OPAC에서 자동분류 열람을 위한 계층 클러스터링 연구

Hierarchic Document Clustering in OPAC

노정순(Jung-Soon Ro)*

초 록

본 연구는 OPAC에서 계층 클러스터링을 응용하여 소장자료를 계층구조로 분류하여 열람하는데 사용될 수 있는 최적의 계층 클러스터링 모형을 찾기 위한 목적으로 수행되었다. 문헌정보학 분야 단행본과 학위논문으로 실험집단을 구축하여 다양한 색인기법(서명단어 자동색인과 통제어 통합색인)과 용어가중치 기법(절대빈도와 이진빈도), 유사도 계수(다이스, 자카드, 피어슨, 코싸인, 제곱 유클리드), 클러스터링 기법(집단간 평균연결, 집단내 평균연결, 완전연결)을 변수로 실험하였다. 연구결과 집단간 평균연결법과 제곱 유클리드 유사도를 제외하고 나머지 유사도 계수와 클러스터링 기법은 비교적 우수한 클러스터를 생성하였으나, 통제어 통합색인을 이진빈도로 가중치를 부여하여 완전연결법과 집단간 평균연결법으로 클러스터링 하였을 때 가장 좋은 클러스터가 생성되었다. 그러나 자카드 유사도 계수를 사용한 집단간 평균연결법이 십진구조와 더 유사하였다.

ABSTRACT

This study is to develop a hiararchic clustering model for document classification and browsing in OPAC systems. Two automatic indexing techniques (with and without controlled terms), two term weighting methods (based on term frequency and binary weight), five similarity coefficients (Dice, Jaccard, Pearson, Cosine, and Squared Euclidean), and three hierarchic clustering algorithms (Between Average Linkage, Within Average Linkage, and Complete Linkage method) were tested on the document collection of 175 books and theses on library and information science. The best document clusters resulted from the Between Average Linkage or Complete Linkage method with Jaccard or Dice coefficient on the automatic indexing with controlled terms in binary vector. The clusters from Between Average Linkage with Jaccard has more likely decimal classification structure.

키워드 : 온라인 목록, 문헌 클러스터링, 계층 클러스터링, 자동분류, 열람, 유사도 계수
OPAC, document clustering, hierarchic clustering, automatic classification, browsing, similarity coefficient

---

* 한남대학교 문헌정보학과 교수(jsr@mail.hannam.ac.kr)

■ 논문접수일 : 2004. 2. 16
■ 게재확정일 : 2004. 3. 10

1 서 론

1.1 연구목적

온라인 도서관목록(Online Public Access Catalog: OPAC)에서 너무 많은 문헌의 검색은 특히 그 속에 포함된 너무 많은 부적합문헌 때문에, 높은 정확률을 요구하는 이용자의 OPAC에 대한 요구를 만족시키지 못하고 있다. 저자나 서명을 알지 못하고 수행하는 주제탐색에서 이용자는 수십 권 혹은 수백 권의 문헌을 검색하지만[1], 이용자가 실제 디스플레이한 문헌은 10권 내외였다.[2] 부적합 문헌의 검색과 낮은 정확률의 문제는 주제탐색 중 주제명단어 탐색보다는 서명단어 탐색에서 더 심각하다고 보고되었다.[3]

너무 많은 문헌을 검색하는 반면에 한편의 문헌도 검색하지 못하는 제로포스팅(Zero Posting)의 문제는 OPAC을 더욱 어렵게 한다. Zink(1991)는 전체 6,118 탐색 중 27.81%의 제로포스팅을, Blecic et al.(1999)은 전체 39,421 탐색 중 35.05%의 제로포스팅을, Cooper & Chen(2001)은 전체 905,970 탐색 중 17.85%의 제로포스팅을 보고하였다.

OPAC 탐색의 너무 많은 문헌의 검색과 낮은 정확률, 높은 제로포스팅의 문제에서 상대적으로 우수한 통제어 주제탐색이 불가능한 대부분 우리나라 한글자료의 OPAC에서는 분류번호 탐색과 서명단어 탐색으로 OPAC의 주제탐색 성능향상을 위해 노력하고 있다. 그러나 이용자가 분류번호를 정확히 알고 있지 못하기 때문에 분류번호 탐색이 거의 이용되지 못하는 현실에서 분류번호 열람은 서명단어 탐색의 단점을 보완해 줄 수 있는 대안이 될 수 있다.

한남대학교의 소장자료 검색시스템인 OROM의 MAESTRO는 분류번호열람 기능을 분류번호탐색 대신 제공하나, 세분된 소주제로까지 안내되지 못할 뿐만 아니라 DDC의 구조와 미분류된 자료 때문에 높은 정확률은 물론 높은 재현율을 위한 탐색에도 도움을 주지 못하고 있다. 문헌정보학(020)의 경우 DDC는 미배정된 두 개의 목(024와 029)을 제외하고 8개의 목으로 주제를 배정하고 있으나, 한남대학교 소장목록에서 문헌정보학 분야 실제 도서는 5개의 소주제로 열람되고있다. 문헌정보학관계 한국어 단행본 총 747건 중 022와 023으로 안내된 자료는 한건도 없고, 도서관과 사회 19건, 도서관운영 390건, 특수도서관 4건, 일반도서

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

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

3) Peters & Kurth(1991)연구에서 주제명탐색은 평균 19건의 문헌을 검색한 반면 서명단어 탐색은 105건을 검색하였다. Carlyle(1996)의 연구에서 서명단어탐색의 정확률은 주제명 키워드탐색보다 더욱 낮아 정확률은 중앙치 7%이었고, 90개 탐색 중 72개 탐색(82%)이 19%이하의 정확률을 보였다.

관 86건, 독서지도 137건으로 안내될 뿐이다. 도서관운영 관련자료가 390건(전체의 52%)이나 되지만 소주제로 세분하여 분류열람을 제공하지 못한다(<표 1> 참조). 뿐만 아니라 단행본과 학위논문, 학술지 기사논문까지 통합한 종합시스템이지만 상당수의 학위논문과 모든 학술지 기사논문은 분류번호로 분류되지 않았기 때문에 분류열람으로는 검색이 불가능하다. 이와 같은 문제점을 해결하기 위한 방안으로 분류번호 대신 자료를 유사도값으로 클러스터링하여 생성된 클러스터를 열람하고 원하는 클러스터를 선택할 때까지 주제를 계층구조로 전개하는 브라우징시스템이 OPAC에서, 특히 자료의 소장위치가 필요없는 가상도서관에서 대안책이 될 수 있을 것이다.

본 연구는 OPAC에서 계층 클러스터링을 응용하여 소장자료를 계층구조로 브라우징하는 시스템의 가능성을 연구하기 위하여 수행되었다. 문헌클러스터링은 원래 효율적인 파일조직기법으로 연구되었으나, 문헌과 지식의 분류, 데이터베이스 또는 검색결과의 브라우징이나 분류 등에 응용되고 있다. 분류를 위한 계층클러스터링은 문헌집단이나 분류자질, 유사도계수, 클러스터링기법 등에 따라 그 결과가 다르게 나타난다. 학술지 논문초록이나 신문기사, 웹문서의 클러스터링에서 역문헌빈도, 자카드나 코싸인 유사도가 좋은 성능을 보이는 것으로 보고되고 있으나, OPAC에서 문헌의 대용물인 서명은 학술논문이나 신문, Web문서의 초록이나 전문과 비교하여 분류자질이 되는 색인어 수가 매우 작고, 단어의 출현빈도가 매우 낮기 때문에 OPAC에 적합한 특히 한글 OPAC문헌에 적합한 클러스터링 모형을 찾는 연구가 필요하다.

본 연구는 클러스터링에 의한 분류가 DDC에 의한 분류보다 실제 도서를 보다 균형적으로 분류할 것이라는 가정 하에 한글 OPAC문헌에 적합한 정적 클러스터링 모형을 찾기 위한 목적으로 수행되었다.

〈표 1〉 한남대학교 분류번호 열람: DDC(020)번호와 문헌수 (2003. 8 현재)

Maestro 분류번호 열람	해당 DDC번호	한글 소장자료	한글 2000년 이후	실험DB
도서관 정보학	020~028	747	166	175
도서관과 사회	021	19	0	6
도서관 운영	025	390	87	87
특수도서관	026	4	0	3
일반도서관	027	86	27	27
독서지도	028	137	35	35

1.2 연구과제

1) 한글문헌을 자동 분류하는데 가장 적합한 클러스터링 모형은 무엇인가?
2) 자동색인과 수동색인은 클러스터링에 영향을 끼치는가?
3) 클러스터링에 의한 한글문헌 분류는 DDC보다 문헌을 더 균형적으로 분류하는가?
4) 자동 추출된 클러스터명은 적합 클러스터를 분별할 분별력을 가지는가?

2 선행연구

클러스터(군집) 분석이란 다차원공간에서 유사한 객체집단을 인지할 수 있게 하는 기법이다. 지난 30년 동안 클러스터링은 문헌클러스터링과 용어클러스터링 두 방법으로 많은 연구가 이루어졌다. 문헌클러스터링에 대한 연구는 파일조직의 한 기법으로 전체 파일을 탐색하는 대신 정보요구와 관련된 문헌클러스터만을 탐색하여 검색효율(efficiency)을 향상시키기 위한 목적으로 알고리즘을 연구하는 것에서 시작하였으나 검색효과(effectiveness)를 위한 계층 클러스터링 연구로 이어졌다. 1980년대에 들어와서 문헌클러스터링은 파일조직보다는 문헌과 지식의 분류에 클러스터링을 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. 계층 클러스터링에서 탐색질문과 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 보고하였다.

문헌집단 전체를 대상으로 클러스터링하는 정적 클러스터링에 관한 연구 중 단행본 도서를 대상으로 연구한 연구는 주로 클러스터링으로 생성된 클러스터를 분류번호와 비교하여 평가하였다. 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 Willett 클러

스터링 기법으로 서명과 목차, 권말색인에 출현하는 단어를 분류자질로 삼아 테스트하였다.

데이터베이스를 브라우징하는데 클러스터링을 응용한 대표적인 연구로는 Scatter/Gether 시스템과 Webcluster 프로젝트가 있다. Scatter/Gether(Cutting et al. 1992)에서는 이용자가 탐색질문을 탐색하는 대신 브라우징을 통해 적합문헌을 찾는다. DB 문헌집단을 몇 개의 클러스터로 나누어(scatter) 클러스터의 요약정보를 이용자에게 제공하면, 이용자는 요약정보를 근거로 몇 개의 클러스터를 선택하고, 선택된 클러스터는 함께 합쳐져(gether) 소집단을 형성하고, 시스템은 다시 이 소집단을 대상으로 클러스터링하여 클러스터를 제공하고, 이렇게 정보요구에 가장 적합한 적정량의 문헌을 얻을 때까지 클러스터링을 반복하는 시스템이다. Webcluster 프로젝트(Mechkour, Harper & Muresan 1998)는 웹탐색을 중재하기 위한 목적으로 source 문헌집단을 클러스터링하여 이용자가 브라우징하게 하고, 이용자가 선택한 클러스터를 탐색질문 대신으로 사용하여 Web 탐색엔진(targer집단)으로 보내 Web문서를 검색하게 하는 시스템을 개발하는 것을 목적으로 하였다.

탐색결과에 클러스터링을 적용하는 연구는 Preece(1973)를 시작으로 Willett(1985), Hearst & Pederson(1996), Leouski & Allan (1998), Tombros, Villa, & Van Rijsbergen (2002)의 연구로 이어졌다. 탐색결과에 클러스터링을 적용한 동적 클러스터링은 정적 클러스터링에 비하여 나쁘지 않았다(Willett 1985, Tombros, Villa, & Van Rijsbergen 2002). 적합군집내의 문헌 n개의 순위는 탐색결과 얻은 전체 상위 n개의 순위보다 우수하였고(Hearst & Pederson 1996), 탐색결과 얻은 Top 50개의 문헌은 다차원공간에서 적합문헌끼리 인접하였다(Leouski와 Allan 1998). 탐색결과 얻은 top n개의 문헌을 클러스터링할 때 크러스터링의 성능은 문헌수 n에 관계없이 집단평균이 가장 좋았고 단일연결이 가장 나쁜 것으로 보고되었다(Tombros, Villa & Van Rijsbergen 2002).

탐색결과를 클러스터링하여 시각적으로 보여주거나 브라우징할 수 있도록 하는 연구는 웹문헌에서도 수행되었고(Zamir & Etzioni 1998, Roussinov & Chen 2001), 탐색결과의 클러스터링은 Altavista, Alltheweb, Vivisimo, QueryServer 등과 같은 여러 상업적인 웹탐색엔진에서 사용되고 있다.

정적 문헌클러스터링에 대한 국내 연구로는 한승희 & 이재윤(1999)이 50건의 신문기사를 대상으로 클러스터링에 사용되는 다양한 유사계수 간의 성능과 연관성을 연구하였고, 정영미 & 이재윤(2001)은 신문기사 1,020건과 학술논문 초록 990건을 대상으로 다양한 용어가중치 공식과 자질 축소방법을 사용하여 계층클러스터링 완전연결기법과 비계층클러스터링 K-means의 성능을 수동분류번호와의 유사도값으로 테스트하였다. K-means보다는 완전연결기법의 성능이 더 우수하였고, 분류자질은 전체의 7.66%만 사용했을 때도 성능저하가 크지 않았다.

3 실험 설계

OPAC 데이타는 한남대학교 분류열람에서 문헌정보학 분야를 대상으로 2000년 이후 출간된 한국어 문헌을 단행본으로 제한하여 추출하였다. 2000년 이후에 출간된 문헌 166권에는 도서관과 사회(021)와 특수도서관(026) 관련 주제가 없었으므로 이 두 세목에 대해서는 1990년 이후 출판된 도서로 탐색을 확대하여 각각 6권과 3권을 추가 검색하여 총 175건의 문헌으로 실험 DB를 구축하였다. 한남대 OPAC에서는 학위논문도 단행본 자료유형으로 목록되기 때문에 검색된 175건의 단행본에는 13건의 학위논문이 포함되었다.

3.1 모형을 위한 변수

본 연구에서는 색인기법과 가중치부여방법, 유사도기법, 군집방법이 변수로 사용되었다.

분류자질로서의 색인기법은 자동색인과 수동통합색인이 사용되었다.

자동색인 : MARC 245필드만을 대상으로 서명에 출현하는 명사만을 추출하였다. 복합명사는 복합명사와 함께 복합명사를 이루는 단일명사로 분리하여 색인하였다. 예를 들어 "정보검색"은 "정보검색"과 함께 "정보", "검색"으로 색인함으로써 "정보검색"이나 "정보 AND 검색"으로도 검색될 수 있게 하였다. 명사 중 "~에 관한 연구"에서처럼 서명 끝에 출현하는 "연구"는 불용어 처리하였지만 "학술연구동향"에서 "연구"는 색인하였다.

수동통합색인 : 수동통합색인은 서명에 출현하는 단어 외에 색인자가 추가하는 색인어를 통합하였다. 추가된 색인어는 본 연구를 위해 연구자가 부여하였다. 분류번호 대신 색인어를 기반으로 하는 클러스터링을 목적으로 색인어를 부여하는 것이기 때문에 동의어나 상위어 통제를 가하였다. "DDC의 이해"에는 "분류"가, "전자도서관 구축"에는 "디지털"이 부여되었다. "기록관", "역사관", "보존소"라는 용어가 출현하는 서명에는 "아카이브"를, "편목"라는 용어가 출현하는 서명에는 "목록"을 통제어로 부여하였다. "활용법", "관리론", "독서론", "독서술", "독서법"과 같은 용어는 "활용", "관리", "독서"와 같이 스테밍한 용어를 색인어로 추가하였다.

용어의 가중치는 자동색인화일에는 절대빈도와 이진빈도가, 수동통합색인에서는 2진빈도만 사용되었다. 학술논문의 초록이나 신문기사, 웹문서에 비해 OPAC에서 서명 단어수와 단어의 출현빈도수는 현저하게 적기 때문에 절대빈도와 이진빈도는 큰 차이가 없을지도 모른다. 그러나 OPAC의 이 특성 때문에 초록이나 본문의 클러스터링과는 다른 OPAC용 클러스터링 기법이 필요한지를 연구하기 위하여 초록이나 전문의 클러스터링에서 주로 사용되는 코싸인 유사도나 피어슨 상관계수를 OPAC에서도 테스트할 필요가 있었다. 코싸인 유사도나 피어슨 상관계수는 변량의 값이 연속형으로 표현된 경우에 사용할 수 있으므로 가중치를 절대빈

도로 부여할 필요가 있었다.

유사도 기법으로는 절대빈도로 가중치를 표현한 경우는 코싸인과 피어슨, 제곱 유클리드 계수가 사용되었고, 이진빈도 백터에서는 자카드(Jaccard)와 다이스(Dice), 제곱 유클리드 계수가 사용되었다.

군집방법으로 완전연결법(Complete Linkage Method)과 군집간 평균연결법(Between Average Linkage Method), 군집내 평균연결법(Within Average Linkage Method)이 사용되었다. 워드(Ward)와 단일연결(Single link 혹은 Nearest Neighber)은 선행연구(Griffiths. Robinson, & Willett 1984, Tombros, Villa, & Van Rijsbergen 2002, Zarma and Etzioni 1998)에서처럼 사전테스트에서도 좋지 않은 것으로 밝혀져 제외되었다. 완전연결법은 기존의 군집에 포함되어 있는 모든 객체에 대해서 일정거리 이내에 들어와야만 동일한 군집에 포함시키는 방법이다. 군집간의 거리는 각 군집에 속해 있는 객체간의 가장 먼 거리로 산정되는 방법으로, SPSS 군집방법 대화상자에서 "가장 먼 항목"으로 표현되어 있다. 군집간 평균연결법은 새로운 객체를 기존의 군집에 포함시킬 때 기존의 군집 내에 있는 모든 객체와의 평균거리가 가장 가까운 군집에 포함시키는 방법이다. SPSS에서 디폴트로 사용하는 방법으로 SPSS에서 "집단-간 연결"로 표현되어 있다. 군집내 평균연결법은 새로운 객체를 기존의 군집에 포함시킬 때 새로운 객체를 포함한 모든 객체간의 평균거리가 최소가 되는 군집에 포함시키는 방법이다. 즉 군집내의 객체들의 응집에 더 중점을 두는 군집화 방법으로 SPSS에서는 "집단-내 연결"로 표현되어 있다.

요약하면 <표 2>에서와 같이 3종류의 문헌용어행렬표가 작성되어 각각 3종류의 유사도 계수와 3종류의 군집기법으로 군집 테스트되었다. 통계패키지 SPSS 10이 사용되었다.

3.2 실험집단 분석

자동색인에서는 절대빈도(tf)와 이진빈도 백터로 각각 175(문헌)×353(용어) 행렬표가, 수동통합색인에서는 이진백터로 175(문헌)×358(용어) 행렬표가 작성되었다. 수동통

〈표 2〉 실험 설계: 분류자질 × 유사도 × 군집방법

문헌용어행렬표	자동tf	자동-이진	수동통합-이진
유사도	제곱 유클리드 코싸인 피어슨	제곱 유클리드 다이스 자카드	
군집기법	완전연결법(완전연결) 군집내 평균연결법(집단내) 군집간 평균연결법(집단간)		

합색인에서는 총 53개의 통제된 색인어가 자동색인어에 추가되었으나 전체 색인어 화일에 새로 추가된 새로운 용어는 5개였다.

용어의 평균 출현빈도는 2.36~2.47로, 한 단어가 자동색인에서는 평균 2.36개, 통제어 통합에서는 2.47개의 문헌에 출현하였다. 가장 많이 출현한 단어는 "도서관"으로 42개 문헌에 총 47번 출현하였다. 문헌당 최소 색인어 수는 1개, 최대 색인어 수는 12개, 문헌당 평균 색인어 수는 5개였다(<표 3> 참조).

3.3 문헌클러스터링 성능 평가 방법

탐색으로 클러스터링의 성능을 평가한 연구에서는 검색된 클러스터 내의 적합문헌수를 기반으로 클러스터링의 성능을 평가하나, 탐색을 포함시키지 않고 전체 문헌집단을 유사한 문헌끼리 소집단으로 얼마나 잘 군집하는지에 대한 평가는 생성된 전체 클러스터의 구조를 주관적으로 평가하거나 (Griffiths, Robinson, & Willett 1984), 주관적 평가가 어려운 큰 문헌집단에서는 수동분류 범주와의 유사도에 근거한 가중평가 클러스터유사도(정영미, 이재윤 2001)와 같은 평가척도를 사용할 수 있다. 본 연구에서 클러스터링의 성능은 수치화된 평가척도보다는 주관적인 판단으로 평가되었다.

기존의 분류번호가 클러스터링 성능의 평가척도가 되기 위해서는 정확한 분류번호가 필수적이다. 그러나 OPAC에서의 분류번호는 일차적으로 서가에서의 문헌의 위치를 지정하기 위한 목적으로 부여된 것이기 때문에 2개 이상의 주제를 담고있는 문헌을 1차(major) 주제에 분류함으로써 2차적인 주제로는 분류번호를 부여하지 못한다. 물론 도서관에 따라 검색의 목적으로 2차 주제의 분류번호를 부출할 수 있겠지만 175건의 실험문헌에는 부출된 분류번호가 없었다. 이 때문에 "세계 주요국의 도서관 및 독서현황 조사 자료집"이란 문헌은 020.9에 분류되어 독서는 반영되지 못하고 있다. 계층 클러스터링에서 바람직한 결과는 이 문헌이 도서관 현황과 독서를 계층구조에서 연결시키는 것일 것이다. 주관적 평가는 이와 같은 문헌이 계층구조에서 두 주제를 연결시키는지를 알 수 있는 장점이 있다.

〈표 3〉 실험문헌집단 분석

	자동tf	자동이진	수동통합이진
용어수	353	353	358
용어 총 출현빈도	856	832	885
용어 평균 출현빈도	2.42	2.36	2.47
용어당 최대 문헌 출현빈도	47	42	44
문헌당 최대 단어 출현빈도	14	12	12
문헌당 평균 단어 출현빈도	4.89	4.75	5.06

실험DB에는 잘못 부여된 분류번호와 DDC 분류표상 아직 전개되지 않은 주제(예, 인터넷심리학)도 적지 않았다. "국가기록물 관리의 발전방안에 관한 연구", "정보학의 실제", "정보화 사회와 도서관 정보네트워크" 등은 021(도서관과 사회)에 잘못 분류되었고, "전자도서관 구축론"은 026(특수도서관)에 잘못 분류되었다.

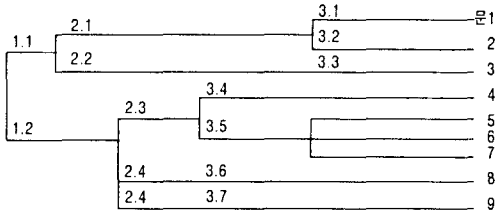
주관적 평가는 기존의 잘못 부여된 분류번호로부터 자유로울 뿐만 아니라 본 실험 DB 175권은 주관적으로 심층 분석하는데 큰 무리가 없는 규모였다.

4 클러스터링 성능 분석

4.1 계층별 군집 수 및 군집내 문헌 수

본 연구에서 계층단계는 DDC 020 아래 나타나는 군집을 1단계 계층군집이라고 하고, 1단계 계층군집이 세분되어 생성한 하위군집을 2단계 계층군집, 2단계 계층군집이 세분되어 생성한 하위군집을 3단계 계층군집이라고 한다. 하나의 문헌으로 독립집단을 이루는 집단이 같은 계층에 2개 이상 존재할 때는 '기타' 집단으로 묶어 한 집단으로 처리하였다. 따라서 9개의 문헌이 <그림 1>에서와 같이 클러스터링되었다면 1단계 계층에서는 2개의 군집이, 2단계 계층에서는 문8과 문9가 '기타'로 처리되어 4개 군집이, 3단계 계층에서는 7개 군집이 형성된 것으로 간주하였다.

<표 4>는 자동tf와 자동2진, 수동통합2진



〈그림 1〉 계층클러스터에서 계층단계와 군집 수

에서 유사도 계수와 군집 기법에 따라 생성된 계층별 군집 수와 군집 내 최소, 최대 문헌 수를 보여준다.

성능이 좋지않다고 알려진 유클리드(Ellis, Furner-Hines, & Willett 1993, Willett 1983)와 마찬가지로, 제곱 유클리드는 세 색인화일 모두에서 군집방법과는 상관없이 좋지 못하며, 집단내 평균연결법 또한 유사도 기법과 상관없이 문헌의 클러스터링에 좋지 못함을 볼 수 있다. 첫 계층에서 2-3개의 군집으로 클러스터링하며, 다음 계층으로 세분할 때마다 거의 두 그룹으로 2진 클러스터링하는 것을 볼 수 있다. 따라서 집단내 평균연결법과 제곱 유클리드 유사도 계수는 심층분석에서 제외되었다.

세 화일 모두에서 집단간 평균연결보다는 완전연결이, 다이스보다는 자카드가 더 여러 집단으로 군집하였다. 완전연결에서는 유사도간에 2단계에서 약간의 차이가 보일 뿐 거의 비슷하였다. 화일간에는 2진백터에서는 수동통합보다는 자동2진이 더 여러 집단으로 군집하였고, 자동화일에서는 tf보다는 2진백터가 더 여러 집단으로 군집하였다. 통합이진-집단간-다이스가 가장 작은 집단(9개)

〈표 4〉 파일별 유사도별 클러스터링기법별 계층 클러스터 수

				집단내			집단간			완전연결		
		계 층		1	2	3	1	2	3	1	2	3
자동 tf	코싸인	군집 수		3	6	12	11	32	49	24	57	
		군집내 문헌 수	최소	47	10	2	2	1	1	2	1	
			최대	76	57	49	42	39	37	34	27	
	피어슨	군집 수		3	6	12	10	28	46	24	57	
		군집내 문헌 수	최소	45	10	2	2	1	1	2	1	
			최대	73	57	49	42	39	37	34	27	
	유클리드	군집 수		2	4	9	2	4	5	2	3	4
		군집내 문헌 수	최소	33	5	1	2	1	1	1	1	1
			최대	142	136	118	173	172	173	174	173	167
자동이진	다이스	군집 수		3	6	12	12	33	53	26	60	
		군집내 문헌 수	최소	26	8	1	2	1	1	2	1	
			최대	103	59	45	43	40	38	31	20	
	자카드	군집 수		4	9	20	15	40	68	26	60	
		군집내 문헌 수	최소	29	1	1	2	1	1	2	1	
			최대	61	49	45	40	38	19	31	18	
	유클리드	군집 수		3	6	11	2	4	5	2	4	7
		군집내 문헌 수	최소	6	1	1	2	1	1	39	1	1
			최대	137	130	119	173	172	171	136	135	70
수동통합이진	다이스	군집 수		3	6	12	9	24	42	20	45	
		군집내 문헌 수	최소	43	12	5	2	1	1	2	1	
			최대	77	58	49	42	37	33	37	17	
	자카드	군집 수		3	6	13	13	29	55	20	47	
		군집내 문헌 수	최소	43	13	5	2	1	1	2	1	
			최대	70	57	45	42	35	25	37	16	
	유클리드	군집 수		2	4	9	2	4	5	2	4	7
		군집내 문헌 수	최소	46	6	1	2	1	1	38	1	1
			최대	129	112	113	173	172	170	137	136	94

으로, 자동이진-완전연결이 가장 많은 집단(26개)으로 분류하였다. 집단간의 3단계 범주를 완전연결에서는 2단계 계층으로 얻을 수 있었다.

4.2 1단계 군집 비교 및 적절성 분석

<표 5>는 파일별, 군집기법별, 유사도별 1단계에서 생성된 클러스터의 항목을 비교한 것이다. 생성된 클러스터는 자동색인과 수동통합색인 간에 근본적인 차이가 있었으며, 사용된 유사도와는 관계없이 모든 파일에서 집단간보다는 완전연결이 더 여러 집단으로 군집하였다. 모든 화일에서 완전연결은 유사도 계수간의 차이가 없이 동일한 클러스터를 형성하였다.

4.2.1 자동tf

먼저 자동색인은 집단간에서 피어슨은 10군집으로, 코싸인은 11군집으로 분류하였다. 코싸인과 피어슨은 코싸인에서 '이해'가 '정보학/문헌'으로부터 분리된 것을 제외하고 문헌의 분류는 동일하였다. 피어슨에서 '정보학/문헌'은 2단계에서 '이해'와 '정보학/문헌'으로 세분되었으나 코싸인에서는 1단계에서 '이해'와 '정보학/문헌'이 군집되었다. 도91(MARC의 이해), 도97(KDC의 이해), 도99(분류의 이해)를 '이해'로 군집시킨 것은 좋은 주제군집이 아닌 것 같다. '한국/학술'은 2단계에서 '색인', '웹 정보원', '한국/학술'로 세분되고, '한국/학술'은 3단계에서 다시 'DDC'와 '한국학술연구동향', '한국'으로 세분된다.

완전연결에서는 코싸인과 피어슨이 동일한 클러스터(24개)를 생성하였다. 집단간-피어슨에서 생성된 '정보학/문헌'과 '한국/학술'은 완전연결에서는 '정보학/문헌', '사상', '이해', '색인', '한국', '문헌', 'DDC'로 세분되었다. 도8(문헌정보학의 철학과 사상: 세라의 사상을 중심으로)와 도20(도서관인 박봉석의 생애와 사상)이 '사상'으로 군집되었다. '정보'는 '웹 정보원', '정보/검색', '인터넷', '학술정보', '실제'로 세분되었다. '기록/관리'는 '기록'과 '관리', '비도서자료'로 세분되었는데, '관리'에는 '정보관리'와 '기록관리'가 포함되었다. '대학/도서관'과 '주제명표목표 개발'이 '도서관'으로부터 분리되었다.

비주제어 '이해' 외에 '실제', '문헌', '한국', '학술정보'라는 범주가 생성되었지만 소속문헌을 보면 바람직한 군집은 아닌 것 같다. 도81(장서관리의 이론과 실제)와 89(목록조직의 실제)가 '실제'로 분류되었다. 도84와 도96(한국문헌자동화 목록), 도98(문헌분류의 실제), 도80(문헌자료조직론)이 도2(문헌정보활용법)와 함께 '문헌'으로 군집되었다.

4.2.2 자동2진

자동색인은 클러스터링 기법과 상관없이 절대빈도보다는 이진백터가 더 여러 항목으로 세분하였다. '정보학/문헌'은 '정보학/문헌', '이해'로 세분되었다. 비주제어 '이해'는 자동tf-코사인에서와 마찬가지로 1단계 군집을 형성했다.

자카드는 다이스의 '한국/학술' 클러스터

〈표 5〉 1단계 클러스터 비교

자동tf			자동2진			통합		
집단간(P)	집단간(C)	완전연결	집단간(D)	집단간(J)	완전연결	집단간(D)	집단간(J)	완전연결
10	11	24	12	15	26	9	13	20
정보학/문헌	정보학/문헌	정보학/문헌	정보학/문헌	정보학/문헌	정보학/문헌	한국/정보학	정보학	정보학
		사상						
	이해	이해	이해	이해	이해		목록/조직	목록
								자료조직
한국/학술	한국/학술	색인	한국/학술	색인	색인		색인	색인
		한국		한국/학술	학술		한국학술연구 동향	한국학술연구 동향
					한국			
		문헌			문헌	분류	분류	분류
		DDC			DDC			
정보	정보	웹 정보원	정보	웹 정보원	웹 정보원	정보	정보	정보검색
		정보/검색		정보	인터넷			
		인터넷			정보			사회
		학술정보			품질			정보
		실제			실제			
기록/관리	기록/관리	관리	비도서자료	비도서자료	관리	기록/관리	기록/관리	관리/기록
		비도서자료	기록관리	기록/관리	비도서자료			
		기록			기록관리			기록
기록학	기록학	기록학	기록학	기록학	기록학			
아카이브	아카이브	아카이브	아카이브	아카이브	아카이브	아카이브	아카이브	아카이브
도서관	도서관	도서관	도서관	도서관	도서관	도서관	도서관	도서관
		대학/도서관			대학/도서관			
		주제명표목개발		개발	개발		개발	주제명표목 개발
독서	독서	독서	독서	독서	독서	독서	독서/책	독서
		독서지도			독서지도			독서지도
책	책	책	책	책	책			책
기타	기타	기타	기타	기타	21세기	21세기	21세기	21세기
					논문집	기타	기타	논문집
					기타			기타

에서 도104(색인, 초록)과 도106(Wilson사 색인 변천, 발달)을 분리하여 '색인'을, 도110(한국 웹기반 인물정보원 연구)와 도113(웹 정보원과 검색엔진)을 분리하여 '웹 정보원'을 독립클러스터로 생성하였고, 도92,93(주제명표목표 개발)과 도111(참고질의시스템 개발)를 '도서관'에서 분리하여 '개발' 군집을 이루었다.

완전연결에서도 자동색인은 절대빈도(24군집)보다는 이진백터(26군집)가 더 여러 항목으로 군집하였다. 완전연결은 전체 12개의 클러스터링 중 가장 많은 26개의 범주를 생성하였다.

자동tf-완전연결에서 생성된 좋지 않은 군집 '이해', '실제', '문헌', '한국', '개발' 등의 군집이 자동2진-완전연결에서도 생성되었다. 자동이진-자카드에서 도92,93(주제명 표목표개발)은 도111(참고질의시스템개발)과 함께 '개발' 군집이 되었으나, 자동2진-완전연결에서는 도163(창의력개발을 위한 독서지도법, 독서신문 만들기)과 함께 '개발' 군집이 되었다.

집단간-자카드로 생성된 '한국/학술'은 완전연결에서 '학술', '한국', '문헌', 'DDC'로 분리되었다. 도14-17(한국학술연구동향)과 도6(학술정보론)이 '학술'로, 도86(한국편목규칙)과 도90(한국의 목록규칙변천사), 도105(한국용어열색인시스템), 도150(책과의 만남: 한국출판인회의선정 이달의 책)이 '한국'에, 도84(한국문헌자동화목록형식)과 도96(한국문헌자동화목록기술규칙), 도2(문헌정보활용법), 도80(문헌자료조직론), 도98(문헌분류의 실제)이 '문헌'으로 분리되었다.

'인터넷'과 '품질', '실제', '관리'가 '정보'로부터 독립되었다. 자동tf-완전연결에서 도33(인터넷심리학)과 도111(인터넷기반 참고질의시스템 개발)은 '인터넷'으로 군집되고 인터넷 정보검색(사)에 관한 문헌은 '정보/검색'에 분류되었지만, 자동2진-완전연결에서는 도111과 도33은 인터넷 정보검색(사)와 함께 '인터넷'에 포함되었다. 도41(KOSIS통계DB 품질평가)과 도107(도서관서비스 품질관리론)이 '품질'로 군집되었다. 자동tf에서는 도81(장서관리 이론과 실제)와 도89(목록조직의 실제)가 '실제'로 군집되었으나, 자동이진에서는 도71(한국국가기록관리의 이론과 실제)도 '실제'에 분류되었다.

4.2.3 수동통합이진

수동통합은 자동이진 파일에 추가된 통제어가 동의어와 하위어를 통제어로 연결해 줌으로써 가장 작은 군집을 이루었다. 자동이진화일과 비교하여 집단간에서 다이스는 12군집을 9군집으로, 자카드는 15군집을 13군집으로, 완전연결은 26군집을 20군집으로 축소하였다.

자동2진-다이스에서의 '정보학/문헌', '비도서자료', '이해', '한국/학술'이 수동통합-다이스에서는 '분류'를 제외하고 '한국/정보학'으로 통합되었으며, '책'은 '독서'로 통합되고, '기록학'도 '기록/관리'로 통합되었다. 그러나 도140(21세기를 움직일 명저 100선)과 도128(21세기 국회도서관 발전전략)이 '기타'에서 분리하여 비주제 군집 '21세기'를 이루었

다. 전체적으로 다른 파일에서 나타난 바람직해 보이지 않은 '이해', '한국', '실제', '문헌' 등이 사라졌고, 이들 군집에 분류되었던 문헌은 집단간-자카드에서 '분류', '색인', '한국학술연구동향', '목록/조직'을 생성하였고, '목록/조직'은 완전연결에서 다시 '목록'과 '자료조직'으로 분리되었다.

'정보'에서 '정보검색'과 '사회'가, '독서'에서 '독서지도'와 '책'이 분리되고, '기록/관리'가 '기록'과 '관리/기록'로 분리되는 등 수동통합-집단간-자카드의 13군집은 수동통합-완전연결에서 20군집으로 세분되었다. 자동색인에서 생긴 '이해'와 '실제'는 없어져서, 도97(KDC 이해)와 도99(분류의 이해)는 'DDC'와 함께 '분류'로, 도91(MARC의 이해)는 '목록'으로 분류되었다. '정보학'에는 정보학과 문헌정보학에 관한 문헌들만 분류되었다. 도81(장서관리의 실제)과 도71(국가기록관리 이론 실제)은 '관리/기록'에, 도89(목록조직의 실제)는 '목록'에 분류되었다. 도24(사회교육과 도서관)과 도27(정보화사회와 도서관 정보네트웍), 도108(정보화사회와 정보이용)이 '정보'에서 분리되어 '사회' 군집을 이루었다. 도92,93(주제명표목표 개발)은 집단간-다이스에서 도111(참고질의시스템 개발)과 함께 '도서관'의 하위군집을 이루었고, 집단간-자카드에서는 '도서관'에서 분리하여 독립군집 '개발'을 이루었으나, 완전연결에서는 자동tf-완전연결에서와 같이 독립군집 '주제명표목표 개발'을 이루었다.

4.3 잘못 분류된 문헌 분석

<표 6>은 군집을 이루는 문헌을 대상으로 1단계 계층에서 생긴 군집에 잘못 분류된 문헌을 분석한 것이다. 각 클러스터링에서 서로 다르게 분류된 도서를 그 도서가 속한 클러스터명과 함께 표시하였다. 문헌 용어행렬표를 작성할 때 비슷한 주제가 서로 인접하도록 175권의 도서를 DDC분류번호순으로 정렬하여 일련번호를 부여하였으므로 도서번호순 리스트는 비슷한 주제의 문헌을 인접시키고 있다. 특히 잘못 분류된 군집명을 짙은 문자로 나타냈다.

4.3.1 자동tf

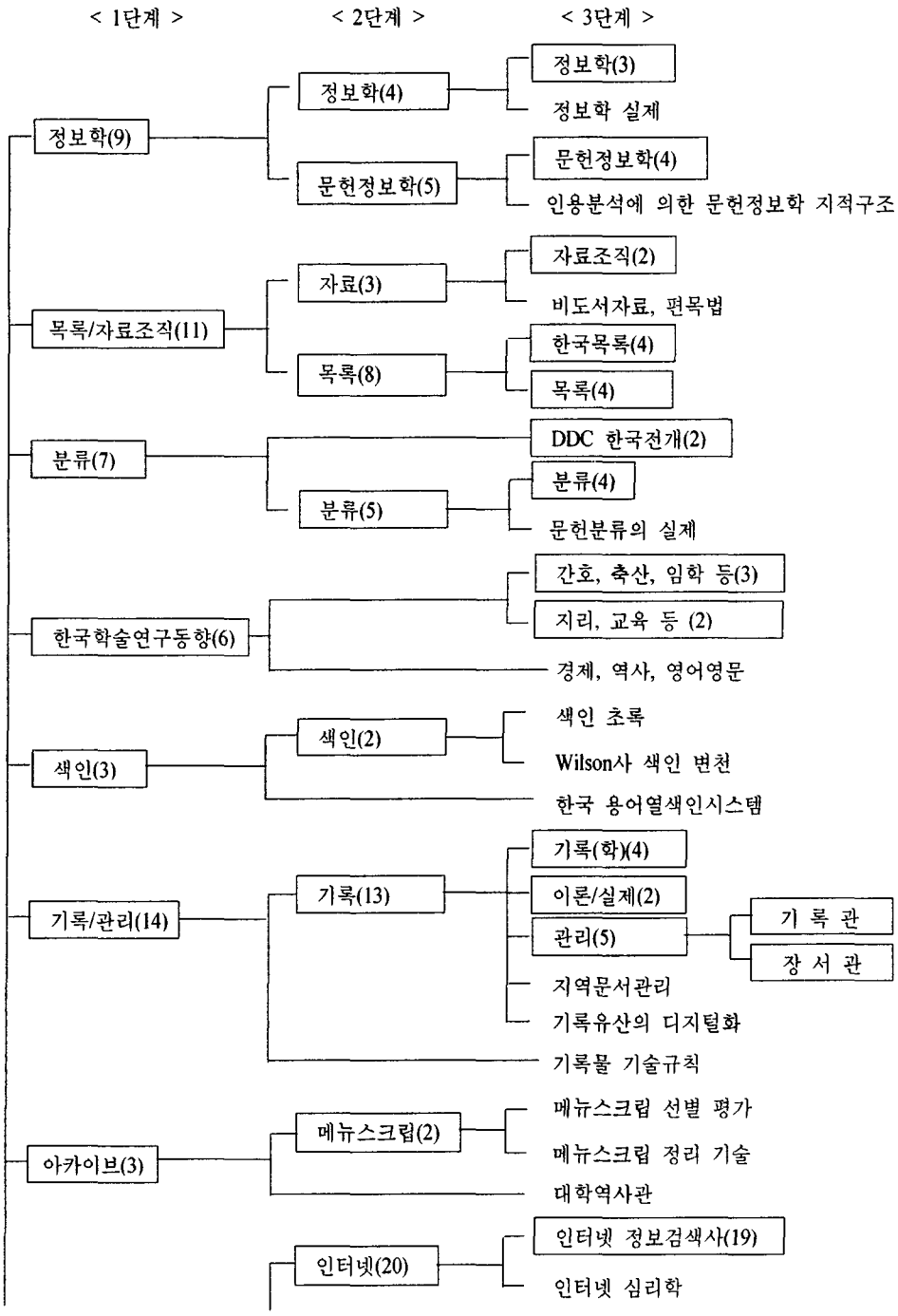
집단간에서 군집내 잘못 분류된 문헌으로는 도81(장서관리의 실제)와 도89(목록조직의 실제), 도94,95(비도서자료 편목,조직)이 '기록/관리'에 분류되었고, 도79(대학역사기록관 설립운영)과 도85(목록학)이 '기타'에 분류되었다. 도41(KOSIS통계DB 품질평가)가 도77,78(아카이브와 메뉴스크립트의 ...)와 함께 '아카이브'에 군집되었는데, 도77(아카이브와 메뉴스크립트의 선별과 평가)에 출현하는 색인어 "평가" 때문인 것으로 해석된다. 도87(미래도서관 목록법이론)은 '도서관'에 분류되었다.

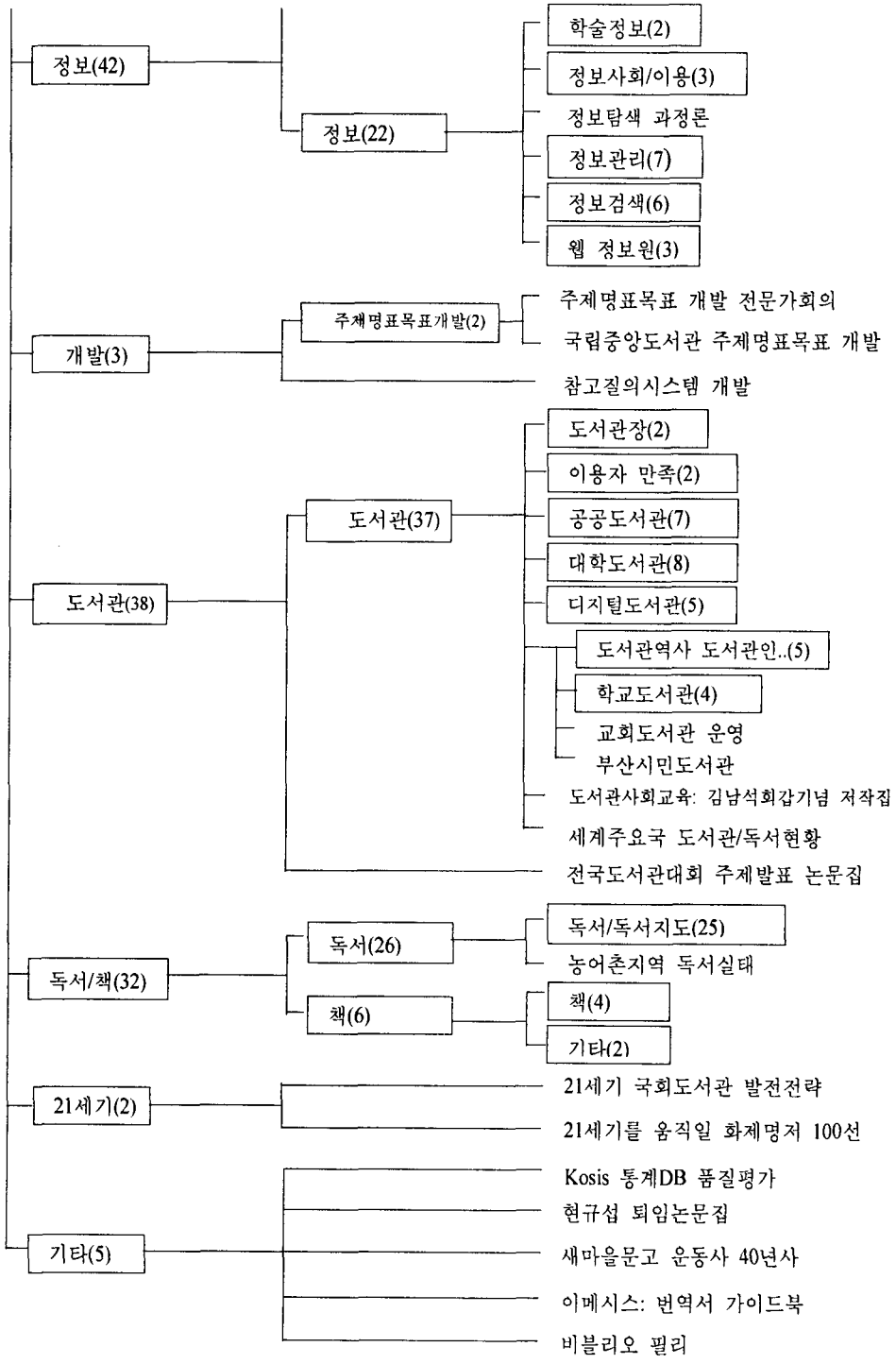
완전연결에서는 도105(한국용어열색인시스템)이 '색인'에 분류되지 못하고, 도12-17(한국학술연구동향), 도86(한국편목규칙), 도90(한국목록규칙변천사)와 함께 '한국'에 분류되었다. 도6(학술정보론)이 도2(문헌정보활용법)과 함께 분류되지 못하고 도112(전자공학대학원 학생 학술정보 이용형태)과 함께

〈표 6〉 문제가 된 군집 및 분류

	자동tf		자동이진		수동통제통합		
	집단간(코싸인)	완전연결	집단간(J)	완전연결	다이스	자카드	완전연결
도2 (문헌정보활용법)	정보학/문헌	문헌	정보학/문헌	문헌	정보	정보	정보
도6 (학술정보론)	한국/학술	학술정보	한국/학술	학술	정보	정보	정보
도12-17 (한국학술연구동향)	한국/학술	한국	한국/학술	학술	한국/정보학	한국학술연구 동향	한국학술 연구동향
도22 (국가기록물관리)	기록관리	**관리**	기록관리	기록관리	기록/관리	기록관리	관리/기록
도33 (인터넷 심리학)	정보-검색	인터넷	정보	인터넷	정보	정보	정보검색
도41 (KOSIS 통계DB 품질 평가)	**아카이브**	기타	**아카이브**	품질	**아카이브**	기타	기타
도 71 (국가기록관리의 이론과 실제)	기록/관리	기록	기록/관리	**실제**	기록/관리	기록/관리	기록/관리
도73 (지역문서관리)	기록관리	**관리**	기록관리	기록관리	기록/관리	기록관리	관리/기록
도74 (한국기록물관리 이해)	기록관리	**관리**	기록관리	기록관리	기록/관리	기록관리	관리/기록
도79 (대학역사기록관 설립)	**기타**	대학도서관	**기타**	대학도서관	아카이브	아카이브	아카이브
도80 (문헌자료 조직론)	정보학/문헌	문헌	정보학/문헌	문헌	한국/정보학	목록/조직	자료조직
도81 (장서관리 이론 실제)	**기록/관리**	**실제**	**기록/관리**	**실제**	**기록/관리**	**기록/관리**	**관리/기록**
도84,96 (한국문헌자동화목록)	한국/학술	문헌	한국/학술	문헌	한국/정보학	목록/조직	목록
도85 (목록학)	**기타**	**기타**	**기타**	**기타**	한국/정보학	목록/조직	목록
도86 (한국편목규칙)	한국/학술	한국	한국/학술	한국	한국/정보학	목록/조직	목록
도87 (미래도서관목록법이론)	**도서관**	**도서관**	**도서관**	**도서관**	한국/정보학	목록/조직	**도서관**
도88 (기록물 기술규칙(ISAD))	한국/학술	**기타**	**기타**	**기타**	한국/정보학	목록/조직	목록
도89 (목록조직 실제)	**기록/관리**	**실제**	한국/학술	**실제**	한국/정보학	목록/조직	목록
도90 (한국목록규칙 변천사)	한국/학술	한국	한국/학술	한국	한국/정보학	목록/조직	목록
도91 (MARC 이해)	**이해**	**이해**	**이해**	**이해**	한국/정보학	목록/조직	목록
도92,93 (국립중앙도서관 주제명표목표 개발)	도서관-개발	주제명 표목표개발	개발	개발	도서관-개발	개발	주제명 표목개발
도94,95 (비도서자료 편목, 조직, 이론)	**기록/관리**	비도서자료	비도서자료	비도서자료	한국/정보학	목록/조직	목록
도97 (KDC 이해)	**이해**	**이해**	**이해**	**이해**	분류	분류	분류
도98 (문헌분류의 실제)	정보학/문헌	문헌	정보학/문헌	문헌	한국/정보학	분류	분류
도99 (분류의 이해)	**이해**	**이해**	**이해**	**이해**	분류	분류	분류
도105 (한국용어열색인)	한국/학술	한국	색인	**한국**	한국/정보학	색인	색인
도111 (인터넷 참고질의시스템 개발)	도서관-개발	인터넷	개발	인터넷	도서관-개발	개발	정보검색
도150 (한국출판인회의 선정 이달의 책)	책	책	책	**한국**	책	책	책
도163 (창의력개발을 위한 독서 지도법과 독서신문 만들기)	독서	독서	독서	**개발**	독서	독서	독서

〈표 7〉 수동통합-집단간-자카드 군집분류




정보(42)
정보(22)
학술정보(2)
정보사회/이용(3)
정보탐색 과정론
정보관리(7)
정보검색(6)
웹 정보원(3)
개발(3)
주제명표목표개발(2)
주제명표목표 개발 전문가회의
국립중앙도서관 주제명표목표 개발
참고질의시스템 개발
도서관(38)
도서관(37)
도서관장(2)
이용자 만족(2)
공공도서관(7)
대학도서관(8)
디지털도서관(5)
도서관역사 도서관인..(5)
학교도서관(4)
교회도서관 운영
부산시민도서관
도서관사회교육: 김남석회갑기념 저작집
세계주요국 도서관/독서현황
전국도서관대회 주제발표 논문집
독서/책(32)
독서(26)
독서/독서지도(25)
농어촌지역 독서실태
책(6)
책(4)
기타(2)
21세기(2)
21세기 국회도서관 발전전략
21세기를 움직일 화제명저 100선
기타(5)
Kosis 통계DB 품질평가
현규섭 퇴임논문집
새마을문고 운동사 40년사
이메시스: 번역서 가이드북
비블리오 필리

독립군집 '학술정보'을 이루었다. 도22(기록물관리), 도73(지역문서관리), 도74(한국기록물 관리와 이해)가 '기록'에 분류되지 못하고 도42-48(정보관리강좌), 도115(연구단지 정보관리총람)와 함께 '관리'에 분류되었다. 도79(대학역사기록관 설립)은 '아카이브'보다는 '대학도서관'에 분류되었고, 도87(미래도서관 목록법이론)도 '도서관'에 분류되었다. 도85(목록학)과 도88(기록물 기술규칙; ISAD, APPM2, RAD, ...)은 '기타'로 분류되었다.

4.3.2 자동2진

자동2진에서 다른 군집에 잘못 분류된 문헌으로는 집단간에서는 도81(장서관리)가 '기록/관리'에, 도41(KOSIS통계DB품질)이 '아카이브'에, 도87(미래도서관 목록)이 '도서관'에 분류되었다. 도85(목록학)과 도88(기록물 기술규칙...), 도79(대학기록관 설립)은 '기타'에 분류되었다.

완전연결에서 잘못 분류된 문헌으로는 도81(장서관리), 도89(목록조직) 외에 도71(국가기록관리)가 '실제'에 분류되었고, 도87(미래도서관 분류 목록법)은 '도서관'에 분류되었다. 도105(한국용어열색인)은 '색인'이 아닌 '한국'에 분류되었고, 도150(한국출판인협회 선정 이달의 책)이 '책'이 아닌 '한국'에 분류되었다. 도85(목록학)과 도88(기록물 기술규칙...)은 '기타'에 분류되었다.

4.3.3 수동통합이진

수동통합이진에서 잘못 분류된 문헌으로는 도41(KOSIS통계DB 품질평가)가 집단간-다이스에서 '아카이브'에 분류되고, 도81(장서관리 이론과 실제)가 집단간과 완전연결 모두에서 '기록/관리'에 포함된 것을 제외하고는 큰 무리는 없어 보였다. 도81(장서관리 이론과 실제)은 12종류의 클러스터 모두에서 '기록/관리' 혹은 '실제'에 잘못 분류되었다. 이는 장서관리에 대한 다른 문헌이 없었기 때문에 장서관리로는 군집되지 못하고 "관리" 혹은 "실제"가 출현하는 문헌들과 군집되었기 때문이다.

전체적으로 12종류의 군집 중 수동통합-집단간-자카드 혹은 수동통합-완전연결이 가장 주제적으로 분명하고 균형적이고 우수한 클러스터링을 생성했으나, 수동통합-집단간-자카드가 보다 10진구조와 유사하였다. <표 7>은 가장 바람직해 보이는 수동통합-집단간-자카드가 생성한 클러스터를 계층별로 3단계까지 전개한 것이다. 사각형으로 표시된 것은 군집명이고, 하나의 문헌으로 이루어진 군집은 문헌명을 사각형 없이 그대로 표시하였다. 괄호 안의 숫자는 군집에 분류된 문헌수를 나타낸다.

5 DDC 분류와 클러스터링 범주 비교

클러스터링에 의한 한글문헌 분류는 DDC보다 문헌을 더 균형적으로 분류하는가?

<표 8>은 175건의 실험집단의 DDC 분류번호를 가장 바람직해 보이는 수동통합-집단간-자카드 클러스터링으로 생성된 클러스터

〈표 8〉 DDC 분류와 수동통합-집단간-자카드 범주 비교

DDC		수동통합-집단간-자카드	
DDC 분류번호	분류건수	범주명	분류건수
도서관 정보학 일반(020)	21	정보학	9
		한국학술연구동향	6
도서관과 사회(021)	6		
도서관 건물/설비(022)	1		
도서관 운영(025)	85	목록/자료조직	11
		분류	7
		색인	3
		기록/관리	14
		아카이브	3
		정보	42
		개발	3
특수도서관(026)	3	도서관	38
일반도서관(027)	24		
독서지도(028)	35	독서/책	32
		21세기	2
		기타	5

와 분류건수를 비교한 것이다.[4)]

DDC 분류와 비교할 때 먼저 020(문헌정보학 일반)에 분류된 21건(도1~도21까지) 중 7건은 "정보학" 범주로, 6건은 "학술연구동향"으로 클러스터링되었지만, 나머지 8권은 다른 범주에 분류되었다. 도11(정년퇴임 논문집)은 "기타"에, 도10(도서관대회 주제발표 논문집)과 도19(세계주요국 도서관, 독서현황), 도20(도서관인 박봉석 생애와 사상), 도21(중국의 도서관과 도서관사업)은 "도서관"에, 도2(문헌정보활용법)는 "정보 -정보사회/이용"에, 도6(학술정보론)은 "정보 -학술정보"에 클러스터링되었다. 020에 잘못 분류된 것으로 보이는 도18(대학도서관 웹페이지 평가 개선방안)은 "도서관 -대학도서관"에 클러스터링되었다.

021(도서관과 사회)에 분류된 6권(도22~도27까지) 중 021에 잘못 분류된 것으로 보이는 도22(국가기록물 관리 발전방안)은 "기록관리"에, 도26(정보학의 실제)은 "정보학"에, 도27(정보화사회와 도서관네트웍)은 "정보 -정보사회/이용"에, 도23(도서관 어제와

4) 분류열람으로 검색된 175 문헌의 분류번호를 분석한 결과 실제 분류번호는 분류열람에서 5범주로 분류된 것과는 약간의 차이가 있었다. 분류열람에서 열람되지 않았던 022(도서관 건물)에 분류된 문헌도 있었고, 025와 027의 문헌수도 일치하지 않았다.

오늘)과 도24(김남석 회갑기념저작집 사회교육과 도서관)와 도25(세계 국립도서관 협력)는 "도서관"에 클러스터링되었다.

022(도서관 건물 및 시설)에 분류된 도28(대구지역 이동도서관 이용자만족도)는 "도서관 -이용자/만족"에 클러스터링되었다.

025(도서관 운영)에 분류된 85건의 문헌(도29~도113까지)은 "목록", "분류", "색인", "기록/관리", "아카이브", "개발", "정보"로 클러스터링되었다. 다만 도107(도서관서비스 품질관리: 고객만족)과, 도32(디지털도서관), 도109(디지털도서관 정보서비스), 도83(디지털도서관 구축을 위한 XML 스키마데이타), 도66(초,중,대학도서관 발전 종합계획)이 "도서관"에 클러스터링되었다.

026(특수도서관)과 027(일반도서관)에 분류된 27건(도114~도140까지)은 도115(연구단지정보관리총서)가 "정보"에, 도128(21세기 국회도서관)이 "21세기"에, 도129(새마을문고 운동40년사)가 "기타"에 클러스터링된 것을 제외하고 모두 "도서관"에 클러스터링되었다. 특수도서관은 도서관명칭의 고유성 때문에 통제어를 추가하지 않는 한 하나의 클러스터로 군집되기 어려워 보였다.

028(독서지도)에 분류된 35건의 문헌(도141~175까지)은 도147(이메시스)과 도148(비블리오필리)가 "기타"에, 도149(21세기 명저 100선)이 "21세기"에 클러스터링된 것을 제외하고 모두 "독서/책"으로 클러스터링되었다.

이상과 같이 집단간 클러스터링은 DDC와 비교하여 "개발"과 "21세기"를 제외하고 문헌을 1차 주제로 분류하였으며, 같은 주제를 같은 범주로, 잘못된 DDC분류번호로 분류된 문헌까지 비교적 잘 클러스터링하였다. 십진이라는 제한이 없기 때문에 DDC에서처럼 도서관 업무와 관종이라는 지식 분류에 억매이지않고 "분류"나 "편목", "색인", "기록관리" 등과 같은 주요 연구분야가 1차 클러스터로 생성되었다. 이로써 DDC에서는 하나의 범주에 최고 85건의 문헌이 분류되어 있으나, 집단간 클러스터링으로는 최고 42건의 문헌이 군집됨으로써 주제를 보다 균형적으로 군집한 것을 볼 수 있다.

6 군집명 분석

자동 추출된 클러스터명은 적합문헌 클러스터를 분별할 분별력을 갖는가?

<표 5>에 표시된 클러스터명은 클러스터에 분류된 모든 문헌에 공통으로 출현하는 단어로 이름을 붙였다. 그러나 클러스터 내 모든 문헌에 공통으로 출현하는 단어가 없을 때는 최다출현 두 단어를 출현빈도수 순서대로 '/'로 연결하여 명명하였다. 그러므로 "기록"과 "관리" 두 단어가 군집 내 모든 문헌에 출현하면 '기록관리'로, 모든 문헌에 출현하지 않지만 "기록"의 빈도수가 최고이고 "관리"가 그 다음이면 '기록/관리'로, "관리"가 최고이고 "기록"이 그 다음이면 '관리/기록'으로 표현하였다. 군집명은 '한국/정보학', '정보학/문헌', '한국/학술'을 제외하고 군집에 분류된 문헌의 주제를 비교적 잘 반영하고 있다.

수동통합-집단간-다이스에서 '한국/정보학'은 다른 군집에 속하지 않은 정보학과 문헌정보학, 목록 관련 20개의 문헌이 모인 군집이다. 그러나 "한국"이 14번, "정보학"이 9번, "문헌"이 8번의 순서로 출현하였으므로 군집명은 '한국/정보학'이 되었다. 자동tf-집단간에서 '한국/학술' 군집에 속한 문헌은 색인, 웹, 분류, 목록, 한국학술연구동향 등의 주제에 관한 문헌으로 집단명은 소속 문헌의 내용을 잘 표현하지 못하고 있다.

'정보학' 집단은 정보학 혹은 문헌정보학이 출현하는 문헌으로 이루어졌으나, "문헌정보학"이 "문헌정보학"과 "문헌", "정보학"으로 색인되었기 때문에 "문헌정보학"과 "정보학"의 모든 문헌에 출현하는 '정보학'이 군집명이 되었다.

7 결론 및 제한점

결론적으로 본 연구의 결과를 기술하면 다음과 같다.

1) 문헌정보학 분야의 계층 클러스터링에서 집단내 평균연결법과 제곱유클리드 유사도를 제외하고 12종류의 계층클러스터링(3화일×2군집×2유사도)은 모두 비교적 좋은 군집을 형성하였다. 1차 계층클러스터로 생성된 군집 수와 군집명, 군집에 분류된 문헌들의 주제를 분석한 결과, 통제어가 추가된 수동통합색인에서 집단간-자카드와 완전연결이 문헌정보학 관련 주제를 가장 균형있고 정확하게 분류하였으나, 집단간-자카드가 보다 10진 구조와 유사하였다.
2) 생성된 클러스터는 자동색인과 수동통합색인 간에 근본적인 차이가 있었다. 자동색인보다는 통제어를 추가시킨 통제통합색인에서 가장 좋은 군집이 형성되었다. 색인어의 가중치는 이진빈도가 절대빈도(tf)보다 약간 더 많은 클러스터를 형성하였지만, 내용은 매우 유사하였다.
3) 다이스보다는 자카드가, 피어슨보다는 코싸인이 더 많은 집단으로 군집하였다. 집단간 평균연결에서는 이진빈도에 사용된 다이스와 자카드가 절대빈도에 사용된 피어슨과 코싸인보다 더 많은 클러스터로 군집하였다. 완전연결에서는 유사도 방법이 클러스터링에 영향을 주지 못하였다.
4) 모든 파일에서 사용된 유사도계수와 관계없이 집단간보다는 완전연결이 보다 많은 군집으로 클러스터링하였다. 집단간 평균으로 생성되는 3단계 클러스터는 완전연결에서는 2단계 계층단계에서 생성되었다. 1차 계층클러스터링에서 집단간은 9~15개의 클러스터를, 완전연결은 20~26개의 클러스터를 생성하였는데, 집단간이 보다 10진구조에 가까웠다.
5) 최적의 모델로 제시된 수동통합-집단간-자카드 기법으로 생성된 클러스터는 DDC에 비해 학문연구분야를 보다 직접적으로 표현하며, 한 클러스터에 치

우치지 않고 보다 균형적으로 문헌을 클러스터링하는 것으로 보였다.

6) 군집 내 단어의 출현빈도에 의해 부여된 군집명은 '한국/정보학', '정보학/문헌', '한국/학술'을 제외하고 군집내용을 비교적 잘 표현하였다.

그러나 연구결과에 대한 해석은 다음과 같이 제한된다.

1) 서명출현 단어로만 분류의 자질을 삼는 것보다는 통제어를 추가하는 것이 보다 좋은 클러스터링을 가져왔다. 본 연구에서는 문헌당 평균 1개 미만의 통제어가 추가되었으나 도서관에서 실제 653 필드에 부과한 통제어의 수(포괄성)와 특정성은 클러스터링에 어떤 영향을 끼칠지 알 수 없다.

2) 자동색인에서 비주제 군집을 이루는 '이해'나 '실제'같은 용어의 추출은 바람직해 보이지 않았다. 도서 "MARC의 이해"와 "KDC의 이해"에서 "MARC"와 "KDC"는 다른 문헌에 출현하지 않지만 "이해"는 출현했기 때문에 '이해'로 군집되었다. 그러나 수동통합에서는 "분류"나 "목록"을 추가함으로써 주제로 군집되었다. 문헌의 크기가 증가하여 MARC나 KDC와 같은 주제의 문헌이 더 있을 경우에도 자동색인에서 군집 '이해'를 생성할지 알 수 없다. 자동색인시 비주제어를 제외시키는 것이 좋을지는 문헌집단의 크기를 늘려 실험해 볼 필요가 있겠다.

3) 집단간보다는 완전연결이, 다이스보다는 자카드계수가 보다 짧은 계층단계에서 보다 많은 군집을 생성하였다. 이는 집단의 동질성이 높은 한 학문 분야 소규모의 실험집단으로 테스트한 결과이다. 동질성이 보다 낮은 여러 학문분야를 통합한 문헌집단이나 동질성이 보다 높은 탐색결과를 클러스터링할 때도 같은 결과를 얻을지는 알 수 없다.

4) 군집에 출현하는 단어의 출현빈도수를 기반으로 군집명을 부여하는 것은 몇 개의 군집을 제외하고 적절하였다. 그러나 단순히 단어의 출현빈도에 근거하여 1~2개의 단어를 추출하는 것보다 주제를 보다 잘 나타낼 수 있는 복합명사나 구의 생성에도 관심을 가질 필요가 있겠다.

5) 본 연구의 결과는 학위논문과 단행본 도서를 대상으로 한 결과이다. 따라서 본 연구의 결론은 학위논문과 단행본 도서에 제한될 것이다. 클러스터링 대상 문헌의 동질성과 문헌집단의 크기가 클러스터링에 어떤 영향을 끼치는지는 흥미로운 후속 연구과제가 될 것이다.

본 연구에서 보고된 최적의 정적 클러스터링 모형은 탐색결과를 클러스터링하는 동적 클러스터링 모형으로도 사용할 수 있는지 후속연구가 진행 중이다. 불리언 OPAC시스템에서 서명단어로 탐색하여 얻은 검색결과를 계층클러스터링하여 제공되는 클러스터 중 최적의 클러스터를 선택할 때 선택된

클러스터의 성능을 분석하는 연구이다. 특정 주제로 탐색하여 얻는 탐색결과는 하나의 학문분야보다는 문헌의 동질성이 더 높기 때문에 모형에서와 같은 결과를 얻게될 지 흥미롭다.

참 고 문 헌

정영미, 이재윤. 2001. 지식분류의 자동화를 위한 클러스터링 모형 연구. 『정보관리학회지』, 18(2): 203-230.

한승희, 이재윤. 1999. 문헌클러스터링을 위한 유사계수 간의 연관성 측정. 『제6회 한국정보관리학회 학술대회 논문집』, 8: 25-28.

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

Blecic, D. D., J. L. Dorsch, M. H. Koenig, and N. S. Bangalore. 1999. "A Lorgitudinal study of the effects of OPAC screen changes on searching behavior and searcher success." *College & Research Libraries,* 60(Nov.): 515-530.

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.

Cooper, M. D. & Hui-Min Chen. 2001. "Predicting the relevance of a library catalog search." *Journal of the American Society for Information Science and Technology,* 52(10: 812-827.

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

Cutting, D. R., J. O. Pedersen, D. Karger, and J. W. Tukey. 1992. "Scatter/Gather: a cluster-based approach to browsing large document collections." *Processing of the 15th Annual International ACM SIGIR Conference on Research and development in Information Retrieval:* 318-329.

El-Hamdouchi, A. and P. Willett. 1989. "Comparison of hierarchic agglomerative clustering methods for document retrieval." *The Computer Journal,* 32(3): 220-227.

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

Mechkour, Harper, & Muresan. 1998. "The WebCluster Project Using clustering for mediating access to the world Wide Web." *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and development in information Retrieval:* 357-358.

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.

Voorbij, Henk J. 1998. "Title key-words and subject descriptors: a comparison of subject entries of books in the humanity and social science." *Journal of Documentation,* 54(4): 466-476.

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 demon-

stration." *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 46-54.

Zink, Steven A. 1991. "Monitoring user success through transaction log analysis: The WolfPAC Example." *Reference Services Review*, 19(spring): 49-56.