1단계와 최종단계 클러스터링 모두에서, 정확률 측면에서는 완전연결이, 재현율 측면에서는 집단간 평균연결이 더 효과적이었으나 통계적으로 유의한 수준은 아니었다. 1단계 클러스터에서 집단간 평균연결이 보다 높은 재현율을 보인 것은 유의하였다. 다이스와 자카드 사이에 차이는 없었다. 최종클러스터가 선택되기까지 집단간 평균연결은 너무 긴 계층군집 단계를 필요로 하여 탐색효율 측면에서 바람직해 보이지 않았다.

ABSTRACT

This study evaluated the applicability of the static hierarchic clustering model to clustering query results in OPAC. Two clustering methods(Between Average Linkage(BAL) and Complete Linkage(CL)) and two similarity coefficients(Dice and Jaccard) were tested on the query results retrieved from 16 title-based keyword searchings. The precision of optimal clusters was improved more than 100% compared with title-word searching. There was no difference between similarity coefficients but clustering methods in optimal cluster effectiveness. CL method is better in precision ratio but BAL is better in recall ratio at the optimal top-level and bottom-level clusters. However the differences are not significant except higher recall ratio of BAL at the top-level cluster. Small number of clusters and long chain of hierarchy for optimal cluster resulted from BAL could not be desirable and efficient.

키워드: 온라인 목록, 문헌 클러스터링, 계층 클러스터링, 탐색결과, 유사도
OPAC, Document Clustering, Query Result, Similarity, Hierarchic Clustering

---

* 이 논문은 2002년도 한남대학교 학술연구조성비 지원에 의하여 연구되었음.
** 한남대학교 문헌정보학과 교수(jsr@mail.hannam.ac.kr)
논문접수일자 2004년 2월 4일
게재확정일자 2004년 3월 12일

1. 서 론

1. 1 연구목적

정보가 증가할수록 정보검색시스템에서 특정 정보요구에 대해 너무 많이 검색되는 문헌은 시스템 이용자나 운영자 모두에게 고민거리가 되고 있다. 논문의 preprint를 대상으로 검색을 제공하는 IR시스템을 검색하기 위해 Web 탐색엔진 Google에서 "preprint database*"를 탐색한 결과 71,000건이 검색되었다. 보다 보편화된 주제 "digital libraries"는 4,130,000건이 검색되었다(2004. 1. 30 현재).

온라인 도서관목록(Online Public Access Catalog: OPAC)에서도 너무 많은 문헌의 검색은 특히 그 속에 포함된 너무 많은 부적합 문헌 때문에, 높은 정확률을 요구하는 이용자의 OPAC에 대한 요구를 만족시키지 못하고 있다. 저자나 서명을 알지 못하고 수행하는 주제탐색에서 이용자들이 원하는 문헌 수가 10권 내외임을 고려하면[1], 수십 권 혹은 수백 권의 문헌이 검색되어[2], 출판년도순, 저자명순 혹은 서명순으로 정렬되어 제공되는 검색결과 리스트는 이용자로 하여금 적합문헌을 찾기 위해 각 문헌들을 하나 하나 확인하는데 너무 많은 수고와 시간을 쏟게 한다.

주제탐색 중에서도 주제명키워드 탐색보다는 서명단어 탐색에서 부적합 문헌의 검색은 심각한 것으로 보고되었다. 평균 19개의 문헌을 검색한 주제명 탐색에 비해 서명단어탐색은 105개의 문헌을 검색하였다(Peters & Kurth 1991). 서명단어탐색의 정확률은 중앙치 7%이었고, 90개 탐색 중 72개 탐색(82%)이 19%이하의 정확률을 보였다(Carlyle 1996).

검색결과를 축소하는 기법으로는 적합성 순위화기법에 기반하여 상위 문헌만을 제공하거나, 검색결과를 유사한 문헌끼리 클러스터링하여 제공하여 이용자가 적합클러스터를 선택하게 하는 기법 등이 연구되었다.

문헌클러스터링은 원래 효율적인 파일조직 기법으로 연구되었으나, 문헌과 지식의 분류, 데이터베이스 또는 검색결과의 브라우징이나 분류 등에 응용되고 있다. 분류를 위한 계층 클러스터링은 문헌집단이나 분류 자질, 유사도 계수, 클러스터링 기법 등에 따라 그 결과가 매우 다르게 나타난다. 학술지 논문초록이나 신문기사, 웹문서의 클러스터링에서 단순 빈도보다는 역문헌빈도, 자카드나 코싸인 유사도가 좋은 성능을 보이는 것으로 보고되고 있다. OPAC에서는 문헌의 대용물인 서명이나 목차, 권말색인에 출현하는 단어를 분류자질로 삼아 문헌을 클러스터링하여 생성된 클러스터를 분류번호와 비교한 연구가 수행되었다.

본 연구는 학술논문이나 신문, web문서의 초록이나 전문과 비교하여 색인어수가 매우 작고 단어의 출현빈도수가 매우 작은 OPAC에 특히 한글 OPAC에 적합한 클러스터링 모형을 연구한 연구(노정순 2004)의 후속연구로, 선행실험에서 연구된 한글 OPAC에서

1) Larson(1986)의 연구에서는 9.1건, Wiberley, Daugherty, & Danowski(1995)의 연구에서는 9건, Barbuto & Cevallos(1991) 연구에서는 이용자의 54%가 1-20건이 검색되기를 원하였다.

2) 이용자는 평균 77.5건(Larson 1986), 179.6건(Lynch 1989), 200건(Larson 1991), 90건(Wiberley, Daugherty & Danowski 1995)의 문헌을 검색하였다.

문헌의 분류 브라우징에 최적으로 제시된 정적 클러스터링 모형이 서명단어 탐색으로 검색된 문헌을 클러스터링하는데도 효과적인지를 연구하기 위해 수행되었다.

1. 2 연구과제

1) OPAC 문헌의 브라우징에서 연구된 최적의 클러스터링 모형은 서명단어 탐색으로 검색된 문헌을 유사한 문헌끼리 그룹화하는데 효과적인가?

2) 서명단어 검색결과를 클러스터링하는 것은 검색의 성능을 향상시키는가?

3) 서명단어 검색결과를 클러스터링하여 얻은 1단계 클러스터나 최종단계의 클러스터에는 부적합문헌이 어느 정도 제거되고 적합문헌은 어느 정도 포함되는가?

4) OPAC 문헌의 브라우징 모형에서 최적의 클러스터링 기법으로 제안된 집단간 평균연결 기법과 완전연결 기법은 탐색결과의 클러스터링에서 어떤 차이를 가져오는가?

5) OPAC 문헌의 브라우징 모형에서 유용했던 다이스와 자카드 유사도계수는 탐색결과의 클러스터링에서 어떤 차이를 가져오는가?

2. 선행연구

클러스터링에 대한 연구는 파일조직의 한 기법으로 전체 파일을 탐색하는 대신 정보요구와 관련된 문헌클러스터만을 탐색하여 검색효율(efficiency)을 향상시키기 위한 목적으로 클러스터링 알고리즘을 연구하는 것에서 시작하였으나 검색의 효과(effectiveness)를 위한 계층 클러스터링에 대한 연구로 이어졌다. 1980년대에 들어와서 문헌클러스터링은 파일조직보다는 문헌과 지식의 분류에 문헌클러스터링을 적용하는 연구로 이어졌고, 최근에는 데이터베이스 또는 검색결과를 브라우징하기 위한 목적의 연구가 증가하였다.

클러스터링 기법은 접근방법상 계층클러스터링 기법과 자기발견적 기법이 있다. 빠른 계산방법 때문에 자기발견적 기법이 먼저 사용되었으나 탐색의 효과는 비클러스터 파일보다 못하였다(Salton 1971). 이 기법을 정보검색에서 응용한 최근의 연구(Silverstein & Pedersen 1997, Zamir & Etzioni 1998) 역시 효과보다는 효율성을 강조하고 있다.

계층 클러스터링은 하위 클러스터내의 문헌간 유사도가 상위 클러스터내의 유사도 값보다 커지도록 문헌집단을 계층분류 형식으로 나누는 것이다. 계층군집에서 검색질문과 유사도가 가장 큰 군집을 선택하는 방법으로는 top-down과 bottom-up 탐색방법이 있으나, bottom-up방식이 top-down보다 더 효과적인 것으로 보고되었다(Croft 1980, El-Hamdouchi & Willett 1989). 클러스터링으로 생성된 계층군집에서 탐색질문과의 유사도가 가장 큰 군집을 선택하는 시스템에서 Griffiths, Robinson & Willett(1984)은 이전까지 주로 사용된 단일연결(single link) 기법이 완전연결, 집단평균, 워드(Ward) 기법에 비하여 성능이 좋지 못함을 밝혔다.

탐색결과에 클러스터링을 적용하는 연구는 Preece(1973)를 시작으로 Willett(1985), Hearst & Pederson(1996), Leouski &

Allan(1998), Tombros, Villa, & Van Rijsbergen(2002)의 연구로 이어졌다. Willett (1985)은 탐색결과를 클러스터링하는 동적 클러스터링과 탐색질문을 사용하지 않고 전체 문헌집단을 대상으로 클러스터링하는 정적 클러스터링을 비교하였는데, 동적 클러스터링은 정적 클러스터링에 비해 분명하게 나쁘지는 않았다. Hearst & Pedersen(1996)은 탐색결과 검색된 상위 n개 문헌(100, 250, 500, 1000)을 대상으로 Scatter/Gather 시스템을 테스트하였다. 5개의 분할 클러스터링 방법이 사용되었는데, 가장 좋은 클러스터링에서 선택된 모든 군집이 최소한 전체 적합문헌의 50%를 담고 있었다. 또한 가장 좋은 군집내의 문헌 n개의 순위는 전체 상위 n개에서의 순위보다 우수하였다. Leouski와 Allan(1998)은 도치색인화일에서의 탐색결과 얻은 Top 50개 문헌을 다차원공간에서 적합문헌들이 어떻게 인접하고 있는지를 조사하였다. 문헌간의 유사도를 공간거리로 표현한 다차원공간 분석에서 적합문헌들은 서로 밀접하게 위치하였다. Tombros, Villa, & Van Rijsbergen(2002)은 5개의 문헌집단(문헌 수는 1,033~74,520)에서 단일링크와, 완전연결, 집단평균, 워드 기법을 테스트하였다. 질문과 유사도 순으로 순위화된 탐색결과로부터 top n(100, 200, 350, 500, 750, 1000)개의 문헌을 대상으로 클러스터링하였다. 문헌 수 n은 클러스터링 성능에 영향을 끼치지 않았고, 정적 클러스터링은 동적 클러스터링보다 모든 계층단계에서 성능이 떨어졌다. 4종류의 클러스터링 중 집단간 평균이 가장 좋았고 단일연결이 가장 나빴다.

검색결과를 클러스터링하여 시각적으로 보여주거나 브라우징할 수 있도록 하는 연구는 특히 웹문헌을 대상으로 수행되었고, 탐색결과의 클러스터링은 Altavista, Alltheweb, Vivisimo, QueryServer 등과 같은 여러 상업적인 웹탐색엔진에서 사용되고 있다. Zamir & Etzioni(1998)는 MetaCrawler로 검색된 Web문서를 대상으로 6가지 클러스터링 기법을 비교하였다. 성능은 STC, 집단평균, Fractionation, Buckshot, K-평균, Single Pass 순서로, STC가 가장 좋은 정확률을 보였으며, Single Pass가 가장 성능이 좋지 못하였다. Roussinov & Chen(2001)은 Altavista로 검색된 상위 200개의 결과를 클러스터링하여 클러스터를 제공하면, 이용자가 적합클러스터를 선택하고, 선택된 클러스터를 요약하는 단어리스트 중 이용자로 하여금 용어를 선택 혹은 삭제하게 하여 얻은 용어 적합성피드백으로 질의를 확장하는 중계 레이어를 실험하였다.

한편 단행본도서를 대상으로는 동적 클러스터링보다는 정적 클러스터링 연구가 수행되어, 클러스터링으로 생성된 클러스터를 분류번호와 비교하였다. Garland(1983)는 LC분류번호 Q(과학)분야 416권의 도서를 단일연결을 사용하여 LCSH과 서명에 출현한 단어로 클러스터링하여 LC분류번호와 비교하였다. Enser (1985)는 DDC 001.4~001.6과 330.519, 658.4분야의 도서 250권을 Willett알고리즘을 사용하여 분류자질로서의 서명과 목차, 권말색인을 비교하였다.

3. OPAC 분류 열람용 클러스터링 모형의 요약

OPAC에서의 분류 열람을 위한 정적 클러스터링 모형을 연구한 1차 연구(노정순 2004)에서는 문헌정보학(DDC 020)분야 단행본(학위논문 포함) 175권을 대상으로 색인방법과 색인어 가중치 기법, 유사도 계수, 계층 클러스터링 기법이 변수로 사용되었다. 서명단어를 대상으로 절대빈도와 이진빈도를 각각 가중치로 부여한 자동색인과, 이진빈도의 서명단어에 색인자가 부여한 통제어를 통합시킨 통제어 통합색인 파일에서, 집단내 평균연결과, 집단간 평균연결, 완전연결 클러스터링 기법이 테스트되었다. 유사도 계수는 절대빈도에서는 피어슨과 코싸인, 제곱 유클리드 계수가, 이진빈도에서는 다이스와 자카드, 제곱 유클리드 계수가 사용되었다.

실험결과 모든 파일에서 집단내 평균연결 클러스터링과 제곱 유클리드 유사도 계수를 제외하고 다른 군집기법이나 유사도 계수는 비교적 좋은 클러스터를 생성하였다. 자동색인보다는 통제어 통합색인에서 보다 좋은 클러스터가 생성되었으며, 특히 통제어 통합색인에서 집단간 평균연결-자카드 유사도 기법과 완전연결(자카드, 다이스 상관없이)이 가장 주제적이고 분명하고 우수한 클러스터를 생성하였다. 집단간 평균연결-자카드로 생성된 클러스터가 보다 10진구조와 유사하였다.

4. 실험설계

본 연구에서는 선행실험에서 OPAC 분류의 자동화에 가장 적합한 클러스터링 모형으로 제시된 통제어통합 이진백터에서 다이스(D)와 자카드(J) 계수, 집단간 평균연결과 완전연결 클러스터링 기법이 테스트되었다.

실험을 위해 서명단어탐색은 2003년 2학기 한남대 문헌정보학과 3학년 교과목 "색인 및 시소라스"를 수강한 학생 24명이 탐색질의를 가지고 서명단어탐색을 수행하였다. MAESTRO 시스템의 논리합 정규형(Disjuntive Normal Form: DNF) 질의표현의 인터페이스의 문제점은 학생들에게 충분히 숙지되었다.

탐색은 AND와 OR 연산자를 하나 이상 사용하도록 권고되었고 탐색결과는 한국어로 제한시켰다. 검색건수가 너무 많으면 50건 내외가 되도록 문헌의 종류나 출판년도를 제한하도록 하였다.

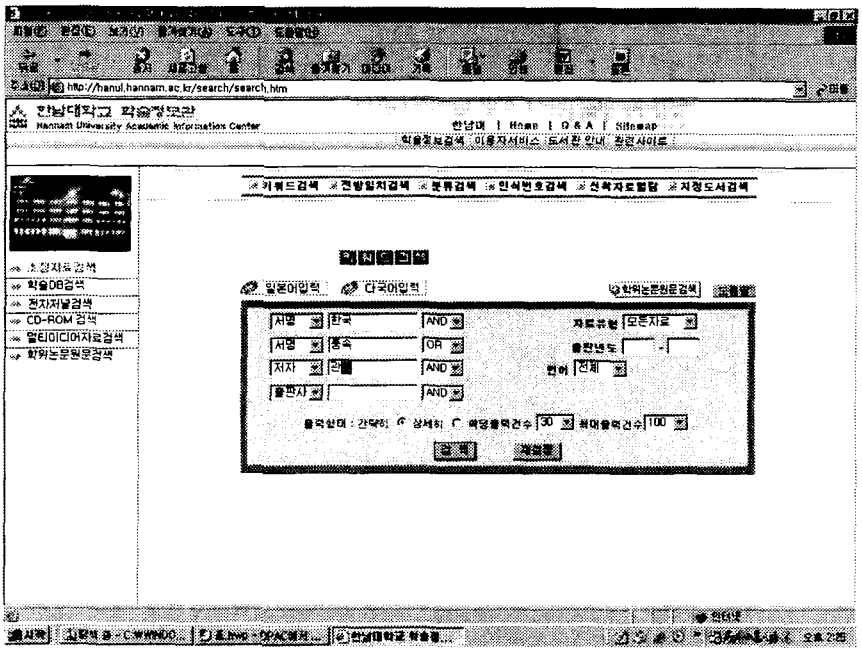
검색된 문헌은 서명순 간략정보 리스트를 출력하도록 하였다. 서명순 리스트에서 적합성을 판정하고, 서명에 출현하는 명사에 동그라미를 치고, 추가할 통제색인어를 표기한 후, SPSS에서 이진백터 문헌용어행렬표를 만들게 하였다. 클러스터링 모형으로 선정된 다이스와 자카드유사도를 사용하여 집단간 평균연결(SPSS 군집방법 대화상자에서는 "집단-간 연결")과 완전연결(SPSS에서는 "가장 먼 항목")으로 군집하도록 하였다. 군집분석 결과 생성된 덴드로그램에서 적합문헌에 체크표시를 한 후 가장 체크표시가 많은 군집을 선택하도록 하였다. 선택된 군집의 분류단계는 제한하지 않고 원하는 계층에서 원하는 군집을

선택하도록 하였다.

탐색과 군집분석을 끝낸 후 학생들은 보고서를 제출하였는데, 보고서에는 탐색질문, 탐색식, 적합성 여부가 표시된 서명순 리스트(이 리스트에는 문헌용어행렬표에 입력될 서명에 출현하는 단어들이 동그라미 쳐있고 추가할 통제어는 서명 옆에 부기되어 있다)와, 적합성과 최종 선택 군집이 마크된 덴드로그램, 초기탐색과 최종 선택된 군집의 정확률과 재현율, 군집분석 결과 선택된 클러스터에 대한 만족도 등과 함께 문헌용어행렬표 데이터까지 제출하도록 하였다.

제출된 보고서와 데이터는 연구자에 의해 그 정확도가 면밀히 조사되었다. 논리합 정규형의 문제점은 충분히 숙지되었고, 불연산자에 대해 충분한 이해가 있을 것으로 예상된 문헌정보학과 3학년 학생임에도 불구하고 여러 학생들은 여전히 탐색식 표현에 어려움을 겪고 있었다. 〈그림 1〉과 같은 MAESTRO 탐색화면에서 AND는 OR보다 우선하고 스페이스(space)는 AND기능을 하기 때문에 한국의 풍속이나 관습에 관한 문헌을 검색하기 위해서 탐색식은 "한국 풍속 OR 한국 관습"이 되어야 한다. 그러나 "한국 AND 풍속 OR 관습" 혹은 "풍속 OR 관습 AND 한국"으로 잘못 탐색하는 학생들이 있었다. 탐색자가 판정한 질문에 대한 문헌의 적합성은 연구자에 의해 리뷰되었고, 이의가 있는 경우는 탐색자와 논의하여 결정하였다.

탐색식이 잘못된 경우, 서명순 리스트와 덴드로그램에서 적합성 체크가 동일하지 않는 경우, 정확률과 재현율 계산이 잘못된 경우, 문헌 용어행렬표 데이터가 잘못된 경우 등과 같이 제출된 보고서에 문제가 있는 학생은 설



〈그림 1〉 한남대학교 OPAC 키워드검색 화면

명과 함께 보고서에 문제점을 표시하여 돌려주고 다시 제출하도록 하였다. 탐색질문과 검색식이 일치하지 않은 학생 중 논리합 정규형 탐색식은 이해하나 검색된 문헌이 너무 적어 확대탐색을 하였다고 한 4명의 학생은 옳은 탐색을 한 것으로 간주하였다. 다시 제출하지 않은 학생, 다시 제출하였음에도 여전히 문제가 있는 학생은 제외하고 총 16건의 탐색을 대상으로 클러스터링의 성능이 분석되었다.

5. 성능분석 척도

서명단어탐색의 성능은 정확률로, 클러스터링의 성능은 정확률(P), 재현율(R)과 함께 단순척도 E값으로 분석되었다. 사용된 재현율은 클러스터링 이전 서명단어탐색에서 검색된 적합문헌이 선택된 군집에 어느 정도 포함되는지를 표현하는 것으로 정의되었다.

$$\text{재현율} = \frac{\text{선택된군집내적합문헌수}}{\text{클러스터링이전서명단어탐색으로검색된적합문헌수}}$$

E값은 탐색질문으로 선택된 클러스터를 평가하는 연구에서 자주 사용되는 것(Enser 1985, Griffiths, Robinson, & Willett 1984, Tombros, Villa, & Van Rijsbergen 2002)으로 다음과 같다.

$$E = 1 - \{ (\beta^2 + 1) PR / (\beta^2 P + R) \}$$

매개변수 β는 정확률(P)와 재현률(R)의 비율을 나타내는 것으로 1과 0.5, 2가 주로 사용된다. 1은 P와 R의 중요도를 동등하게, 0.5는 P를 R보다 2배 더 중요하게, 2는 R을 P보다 2배 더 중요하게 취급하는 것으로, 최저의 E값을 갖는 클러스터가 가장 좋은 클러스터임을 의미한다.

6. 서명단어 탐색결과 분석

16개의 탐색은 총 759건의 문헌을 검색하였고 그 중 적합문헌은 273건이었다. 탐색결과를 50건 내외로 제한했기 때문에 많은 탐색이 자료유형을 단행본으로 제한하여 학위논문과 단행본도서를 검색하였다(한남대 시스템에서 학위논문은 자료유형이 단행본으로 정의되어 있음). 그러나 자료유형을 제한하지 않은 탐색에서는 학술지 기사논문까지 검색되었다. 문헌당 검색된 문헌 수는 최대 61, 최소 35, 평균 47.44건이었으며, 검색된 적합문헌 수는 최대 29, 최소 8, 평균 17.06건으로 시스템 정확률은 .3597이었다. 탐색당 정확률은 최고 .6905, 최소 .1481, 평균정확률은 .3707이었다(표 1 참조).

7. 클러스터링 성능 분석

학생들에게는 클러스터링의 계층단계를 무시하고 원하는 클러스터를 덴드로그램에서 bottom-up 방식으로 선택하도록 하였지만, 본 분석에서는 탐색자가 최적으로 선택한 클러스터를 최종 클러스터라고 정의하고, 최종 클러스터가 소속된 최상위 1단계 계층의 클러

〈표 1〉 초기탐색 결과 분석

검색된 문헌 총수	759개
검색된 적합문헌 총수	273개
탐색당 검색된 문헌 수(평균)	47.44개
탐색당 검색된 적합문헌 수(평균)	17.06개
시스템중심 정확률	.3597
이용자중심 정확률	.3707

스터(top cluster)는 1단계 클러스터로 정의하여, 두 클러스터 모두를 분석하였다.

탐색12의 완전연결 클러스터링에서 탐색자는 1단계에서 생성된 9개의 군집 중 두 개의 군집을 최종집단으로 선택하였다. 탐색12로 검색된 문헌 수가 45개이고 그 중 적합문헌은 15건인데, 1단계 계층클러스터링에서 9개의 군집으로 나뉘었다. 선택된 두 군집에는 각각 5건과 4건이 분류되었는데 모두 적합문헌이었으므로 선택된 2 군집 모두를 최적의 최종군집으로 인정하기로 하였다.

7. 1 1단계 클러스터의 성능

〈표 2〉는 1단계 계층클러스터링에서 생선된 군집 수와 선택된 1단계 클러스터의 정확률과 재현율을 클러스터링 기법과 유사도 기법별로 나타낸 것이다. 군집 수는 독립군집을 이루고 있는 1개의 문헌이 2개 이상일 때는 '기타' 군집으로 묶어 하나로 계산하였다. 모형 연구에서와 마찬가지로 집단간보다는 완전연결이 더 많은 군집으로 세분하였다. 완전연결에서 다이스와 자카드 유사도 간에는 차이가 없었으나, 집단간에서는 탐색12에서 자카드가 다이스보다 하나 더 많은 군집으로 분류하였다. 1단계 계층에서 평균 군집 수는 집단간-D 2.81 군집, 집단간-J 2.88 군집, 완전연결 4.69 군집이었다.

1단계 최적 클러스터의 평균정확률은 서명단어탐색보다 .0943(집단간-D), .0887(집단간-J), .1789(완전연결) 증가하여, 집단간-D .4650, 집단간-J .4594, 완전연결 .5496이었다. 유사도 계수와 클러스터링 기법과 관계없이 4방법 모두 클러스터링전과 비교하여 유의한 차이를 보였다(표 4 참조). 평균재현율은 집단간-D .9787, 집단간-J .9621, 완전연결 .7964였다. 이는 1단계 클러스터에서 선택된 클러스터에 들어있는 문헌 중 45.95%(집단간-J)~54.96%(완전연결)는 적합문헌이며, 서명단어로 검색된 전체 적합문헌의 79.64%(완전연결)~97.87%(집단간-D)가 이 클러스터에 들어있음을 의미한다. 집단간-D에서는 정확률이 25% 증가한 대신 재현율은 2.13% 감소하였고, 집단간-J에서는 정확률이 24% 증가한 대신 재현율이 3.79% 감소하였으며, 완전연결에서는 정확률이 48% 증가한 대신 재현율은 20.36% 감소하였다(표 4 참조).

1단계 최적 클러스터의 정확률과 재현율은 유사도 계수 간에는 차이가 없었으나, 클러스터링 기법 사이에는 차이가 있었다. 정확률은 완전연결이, 재현율은 집단간 평균연결이 더 높았으며, 재현율 간의 차이는 통계적으로 유

의한 수준이었다(표 6 참조).

E값은 β=0.5, 1, 2 모두에서 유사도값과 클러스터링 기법에 관계없이 모두 0.50이하의 E값을 보임으로써 선행연구(Enser 1985, Griffiths, Robinson, & Willett 1984, Tombros, Villa, & Van Rijsbergen 2002)에서보다 더 좋은 결과를 보였다. β=1에서는 4종류의 클러스터링이 비슷하나, β=0.5에서는 완전연결이 가장 좋은 클러스터를 생성하였고, β=2에서는 집단간이 가장 좋은 클러스터를 생성하였다. 그러나 그 차이는 통계적으로 유의한 수준은 아니었다(표 6 참조).

7. 2 최종클러스터 성능

덴드로그램에서는 bottom-up 방식으로 최적의 클러스터를 선택하였지만, 검색된 문헌을 계층 클러스터로 제공하는 Vivisimo와 같은 검색시스템에서는 Top-down 방식으로 최적의 클러스터를 선택하도록 하고 있다. Top-down 방식으로 최적의 클러스터를 선택한다고 할 때, 최종 클러스터가 선택되기까지 거치는 계층 분류단계는 모형연구에서와 같이 집단간 평균연결이 완전연결보다, 다이스가자카드보다 더 계층적이다(표 3참조). 집단간(D)은 평균 5.44단계, 집단간(J)은 5.25단계, 완전연결(D)은 3.19단계, 완전연결(J)은 3.13 계층단계에서 최적의 클러스터가 선택되었다. 집단간 평균연결에서는 완전연결보다 평균 2단계 더 하위계층에서 최종군집이 선택되었으나, 탐색9에서는 10단계 계층에서, 탐색2와 탐색5, 탐색6에서는 8단계 계층에서 최종군집이 선택되었다. 원하는 최종군집을 얻기까지 너무 많은 계층 분류단계를 필요로 하였다. 완전연결 클러스터링에서는 두 탐색(탐색10과 탐색12)이 1단계 계층분류에서 최종클러스터를 선택하였다.

최종클러스터의 정확률은 서명단어탐색보다 .3934(집단간), .4196(완전연결-D), .4270(완전연결-J) 향상되어, 집단간 .7641, 완전연결-D .7903, 완전연결-J .7977이었다. 재현율은 집단간 .7041, 완전연결 .5893이었다. 이는 최적의 클러스터에 있는 문헌의 76.41%(집단간 평균연결)~79.77%(완전연결)는 적합문헌이고, 서명단어탐색으로 검색된 전체 적합문헌의 58.93%(완전연결)~70.41%(집단간 평균연결)가 이 클러스터로 검색됨을 의미한다. 집단간 평균연결에서는 정확률이 106% 증가한 대신 재현율은 29.59% 감소하였고, 완전연결-D에서는 정확률이 113% 증가한 대신 재현율이 41.07% 감소하였고, 완전연결-J에서는 정확률이 115% 증가한 반면 재현율이 41.07% 감소하였다(표 4 참조).

최종 클러스터에서도 정확률과 재현율은 유사도 계수간에는 차이가 없었고, 클러스터링 기법간에는 완전연결이 정확률 측면에서, 집단간 평균연결이 재현율 측면에서 효과적이었으나 유의한 수준은 아니었다.

최적으로 선택된 클러스터의 E값은 1단계에서보다 더 낮아져서 보다 만족스러운 성능을 보였다(표 5 참조). β=0.5, 1, 2 모두에서 집단간이 완전연결보다 우수하였으나 통계적으로 유의한 수준은 아니었다(표 6 참조).

〈표 2〉 1단계 클러스터링 성능

탐색	군집기법	군집 수	선택된 1단계 군집				클러스터링 이전 초기탐색		
			군집내 문헌 수	군집내 적합문헌 수	P	R	검색문헌 수	적합문헌 수	P
1	집단간	6	44	22	.5000	.9565	61	23	.3770
	완전연결	10	21	17	.8095	.7391			
2	집단간	2	37	18	.4865	1.0000	43	18	.4186
	완전연결	5	20	12	.6000	.6667			
3	집단간	2	43	15	.3488	1.0000	49	15	.3061
	완전연결	4	39	14	.3590	.9333			
4	집단간	2	39	20	.5128	1.0000	41	20	.4878
	완전연결	3	34	19	.5588	.9500			
5	집단간	3	45	8	.1778	1.0000	49	8	.1633
	완전연결	4	37	7	.1892	.8750			
6	집단간	2	38	20	.5263	1.0000	41	20	.4878
	완전연결	4	22	12	.5455	.6000			
7	집단간	3	36	15	.4167	1.0000	48	15	.3125
	완전연결	5	22	13	.5909	.8667			
8	집단간	2	44	21	.4773	1.0000	58	21	.4375
	완전연결	6	31	14	.4516	.6667			
9	집단간	2	49	8	.1613	1.0000	50	8	.1600
	완전연결	3	40	8	.2000	1.0000			
10	집단간	3	24	23	.9583	.9200	50	25	.5000
	완전연결	3	25	23	.9200	.9200			
11	집단간	3	29	19	.6552	1.0000	35	19	.5400
	완전연결	3	29	19	.6552	1.0000			
12	집단간(D)	4	27	13	.4815	.8667	45	15	.3333
	집단간(J)	5	23	9	.3913	.6000			
	완전연결	9	9	9	1.0000	.6000			
13	집단간	2	41	29	.7073	1.0000	42	29	.6905
	완전연결	4	17	15	.8824	.5172			
14	집단간	5	21	11	.5238	.9167	53	12	.2264
	완전연결	7	13	7	.5385	.5833			
15	집단간	2	53	8	.1509	1.0000	54	8	.1481
	완전연결	2	53	8	.1509	1.0000			
16	집단간	2	48	17	.3542	1.0000	50	17	.3400
	완전연결	3	41	14	.3415	.8235			
평균	집단간(D)	2.81	38.63	16.69	.4650	.9787	47.44	17.06	.3707
	집단간(J)	2.88	38.38	16.44	.4594	.9621			
	완전연결	4.69	28.31	13.19	.5496	.7964			

〈표 3〉 최종 클러스터의 성능

탐색	군집방법	최종클러스터 계층단계	선택된 최종 군집				적합문헌 총수
			군집문헌 수	군집 적합문헌 수	P	R	
1	집단간	5	18	14	.7778	.6087	23
	완전연결	2	12	12	1.0000	.5217	
2	집단간(D)	8	21	15	.7143	.8333	18
	집단간(J)	6	21	15	.7143	.8333	
	완전연결	4	15	12	.8000	.6667	
3	집단간	4	11	9	.8182	.6000	15
	완전연결(D)	5	7	5	.7143	.3333	
	완전연결(J)	5	6	5	.8333	.3333	
4	집단간	5	24	17	.7083	.8500	20
	완전연결	4	19	15	.7895	.7500	
5	집단간	8	10	5	.5000	.6250	8
	완전연결	4	17	6	.3529	.7500	
6	집단간	8	13	10	.7692	.5000	20
	완전연결	6	9	8	.8889	.4000	
7	집단간	5	19	14	.7368	.9333	15
	완전연결(D)	4	16	12	.7500	.8000	
	완전연결(J)	3	16	12	.7500	.8000	
8	집단간	7	15	10	.6667	.4762	21
	완전연결	3	11	8	.7273	.3810	
9	집단간	10	6	6	1.0000	.7500	8
	완전연결	4	6	6	1.0000	.7500	
10	집단간	1	24	23	.9583	.9200	25
	완전연결	1	24	23	.9583	.9200	
11	집단간	4	14	14	1.0000	.7368	19
	완전연결	2	16	14	.8750	.7368	
12	집단간(D)	3	5	5	1.0000	.3333	15
	집단간(J)	2	5	5	1.0000	.3333	
	완전연결	1	5	5	1.0000	.3333	
13	집단간	5	27	24	.8889	.8276	29
	완전연결	1	17	15	.8824	.5172	
14	집단간	4	12	9	.7500	.7500	12
	완전연결	2	8	7	.8750	.5833	
15	집단간	5	28	7	.2500	.8750	8
	완전연결	5	26	6	.2308	.7500	
16	집단간	5	16	11	.6875	.6471	17
	완전연결	3	5	4	.8000	.2353	
평균	집단간(D)	5.44	16.44	12.06	.7641	.7041	17.06
	집단간(J)	5.25	16.44	12.06	.7641	.7041	
	완전연결(D)	3.19	13.31	9.88	.7903	.5893	
	완전연결(J)	3.13	13.25	9.88	.7977	.5893	

〈표 4〉 선택된 1단계/최종단계 클러스터의 성능 변화

		집단간(D)	집단간(J)	완전연결(D)	완전연결(J)
서명단어탐색 정확률P		.3707	.3707	.3707	.3707
1단계 클러스터	정확률P	.4650	.4594	.5496	.5496
	P향상	+.0943	+.0887	+.1789	+.1789
	상승률(%)	25	24	48	48
	t검증(t값)	3.062	2.892	3.679	3.679
	유의도	.008**	.011*	.002**	.002**
	재현율R	.9787	.9621	.7964	.7964
	R감소	-.0213	-.0379	-.2036	-.2036
최종단계 클러스터	정확률P	.7641	.7641	.7903	.7977
	P향상	+.3934	+.3934	+.4196	+.4270
	상승률(%)	106	106	113	115
	t검증(t값)	8.437	8.437	8.436	8.510
	유의도	.000***	.000***	.000***	.000***
	재현율R	.7041	.7041	.5893	.5893
	R감소	-.2959	-.2959	-.4107	-.4107

〈표 5〉 클러스터링별 평균 E값

	β	집단간(D)	집단간(J)	완전연결(D)	완전연결(J)
1단계	0.5	.490	.497	.441	.441
	1	.397	.406	.406	.406
	2	.240	.252	.330	.330
최종단계	0.5	.269	.269	.295	.291
	1	.296	.296	.368	.367
	2	.305	.305	.406	.405

〈표 6〉 집단간 평균연결과 완전연결간의 차이 검증

		다이스				자카드			
		평균		t	유의확률 (양쪽)	평균		t	유의확률 (양쪽)
		집단간	완전연결			집단간	완전연결		
1단계	P	.4650	.5496	-1.009	.321	.4594	.5496	-1.075	.291
	R	.9787	.7963	4.180	.000*	.9621	.7963	3.361	.002*
	E(0.5)	.490	.441	.651	.520	.497	.441	.737	.467
	E(1)	.397	.406	-.132	.896	.406	.406	.004	.997
	E(2)	.240	.330	-1.812	.080	.252	.330	-1.499	.144
최종	P	.7641	.7903	-.357	.724	.7641	.7977	-.459	.649
	R	.7041	.5893	1.714	.097	.7041	.5893	1.714	.097
	E(0.5)	.269	.295	-.420	.677	.269	.291	-.363	.719
	E(1)	.296	.368	-1.238	.225	.296	.367	-1.220	.232
	E(2)	.305	.406	-1.692	.101	.305	.405	-1.688	.102

8. 탐색결과 해석 및 제한점

본 연구의 결과를 요약하면 다음과 같다.

1) 1단계에서 선택된 클러스터와 최종단계에서 선택된 클러스터는 유사도값과 클러스터링 기법에 관계없이 모두 0.50이하의 E값을 보임으로써 선행연구(Enser 1985, Griffiths, Robinson, & Willett 1984, Tombros, Villa, & Van Rijsbergen 2002)에서보다 더 좋은 결과를 보였다. E척도에서 정확률을 강조한 경우(β=0.5), 1단계에서는 완전연결이, 최종단계에서는 집단간이 더 좋은 클러스터를 생성하였으며, 재현율을 강조한 측면(β=2)에서는 1단계와 최종단계 모두에서 집단간이 더 우수한 클러스터를 생성하였으나 통계적으로 유의한 수준은 아니었다.

2) 최적의 클러스터의 정확률은 1단계에서는 .0887(집단간-J)~.1789(완전연결) 향상되었고, 최종단계에서는 .3934(집단간)~.4270(완전연결) 향상되었다. 그러나 재현율은 1단계에서 .0213(집단간-D)~.2036(완전연결) 감소되었고, 최종단계에서는 .2959(집단간)~.4107(완전연결-J) 감소되었다. 정확률과 재현율 측면에서 다이스와 자카드 계수 사이에는 차이가 없었으나, 클러스터링 기법간에는 약간의 차이가 있었다. 1단계와 최종단계 모두에서 집단간이 완전연결보다 재현율이 .1741(1단계), .1148(최종단계) 높았으며, 정확률은 1단계에서 완전연결이 집단간보다 약.0874 높았다. 최종단계에서 정확률은 비슷하였다. 1단계에서 완전연결에 비해 집단간 평균연결의 높은 재현율은 다이스, 자카드 모두에서 유의한 차이를 보였다.

3) 전체 16개의 탐색 모두에서 클러스터링 기법이나 유사도 계수와 상관없이 선택된 1단계 군집에는 50%이상의 적합문헌이 존재하였다. 1단계에서 선택된 군집에는 평균 79.64%(완전연결)~97.87%(집단간-다이스)의 적합문헌이 존재하였으며, 최종단계에서 선택된 군집에도 평균 58.93%(완전연결)~70.41%(집단간)의 적합문헌이 존재하였다.

4) 계층구조 측면으로 분석하면, 완전연결이 집단간보다 더 적정수의 클러스터를 생성하고, 보다 짧은 계층단계를 필요로 하였다. 1단계 클러스터링으로 생성된 클러스터는 집단간-다이스가 2.81개, 집단간-자카드가 2.88개, 완전연결은 4.69개였다. 최적의 최종군집을 얻기까지의 계층 단계는 집단간이 평균 5.35 단계를 필요로 하였으나 완전연결은 평균 3.13 단계를 필요로 하였다. 집단간에서 5개의 탐색은 최종집단을 얻기까지 7단계 이상의 계층단계를 필요로 하였다. 심지어 10단계에서 최종집단이 선택되기도 하였다(탐색9). 그러나 같은 탐색에서 같은 결과를 완전연결은 4단계에서 생성하였다.

결론적으로 탐색결과의 클러스터링에서 1단계와 최종단계에서 선택된 클러스터는 만족할 만 하였으며, 두 유사도 계수 다이스와 자카드 사이에는 정확률, 재현율, E값 측면에서 차이가 없었으나, 두 클러스터링 기법간에는 약간의 차이가 있었다. 정확률 측면에서는 완전연결 기법이, 재현율 측면에서는 집단간 평균연결이 보다 우수하였으나, 통계적으로는 1단계에서 집단간 평균연결의 재현율이 더 높은 것을 제외하고는 집단간 차이는 유의하지 않았다. 집단간 평균연결은 너무 긴 계층단계를 필요로 하기때문에 탐색의 효율성 관점에

서 뒤떨어 진 것으로 보인다. 탐색자가 컴퓨터 앞에 앉아 생성된 클러스터를 단계별로 클릭하여 하위계층의 보다 작은 소집단을 선택한다고 할 때, 5번 이상의 심지어는 10번의 클러스터 선택은 효율적이지 못할 것 같다. 이는 문헌집단의 동질성때문인 것으로 보인다. 모형연구에서 사용된 문헌정보학 실험집단은 특정 질문으로 검색된 검색결과문헌보다 동질성이 더 낮아, 문헌정보학 문헌에서 1단계 클러스터링은 같은 수동통합 이진백터를 사용했을 때 집단간 평균연결-다이스는 9개, 집단간 평균연결-자카드는 13개, 완전연결은 20개의 군집을 생성하였다. 그러나 본 연구에서 1단계 클러스터링으로 생성된 클러스터는 집단간 평균연결-다이스가 2.81개, 집단간 평균연결-자카드가 2.88개, 완전연결은 4.69개였다. 집단간 평균연결이 동질성이 높은 문헌집단에서는 바람직하지 않는 것으로 보인다.

본 연구에서는 문헌-용어행렬데이터를 수동으로 입력하는데 편의성과 이용자들이 필요로 하는 적합문헌수에 비교하여 지나치게 많은 문헌을 검색하는 OPAC의 특성 때문에 탐색결과를 50건 내외로 제한시켰다. 이것은 서명단어 탐색의 정확률을 높이는데 영향을 끼쳤겠지만, Tombros, Villa & Van Rijsbergen(2002)의 결론처럼 탐색결과를 클러스터링하는데 클러스터링 기법간의 성능에는 영향을 끼쳤을 것으로 생각지 않는다. 그러나 클러스터링 대상 문헌수는 클러스터링에 영향을 끼치는지 후속 연구해 볼 수 있을 것이다.

끝으로 국립중앙도서관을 위시하여 많은 시스템에서 사용되고 있는 〈그림 1〉과 같은 탐색 인터페이스의 타당성은 재고해 볼 필요가 있다. AND는 OR보다 우선할 뿐 괄호를 사용하여 불연산자의 우선순위를 조정할 수 없는 시스템에서 (A OR B) AND C와 같은 탐색요구는 A AND B OR A AND C와 같은 논리형 정규형탐색식으로 탐색해야 한다는 사실을 아는 일반 이용자는 많지 않을 것이기 때문이다.

참 고 문 헌

노정순. 2004. OPAC에서 자동분류 열람을 위한 계층 클러스터링 연구. 『정보관리학회지』, 21 발표 예정

Barbuto, D. M. & E. E. Cevallos. 1991. "End-user Searching: program review and future prospects." *RQ*, 31(winter): 214-227.

Carlyle, Allyson. 1996. "Ordering author and work records: an evaluation of collection in online catalog displays." *Journal of the American Society for Information Science*, 47(7): 538-554.

Croft, W. B. 1980. "A Model of cluster searching based on classification." *Information Systems*, 5: 189-195.

Enser, P. G. B. 1985. "Automatic classification of book material repre-

sented by back-of-book index." *Journal of Documentation*, 41(3): 135-155.

Garland, K. 1983. "An Experiment in automatic hierachical document classification." *Information Processing and Management*, 19(3): 113-120.

Griffiths, A., L. A. Robinson, and P. Willett. 1984. "Hierarchic agglomerative clustering methods for automatic document classification." *Journal of Documentation*, 40(3): 175-205.

Hearst, M. & J. Pederson. 1996. "Reexamining the cluster hypothesis: Scatter/Gather on retrieval results." *Proceedings of the 19th Annual International ACM SIGIR Conference of Research and developoment in Information Retrieval*: 76-84.

Larson, R. 1986. *Workload Characteristics and Computer System Utilization in Online Library Catalog*. Ph.D. Diss., University of California, Berkeley.

_____. 1991. "The Decline of subject searching: long-term trends and patterns of index use in an online catalog." *Journal of the American Society for Information Science*, 42(3): 197-215

Leouski, A. and J. Allan. 1998. "Evaluating a visual navigation system for a digital library." *Proceedings of the second European Conference of Research and Technology for Digital Libraries, Heraklion, Greece*: 535-554.

Lynch, C. A. 1989. "Large database and multiple database problems in online catalogs." In *OPAC and Beyond: Proceedings of a Joint Meeting of the British Library, DBMIST, and OCLC*, Dublin, Ohio: OCLC, 1989.

Peter, T. A. and M. Kurth. 1991. "Controlled and uncontrolled vocabulary subject searching in an academic library online catalog." *Information Technology and Libraries*, 10(Sept.): 201-211.

Preece. 1973. "Clustering as an output option." *Proceedings of the American Society for information Science*, 10: 189-190.

Roussinov, D. & Chen, H. 2001. "Information navigation on the web by clustering and summarizing query results." *Information processing & Management*, 37: 789-816.

Salton, G. 1971. *The SMART Retrieval System-Experiments in Automatic Document Retrieval*. Englewood Cliffs, NJ: Prentice-Hall.

Silverstein, D. and J. O. Pedersen. 1997. "Almost-constant-time clustering of arbitary corpus subsets." *Proceeding of the 20th annual ACM SIGIR conference, Philadelphia, PA*: 60-66.

Tombros, A., R. Villa, and C. J. Van Rijsbergen. 2002. "The Effectiveness of query-specific hierarchic clustering in information retrieval." *Information Processing and Management*, 38(4): 559-582.

Vivisimo(http://www.vivisimo.com).

Wibereley, S. E., R. A. Daugherty, and J. A. Danowsky. 1995. "User persistence in displaying online catalog posting: LUIS." *LRTS*, 39(3): 247-264.

Willett, P. 1985. "Query specific automatic document classification." *Internation Forum on Information and Documentation*, 10(2): 28-32.

Zamir, O. and Etzioni, O. 1998. "Web document clustering: A feasibility demonstration." *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 46-54.