연구 영역 분석을 위한 디스크립터 프로파일링에 관한 연구*

Descriptor Profiling for Research Domain Analysis

김판준(Panjun Kim)**, 이재윤(Jae Yun Lee)***

초록

본 연구는 연구 영역 분석을 위하여 통제어휘와 비통제어휘를 연계해서 사용하는 새로운 방법을 모색하기 위한 것이다. 동시출현단어분석은 크게 통제어휘와 비통제어휘를 사용하는 경우의 두 가지 유형으로 구분할 수 있는데, 통제어휘를 사용할 경우에는 자료 희귀성 및 색인자 효과가 단점이며, 비통제어휘를 사용할 경우에는 저자의 주관에 따른 단어 선택 및 단어의 중의성이 문제가 된다. 이 연구에서는 양자를 보완할 수 있는 방법으로, 통제어휘인 디스크립터를 비통제어휘인 단어와의 동시출현 정보로 표현하는 디스크립터 프로파일링을 제안하였다. 정보학 분야에 적용해본 결과, 디스크립터 프로파일링은 특정 영역의 최신 동향을 파악하는데 있어 통제어휘와 비통제어휘가 갖는 본질적인 문제점을 어느 정도 보완할 수 있는 것으로 나타났다.

ABSTRACT

This study aims to explore a new technique making complementary linkage between controlled vocabularies and uncontrolled vocabularies for analyzing a research domain. Co-word analysis can be largely divided into two based on the types of vocabulary used: controlled and uncontrolled. In the case of using controlled vocabulary, data sparseness and indexer effect are inherent drawbacks. On the other case, word selection by the author's perspective and word ambiguity. To complement each other, we suggest a descriptor profiling that represents descriptors(controlled vocabulary) as the co-occurrence with words from the text(uncontrolled vocabulary). Applying the profiling to the domain of information science implies that this method can complement each other by reducing the inherent shortcoming of the controlled and uncontrolled vocabulary.

키워드: 디스크립터, 저자, 프로파일링, 연구 동향 분석, 지적 구조, 매핑, 동시출현단어분석, 통제어휘, 비통제어휘, descriptor, author, profiling, domain analysis, intellectual structure, mapping, co-word analysis, controlled vocabulary, uncontrolled vocabulary

---

* 이 논문은 제14회 한국정보관리학회 학술 대회(2007년 8월 17일, 연세대학교)에서 발표한 내용을 보완한 것임.

** 연세대학교 문헌정보학과 시간강사(dpbluesea@hanmail.net), 제1저자.

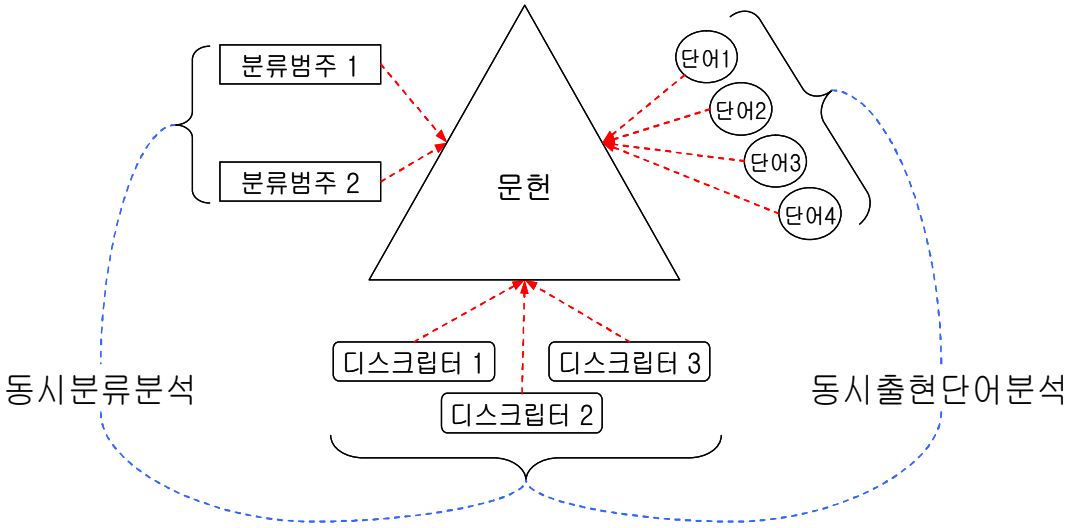
*** 경기대학교 문헌정보학전공 조교수(memexlee@kgu.ac.kr), 교신저자.

1. 서론

1980년대 초반에 처음 개발된 동시출현단어분석(co-word analysis) 기법은 특정 연구 분야의 동향을 파악하기 위한 목적으로 널리 사용되어 왔다(Callan et al. 1983; Callan et al. 1986). 지금까지 이 기법은 여러 연구자에 의해 데이터베이스로부터 지식 구조를 발견하기 위한 강력한 도구로 인정받아 왔다(He, 1999). 또한 이를 응용하는 텍스트 마이닝 기술로 데이터베이스 단층분석법(database tomography; Kostoff 1993; Kostoff et al. 2005), 테크마이닝(tech mining; Porter & Cunningham 2005) 등이 꾸준히 연구되어왔다.

동시출현단어분석은 다음과 같은 단계로 구성된다. 첫째, 대상 영역의 주요 개체(주제, 저자, 학술지 등)의 식별 및 표현, 둘째, 주요 개체 간의 관계(출판, 인용, 공저 등) 규명, 셋째, 관계를 이용한 주요 개체의 매핑.

동시출현단어분석은 분석 대상에 따라 동시분류분석(co-classification analysis; Todorov 1989), 동시출현표목분석(co-heading analysis; Todorov & Winterhager 1990), 동시출현디스크립터분석(co-descriptor analysis; McCain 1995), 동시출현용어네트워크분석(co-term network analysis; Jacobs 2002) 등으로도 일컫는데, <그림 1>에서처럼 디스크립터나 표목의 동시출현을 다룬 경우의 명칭은 '분류'와 '단어'가 혼용되고 있다. 이 연구에서는 혼란을 피하기 위해서 분류범주를 제외한 디스크립터나 단어의 동시출현을 분석하는 경우로 동시출현단어분석의 의미를 제한하였다.



<그림 1> 동시분류분석과 동시출현단어분석

동시출현단어분석은 크게 디스크립터나 주제명표목과 같은 통제어휘를 사용하는 경우와, 자연어 색인이나 자동추출한 색인어와 같은 비통제어휘를 사용하는 경우로 나눌 수 있다. 그런데 통제어휘를 사용하는 분석은 다양한 이형을 통제할 수 있다는 장점이 있지만, 세밀한 주제 표현이 어려우며 빈도가 낮은 통제어휘의 경우에는 데이터 희귀성 문제가 발생한다. 즉 아주 작은 수의 동시출현 만으로 두 디스크립터 사이의 관계가 결정되는 결과를 초래한다. 또한 색인자의 주관에 좌우되는 색인자 효과(Courtial et al. 1984)도 본질적인 문제점으로 지적되고 있다. 반면에 비통제어휘를 사용할 경우에는 저자의 주관에 따른 단어 선

택으로 인한 단어 불일치 및 단어 중의성 문제가 걸림돌이 된다. 또한 색인어휘가 지나치게 많이 추출되기 때문에 시각화하여 분석하기가 어렵다는 문제도 있다. 통제어휘와 비통제어휘를 사용하는 경우에 각각 직면하는 문제점들을 극복하기 위해서 이 연구에서는 통제어휘와 비통제어휘를 함께 상호보완적으로 활용하기 위한 새로운 방법으로 디스크립터 프로파일링을 제안하였다.

2. 선행 연구

지금까지 특정 연구 영역을 분석하여 시각화하기 위한 목적으로 단어동시출현분석을 적용한 연구들은 주제를 표현하는 키워드로서 통제어휘 또는 비통제어휘 중 하나만을 사용하였다. 초기의 연구들은 대부분 색인자가 부여한 주제명표목, 디스크립터, 분류기호 등의 통제어휘를 사용한 경우가 많았다(Callan et al. 1983; Todorov 1990; Tijssen 1992; McCain 1995; Spasser 1997; Noyons and van Raan 1998). 그러나 통제어휘의 본질적인 문제 중의 하나로 잘 알려져 있는 색인자 효과(van Raan 1987)를 극복하기 위하여 표제, 초록 등에 출현한 비통제어휘를 사용하는 경향이 점차 생겨났으며(Bhattacharya and Basu 1998; Widhalm et al. 2001; Watt, Porter, and Minsk 2004), 또한 최근에는 이들 두 가지 유형의 어휘를 함께 사용하거나 문헌의 일부 또는 전문(full-text)에서 추출된 단어 또는 명사구를 사용하려는 시도가 활발히 이루어지고 있다(Voutilainen 1993; Ding et al. 2001; Buter and Noyons 2002; Porter 2005; Kostoff et al. 2005; Janssens et al. 2006).

그러나 이들 연구는 모두 양자 중 하나만을 사용하거나 양자를 하나로 통합하여 사용한 것이라 할 수 있는데, 특히 후자의 경우에는 통제어휘와 비통제어휘를 구분하지 않고 하나의 키워드 집합으로 통합하여 사용하는 경우가 많았다. 다시 말해서, 이는 단순히 통제어휘의 문제점으로 지적된 '색인자 효과'를 감소시키기 위하여 디스크립터, 주제명표목, 통제키워드 등을 표제, 초록, 전문(full-text)에서 추출한 비통제 키워드들과 통합한 키워드 집합을 생성하고, 이들 키워드 각각을 개별적인 주제로 식별하고 표현한 것이다. 이 경우 통제어휘의 장점으로서 다양한 이형을 통제할 수 있는 측면을 전혀 무시하게 되는 결과가 되어 단어 불일치 및 중의성 문제는 그대로 내재하는 결과가 된다.

본 연구에서는 통제어휘와 비통제어휘 각각의 장점은 살리고 단점은 서로 보완할 수 있는 방법으로서 디스크립터 프로파일링(descriptor profiling)을 제안하였다. 디스크립터 프로파일링은 기본적으로는 문헌에 부여된 디스크립터로 주제를 표현하는 한편, 이러한 디스크립터 프로파일(디스크립터 벡터)를 생성하는 과정에서 프로파일의 구성요소로 해당 디스크립터가 부여된 문헌의 표제와 초록에 출현한 단어들을 사용하는 것이다. 이를 데이터베이스 측면에서 보면 문헌의 표제와 초록 필드에서 추출한 단어들은 저자가 생성한 비통제어휘에 포함되고, 디스크립터 필드에서 추출한 용어는 색인자가 부여한 통제언어에 속하는 것이다. 따라서 디스크립터 프로파일링은 특정 연구 영역의 주제를 식별하고 표현하는 과정에서 통제어휘와 비통제어휘를 함께 사용하면서 양자의 장점을 잃지 않고 문제점은 보완할 수 있는 효과적인 방법이 될 수 있을 것이다.

3. 디스크립터 프로파일링

여기서 제안하는 디스크립터 프로파일링은 주제를 대표하는 개체로 통제어휘인 디스크립

터를 사용하되, 각 디스크립터 및 이들 간의 관계는 해당 디스크립터가 부여된 문헌집합의 단어들에 기반하여 표현하는 방법이다. 이와는 달리 디스크립터가 아닌 저자를 개체로 삼고 저자 사이의 관계는 각 저자가 발표한 논문집합에 포함된 단어들을 비교하여 구하는 경우(이재윤 2005)는 저자 프로파일링이라 할 수 있다. 구체적으로는 <그림 2>와 같은 절차를 거쳐서 두 디스크립터 사이의 연관도를 산출한다.

<그림 2> 디스크립터 프로파일링 절차

문헌 벡터 생성 단계부터 상세히 설명하면 다음과 같다.

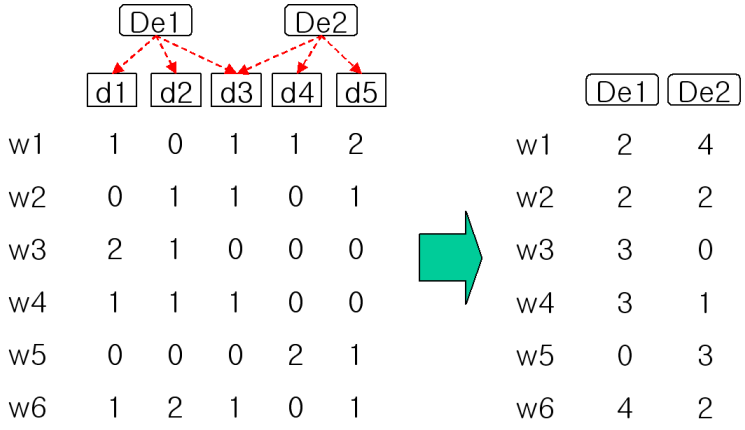
① 문헌 벡터 생성: 분석 대상 문헌(제목 및 초록, 또는 본문)들을 자동색인하고 단어 가중치를 산출하여 각 문헌을 해당 문헌에 출현한 단어의 벡터로 표현한다.

② 디스크립터 프로파일 벡터 생성: 특정 디스크립터가 출현한 모든 문헌 벡터들에 대하여 센트로이드 벡터를 <그림 3>과 같이 산출해서 해당 디스크립터의 프로파일 벡터로 삼는다.

③ 1차 연관도 행렬 생성: 디스크립터 프로파일 벡터 사이의 코사인 유사도를 계산해서 1차 연관도 행렬을 산출한다. 이 과정은 동시출현단어분석이나 저자동시인용분석에서 동시출현행렬을 산출하는 단계에 해당한다.

④ 2차 연관도 행렬 생성: 1차 연관도 행렬의 각 벡터 사이의 피어슨 상관계수나 코사인 계수를 다시 계산하여 2차 연관도 행렬을 산출한다. 이 과정은 동시출현단어분석이나 저자동시인용분석에서 피어슨 상관계수 행렬을 산출하는 단계에 해당한다.

⑤ 다변량분석 및 매핑: 1차 연관도 행렬이나 2차 연관도 행렬에 대해서 군집분석, 다차원척도법, 패스파인더 네트워크 알고리즘 등의 분석 기법을 적용하여 2차원 공간에 매핑한다.

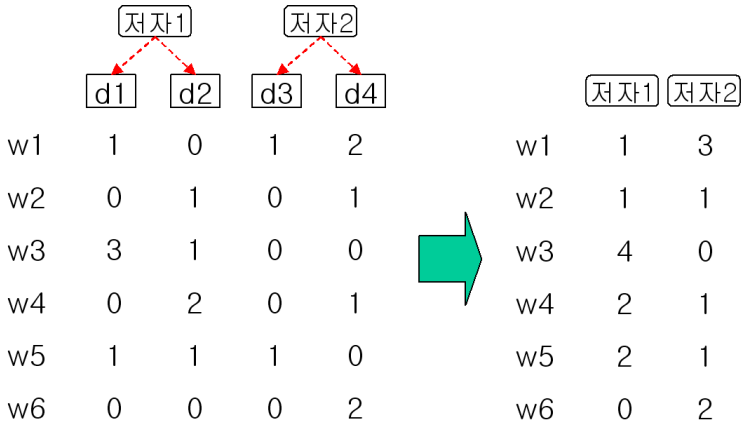


De : 디스크립터, d : 문헌, w : 단어,

↘ : 문헌에 디스크립터 할당

<그림 3> 문헌 벡터로부터 디스크립터 프로파일 벡터 생성 예

디스크립터가 부여된 문헌이 아닌 특정한 저자가 발표한 문헌의 문헌 벡터를 결합하면 <그림 4>와 같이 저자 프로파일을 생성할 수 있다. 이와 같이 저자 프로파일과 디스크립터 프로파일을 생성하면 저자나 디스크립터를 단위로 하는 지적 구조 분석이 각각 가능할 뿐만 아니라, 두 가지 종류의 프로파일 벡터 사이의 유사도를 산출함으로써 저자와 디스크립터 사이의 관계를 파악하는 것도 가능하다. 즉 특정 저자의 주요 연구분야를 파악하거나, 특정 주제의 주요 연구자를 파악할 수가 있다.



d : 문헌, w : 단어,

↘ : 저자가 발표한 문헌

<그림 4> 문헌 벡터로부터 저자 프로파일 벡터 생성 예

4. 디스크립터 프로파일링을 이용한 정보학 분야 분석

디스크립터 프로파일링 기법의 효용성을 확인하기 위해서 정보학 분야를 대상으로 1995년부터 2006년까지 최근 12년간의 지적 구조를 분석하였다.

4.1 데이터 수집 및 가공

분석 대상 문헌과 디스크립터를 얻기 위해서 정보학 분야의 핵심 학술지를 선정하였다. 즉 Journal Citation Reports의 데이터에 기반하여 2001년부터 2006년까지 6년간 영향력 지수, 즉시성 지수, 인용 반감기 측면에서 공통적으로 상위 5위 이내에 있는 *Journal of the American Society for Information Science & Technology*, *Information Processing & Management*, *Journal of Documentation*의 3개 학술지를 정보학 분야의 핵심 학술지로 선정하였다.

실험을 위한 데이터는 LISA 데이터베이스로부터 수집하였다. 먼저 출판년도를 1995년부터 2006년까지로 제한하고 선정된 3개 학술지에 수록된 전체 논문을 검색한 결과로 2,245건의 레코드를 다운로드하였다. 다음으로 각 레코드를 대상으로 디스크립터(DE) 필드에서는 색인자가 부여한 디스크립터(통제어휘: 1,565종)를 추출하였다(<표 1> 참조). 또한 제목(TI)과 초록(AB) 필드에서는 저자가 작성한 것으로 본문의 단어(비통제어휘: 6,433종)를 추출하였으며, 특히 이들 단어에 대해서는 불용어 제거와 스테밍을 통한 사전 처리를 수행하였다(사전 처리 절차에 대해서는 김판준(2006)을 참조). 이외에 저자(AU) 필드에서는 저자 프로파일링에서 사용될 저자 2,908명을 추출하였다.

추출된 전체 디스크립터 중에서 12년간 18회, 즉 연평균 1.5회 이상 사용된 디스크립터 109종을 선정하여, 이중에서 주제 표현력이 미약한 디스크립터 24종을 제외하고 남은 85종을 분석대상으로 하였다. 여기서 제외된 디스크립터는 전체의 40% 이상에 출현하여 특정성이 너무 낮은 'searching'과 'online information retrieval'의 2종과, 'world wide web', 'evaluation', 'research'처럼 주제 중립적인 디스크립터 22종이다.

<표 1> 문헌빈도 상위 디스크립터

디스크립터	빈도
searching	648
online information retrieval	628
world wide web	291
evaluation	148
citation analysis	133
information seeking behaviour	130
computerized information storage and retrieval	111
research	109
automatic text analysis	108
bibliometrics	104

4.2 프로파일 생성 및 프로파일 연관성 산출

추출된 6,433개 단어에 대하여 log TF × IDF 가중치를 산출하여 검색된 2,245건의 문헌 벡터를 표현하였고, 3장에서 제시한 절차에 따라 디스크립터 프로파일링을 수행하였다. 먼저, 디스크립터 프로파일 벡터는 특정 디스크립터가 부여된 문헌 벡터들의 합이 되는 것으로서, 직접 작성한 Visual Foxpro 프로그램으로 <표 2>와 같이 생성하였다. 다음으로 이러한 디스크립터 프로파일 행렬을 SPSS 프로그램에 입력하여 각 프로파일 벡터 간의 코사인 유사도를 산출하여 <표 3>과 같은 디스크립터간 1차 연관성 행렬을 생성하였다. <표 3>의 1차 연관성 행렬에서 각 행(혹은 열) 벡터 사이의 코사인 유사도를 SPSS 프로그램을 통해 다시 산출하여 <표 4>의 2차 연관성 행렬이 생성되었다.

한편 <표 5>는 디스크립터 프로파일링 기법과의 비교를 위해 제시한 것으로 전통적인 동시출현단어분석에서 사용되어 온 디스크립터 동시출현빈도 행렬이다.

<표 2> 디스크립터 5종의 디스크립터 프로파일 (중간 생략)

디스크립터 / 단어	citation analysis	information seeking behaviour	computerized information storage and retrieval	automatic text analysis	bibliometrics
2-dimension	0	0	0	0	5.28
3-dimension	4.47	0	0	0	0
aamulehti	0	0	0	0	0
aat	0	0	0	0	0
ab	0	0	0	0	5.69
abandon	0	0	0	0	0
abbott	0	0	0	0	0
abbrevi	0	0	13.95	11.46	0
abc	0	0	4.95	0	0
abel	0	0	0	0	0
abi	0	0	0	0	0
abid	0	0	0	0	4.69
abil	0	6.39	5.43	4.77	0
abl	1.77	2.65	2.38	3.54	1.77
aboard	0	0	0	0	0
abort	5.69	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮
yule	0	0	4.28	7.24	0
yule-simon	0	0	0	0	8.93
z39	0	0	0	0	0
zadeh	0	0	0	0	0
zagreb	0	0	0	0	0
zatocod	0	0	0	0	0
zeal	0	0	0	0	0
zealand	0	0	0	0	0
zero	4.95	0	0	0	0
zero-order	0	0	0	0	0
zhang	0	0	0	0	0
zipf	9.98	2.63	0	2.63	24.75
zipf-distribut	0	0	0	0	0
zipfian	4.95	0	0	0	4.95
zipf-mandelbrot	0	0	0	0	7.28
zipf-type	5.69	0	0	0	0
zone	0	0	0	0	4.69
zoom	0	0	0	0	0
zprise	0	0	0	0	0
z-score	0	0	0	6.28	0

여기서 기존의 단어동시출현분석에서 일반적으로 사용되어 온 동시출현 빈도행렬과 본 연구에서 제안한 디스크립터 프로파일과의 차이를 살펴보면 다음과 같다. 디스크립터 간의 동시출현 빈도에 기반한 <표 5>에서는 디스크립터 'citation analysis'(DE1)가 단순히 'bibliometrics'(DE5) 및 'performance measures' (DE6)와 7회 동시출현한 것으로 동일하게 나타나는 반면에, 프로파일링 기법으로 산출한 2차 연관성 기반의 디스크립터 프로파일 <표 4>에서는 'performance measures'(DE6)보다 'bibliometrics'(DE5)가 'citation analysis'(DE1)와의 연관성이 더 높은 것으로 나타난다. 따라서 디스크립터 간의 연관성 측면에서 변별력이 더 높은 디스크립터 프로파일이 특정 주제 및 주제들 간의 관계를 보다 적절하게 표현할 수 있는 것이다.

<표 3> 디스크립터간 1차 연관성(디스크립터 프로파일 벡터간 코사인) 행렬 (일부)

Descriptor	ID	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8	DE9	DE10
citation analysis	DE1	1.000	0.173	0.176	0.157	0.462	0.319	0.198	0.220	0.183	0.217
information seeking behaviour	DE2	0.173	1.000	0.257	0.141	0.179	0.310	0.248	0.216	0.307	0.316
computerized information storage and retrieval	DE3	0.176	0.257	1.000	0.363	0.178	0.394	0.280	0.154	0.226	0.174
automatic text analysis	DE4	0.157	0.141	0.363	1.000	0.176	0.230	0.211	0.110	0.171	0.137
bibliometrics	DE5	0.462	0.179	0.178	0.176	1.000	0.300	0.172	0.204	0.188	0.193
performance measures	DE6	0.319	0.310	0.394	0.230	0.300	1.000	0.480	0.151	0.235	0.191
search engines	DE7	0.198	0.248	0.280	0.211	0.172	0.480	1.000	0.110	0.277	0.141
information science	DE8	0.220	0.216	0.154	0.110	0.204	0.151	0.110	1.000	0.176	0.256
internet	DE9	0.183	0.307	0.226	0.171	0.188	0.235	0.277	0.176	1.000	0.246
information communication	DE10	0.217	0.316	0.174	0.137	0.193	0.191	0.141	0.256	0.246	1.000

<표 4> 디스크립터간 2차 연관성(디스크립터간 1차 연관도 행렬의 벡터간 코사인) 행렬 (일부)

Descriptor	ID	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8	DE9	DE10
citation analysis	DE1	1.000	0.672	0.670	0.623	0.895	0.786	0.697	0.726	0.712	0.718
information seeking behaviour	DE2	0.672	1.000	0.751	0.627	0.684	0.809	0.776	0.729	0.840	0.818
computerized information storage and retrieval	DE3	0.670	0.751	1.000	0.865	0.688	0.883	0.838	0.654	0.779	0.691
automatic text analysis	DE4	0.623	0.627	0.865	1.000	0.646	0.773	0.745	0.579	0.697	0.614
bibliometrics	DE5	0.895	0.684	0.688	0.646	1.000	0.785	0.696	0.714	0.719	0.709
performance measures	DE6	0.786	0.809	0.883	0.773	0.785	1.000	0.920	0.691	0.816	0.737
search engines	DE7	0.697	0.776	0.838	0.745	0.696	0.920	1.000	0.634	0.810	0.686
information science	DE8	0.726	0.729	0.654	0.579	0.714	0.691	0.634	1.000	0.719	0.769
internet	DE9	0.712	0.840	0.779	0.697	0.719	0.816	0.810	0.719	1.000	0.791
information communication	DE10	0.718	0.818	0.691	0.614	0.709	0.737	0.686	0.769	0.791	1.000

<표 5> 전통적인 동시출현단어분석에서의 디스크립터간 동시출현빈도 행렬 (일부)

Descriptor	ID	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8	DE9	DE10
citation analysis	DE1	133	0	0	0	7	7	2	2	1	4
information seeking behaviour	DE2	0	130	4	0	0	6	3	1	4	6
computerized information storage and retrieval	DE3	0	4	111	2	0	14	2	0	0	0
automatic text analysis	DE4	0	0	2	108	2	0	1	0	0	0
bibliometrics	DE5	7	0	0	2	104	5	1	1	2	1
performance measures	DE6	7	6	14	0	5	78	18	0	0	0
search engines	DE7	2	3	2	1	1	18	75	0	5	0
information science	DE8	2	1	0	0	1	0	0	72	0	3
internet	DE9	1	4	0	0	2	0	5	0	72	2
information communication	DE10	4	6	0	0	1	0	0	3	2	66

디스크립터 프로파일링과 동시출현단어분석을 통해 각각 산출한 디스크립터 사이의 연관성을 더 구체적으로 비교하기 위해 <표 6>과 <표 7>을 제시하였다.

<표 6>은 디스크립터 'user behaviour'와 연관성이 높은 디스크립터를 두 가지 기준에 따라 각각 제시한 것이다. 이중에서 개념적으로 연관성이 높은 'information seeking behaviour'는 디스크립터 프로파일링 기준에서는 2위로 높게 나타나지만, 동시출현 단어분석 기준으로는 상위에 들지 못하고 12위에 그친다.

<표 6> 디스크립터 'user behaviour'와의 연관성 상위 디스크립터 비교

순위	디스크립터 프로파일링 기준	동시출현 단어분석 기준
1	use statistics	use statistics
2	information seeking behaviour	end users
3	end users	performance measures

<표 7>은 디스크립터 'information science'와 연관성이 높은 디스크립터를 역시 두 가지 기준별로 제시한 것인데, 개념적으로 매우 가까운 'library and information science'가 디스크립터 프로파일링에서는 1위로 나타나지만 동시출현 단어분석 기준으로는 상위에 없으며 17위로 매우 낮게 나타난다.

<표 7> 디스크립터 'information science'와의 연관성 상위 디스크립터 비교

순위	디스크립터 프로파일링 기준	동시출현 단어분석 기준
1	library and information science	science and technology
2	library and information science periodicals	scholarly communication
3	information communication	library and information science periodicals

두 사례에서 보듯이 유사한 개념을 나타내는 디스크립터 사이의 연관성은 동시출현단어분

석에 비해 디스크립터 프로파일링을 적용하는 경우에 더 잘 파악할 수 있는 것으로 나타났다.

4.3 디스크립터 매핑과 관련 저자 분석

디스크립터 프로파일링으로 산출된 디스크립터 간의 2차 연관성 행렬에 기반하여 생성한 디스크립터 네트워크는 <그림 5>와 같다. 여기서 네트워크 생성을 위해서는 Visual Foxpro로 구현한 클러스터링 기반 네트워크(Clustering based Network; CBNet) 생성 알고리즘(이재윤 2007)을 적용하였고, 네트워크의 시각화를 위해서는 네트워크 분석 및 시각화 소프트웨어인 Pajek(de Nooy et al. 2005)을 사용하였다.

또한 디스크립터 간의 연관성 정보에 기초하여 네트워크상의 디스크립터들의 군집을 형성하였는데, 본 연구에서는 SPSS 프로그램의 군집분석에서 제공하는 여러 클러스터링 알고리즘 가운데 지적 구조 분석에 널리 사용되는 Ward 클러스터링 기법을 사용하였다. 분석 대상 클러스터의 수를 결정하기 위해서 일반적인 지적 구조 분석 연구에서와 마찬가지로 Ward 클러스터링 결과의 댄드로그램에서 적절한 구분지점을 찾아본 결과 16개와 3개로 나눠지는 지점이 각각 선택되었다. <그림 6>은 <그림 5>의 네트워크에 Ward 클러스터링 알고리즘으로 구분된 16개 작은 군집을 표현한 것이고, <그림 7>은 이들 군집을 더 결집하여 형성한 3개의 큰 군집을 표시한 것이다. 여기서 노드의 크기는 해당 디스크립터가 12년간 3개 학술지 논문에 부여된 횟수에 비례하도록 설정하였다.

전체를 작은 16개 군집으로 나눈 <그림 6>에서 왼쪽 아래에서부터 오른쪽으로 연결되는 간선에 위치한 디스크립터를 위주로 주제 전개를 살펴보면, 계량서지학(bibliometrics, citation analysis)에서부터 정보학(information science), 인터넷(Internet), 정보탐색행태(information seeking behaviour), 탐색전략(search strategies), 정보 저장 및 검색(computerized information storage & retrieval)으로 나타나고, 이로부터 위쪽의 자동 텍스트 분석(automatic text analysis)과 아래쪽의 데이터베이스(databases) 분야로 확장되는 것으로 나타났다.

<그림 7>에서 전체를 3개 군집으로 나눈 결과는 왼쪽에서부터 계량서지학 및 정보학 군집, 디지털도서관 및 이용자 행태 연구 군집, 정보검색과 데이터베이스 및 자동텍스트분석 군집으로 구분되었다. 이는 각각 정보 및 이론 위주의 정보학, 이용자 위주의 정보학, 시스템 위주의 정보학으로 성격을 구분할 수 있다.

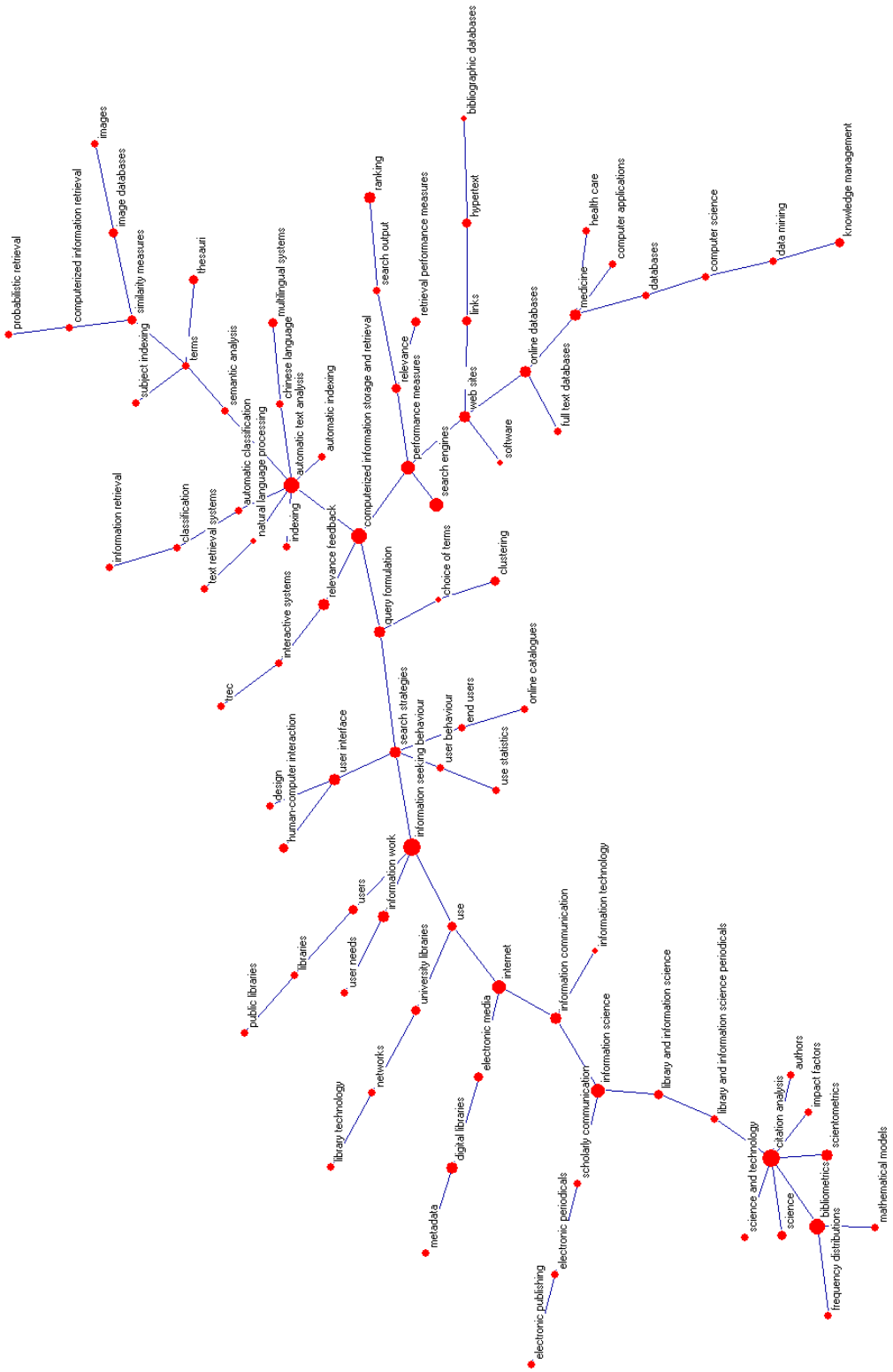
<그림 7>에 표시된 저자는 16개 작은 군집별로 해당 군집의 디스크립터와 연구 주제가 가까운 저자를 파악하여 제시한 것이다. 이를 위해 우선 분석대상 3개 저널에 12년간 6편 이상 발표한 93명의 저자를 추출하여 각 저자의 프로파일 벡터를 구축한 다음, 특정 군집에 속한 디스크립터의 프로파일 벡터와의 연관성이 하나 이상 1위인 저자를 선정하였다. 그림에서 진한 글씨와 밑줄로 표시한 저자는 해당 군집 내에서 둘 이상의 디스크립터와의 연관성이 1위인 저자이다. 예를 들어 계량서지학 및 인용분석에 관련된 군집 C1에서는 H. F. Mode와 R. Rousseau가 2개 이상의 디스크립터에 대해서 연관성이 최고이며 L. Egghe는 하나의 디스크립터에 대해서 연관성이 최고로 나타났다. 저자동시인용분석에서 계량서지학 분야를 대표하는 저자로 항상 등장하는 Price나 Garfield가 빠진 것은 이 연구의 방식이 인용된 저자가 아닌 논문을 발표한 저자를 대상으로 분석하므로 현역 연구자가 아닌 경우는 제외되기 때문이다.

이와 같이 디스크립터 프로파일과 저자 프로파일의 비교를 통해서 파악된 각 저자는 해당

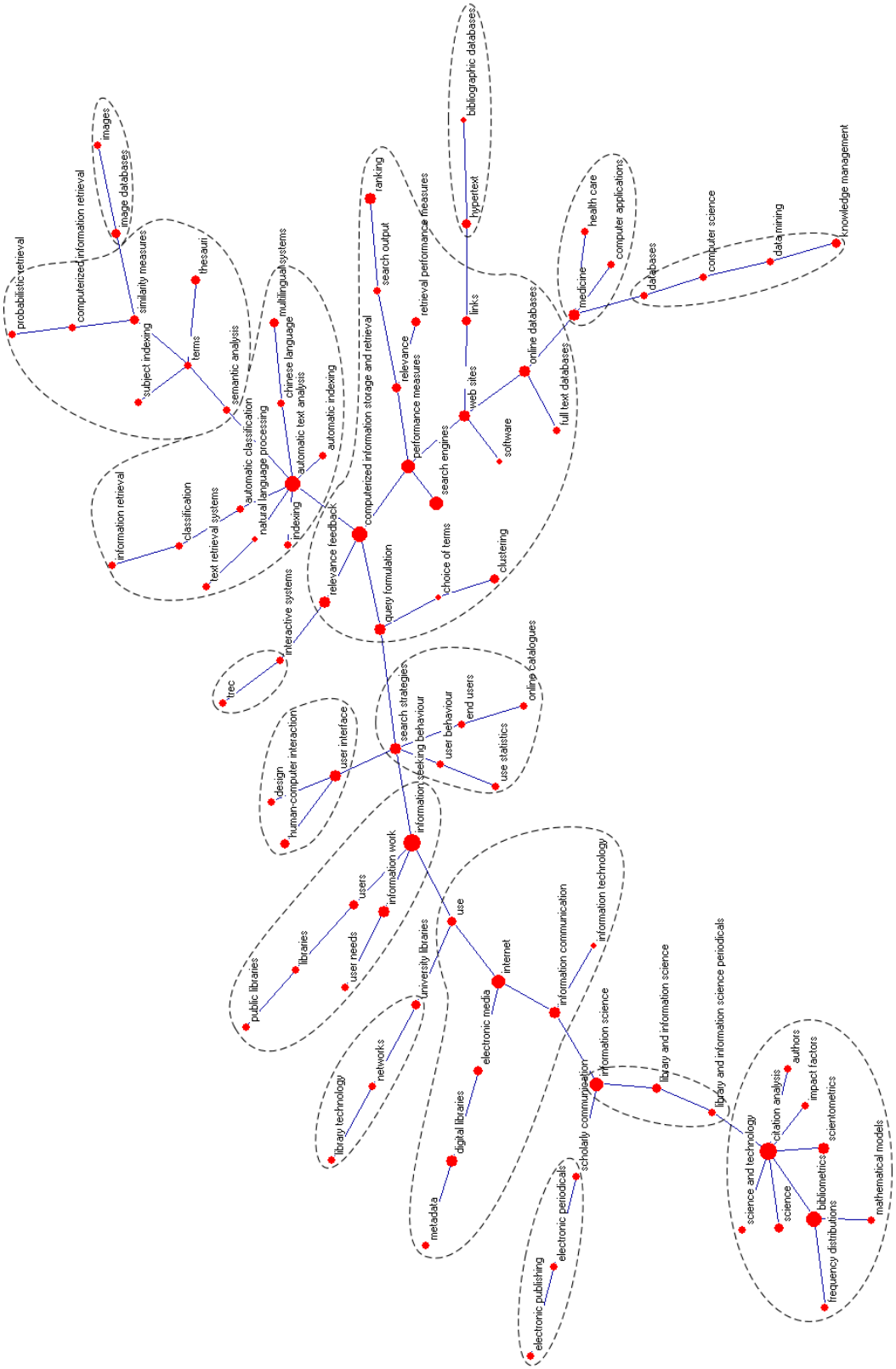
주제 분야에서 활동이 두드러진 현역 연구자라고 할 수 있다. 특정한 주제 군집에서 2개 이상의 디스크립터에 대해서 연관성이 최고로 나타난 저자는 H. Chen을 비롯한 10명으로서 이들을 관련된 디스크립터 군집과 함께 제시하면 <표 8>과 같다. 이 표에 제시된 10명은 앞서 선정된 93명의 저자들 중에서도 특정 정보학 영역에서의 활동이 특히 활발한 연구자라고 할 수 있다.

<표 8> 특정 주제 군집에서 2개 이상의 디스크립터에 대해서 연관성이 최고인 주요 저자와 해당 군집

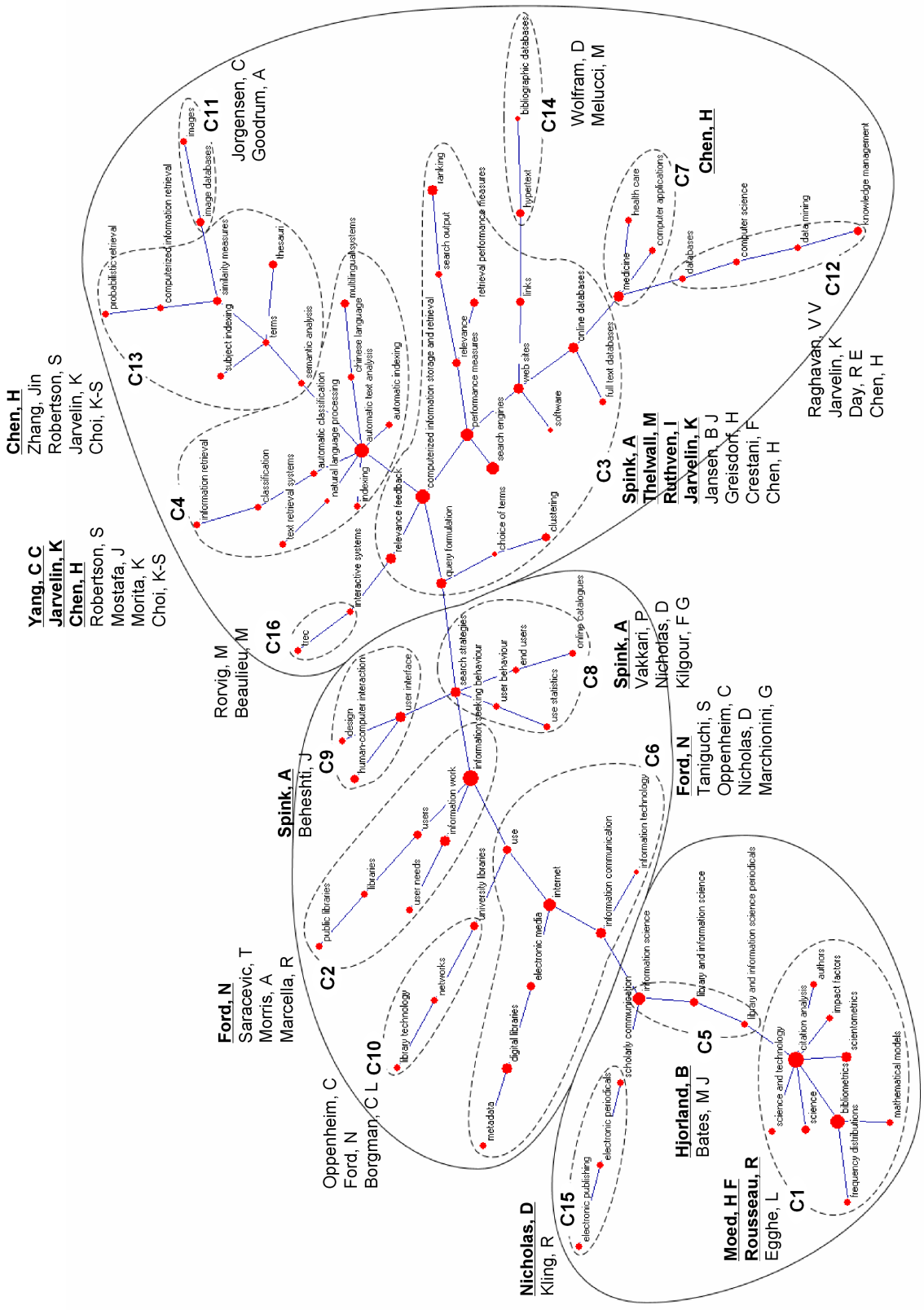
주요 저자	관련 군집 번호 (괄호 안은 군집내 대표 디스크립터)
Chen, H	C4(automatic text analysis), C7(medicine), C13(data mining)
Ford, N	C2(information seeking behavior), C6(Internet, digital libraries)
Jarvelin, K	C3(computerized information storage and retrieval), C4(automatic text analysis)
Moed, H F	C1(bibliometrics, citation analysis)
Nicholas, D	C15(electronic publishing)
Rousseau, R	C1(bibliometrics, citation analysis)
Ruthven, I	C3(computerized information storage and retrieval)
Spink, A	C3(computerized information storage and retrieval), C8(search strategies), C9(user interface)
Thelwall, M	C3(computerized information storage and retrieval)
Yang, C C	C4(automatic text analysis)



<그림 5> 디스크립터 프로파일로 도출한 1995-2006 정보학 분야 디스크립터 네트워크 (CBNet-Ward)



<그림 6> 정보학 분야 디스크립터 네트워크에 Ward 기법에 의한 16개 군집을 표시한 결과



<그림 7> 정보학 분야 디스크립터 네트워크에 Ward 기법에 의한 3개 군집과 16개 군집을 표시하고, 16개 군집별 관련 주요 저자를 표시한 결과

5. 결론

이 연구에서는 통제어휘인 디스크립터를 비통제어휘인 단어와의 동시출현 정보로 표현하는 디스크립터 프로파일링을 제안하였다. 지금까지 특정 주제분야의 지적 구조를 시각화하기 위한 목적으로 단어동시출현분석을 적용한 연구들에서는, 통제어휘(디스크립터, 주제명표목, 분류기호, 통제 키워드 등)와 비통제어휘(표제, 초록, 전문의 단어, 비통제 키워드 등)의 두 가지 상이한 특성을 갖는 자질을 각각 개별적으로 사용하였다. 그러나 본 연구에서는 주제를 표현하기 위한 자질로서 통제어휘인 디스크립터를 사용하지만, 이러한 디스크립터를 비통제어휘인 표제와 초록의 단어로 표현하는 디스크립터 프로파일링을 제안하였다.

제안된 디스크립터 프로파일링 기법의 유용성을 확인하기 위해서 1995년부터 2006년까지 12년 동안의 정보학 분야 핵심 학술지 3종에 게재된 논문 2,245편을 대상으로 분석하였다. 디스크립터 프로파일링 기법으로 파악된 디스크립터 사이의 연관성을 동시출현단어분석 기법으로 파악하는 경우와 비교해본 결과, 기존의 기법으로는 제대로 구분되지 못하는 디스크립터 사이의 관계가 디스크립터 프로파일링 기법으로 파악되는 것으로 나타났다.

디스크립터 프로파일링으로 파악된 디스크립터 사이의 관계에 기반하여 정보학 분야의 전체 지적 구조를 나타내기 위해서 Ward 클러스터링 기법과 CBNet 알고리즘을 적용한 결과, 정보학 분야의 16개 세부 주제와 3개 대주제를 파악하고 네트워크로 표현할 수 있었다. 또한 저자 프로파일과 디스크립터 프로파일의 비교를 통해서 연구 주제와 주요 연구자간의 연계 정보를 함께 파악하여 네트워크 지도에 표현하였다. 이를 통해서 정보학의 주요 주제마다 연구 활동이 활발한 연구자를 파악할 수 있었으며, 이 중에서 특히 연구 활동이 두드러진 10명을 해당 분야와 함께 별도로 제시하였다.

디스크립터 프로파일링 기법은 LISA와 같이 통제어휘가 포함된 색인초록 DB의 데이터에 적용하기가 용이하면서 통제어휘 사용의 문제점을 보완할 수 있으므로, 향후 다양한 분야에 대한 연구 동향 분석에 활용될 수 있을 것이다. 또한 저자 프로파일링 기법을 함께 사용할 경우에는 연구 주제와 관련 연구자를 동시에 파악할 수 있다는 장점이 있어 연구 동향 분석에 더욱 실용적인 도움이 될 것으로 기대된다.

참고문헌

김판준. 2006. 로치오 알고리즘을 이용한 학술지 논문의 디스크립터 자동부여에 관한 연구. 『정보관리학회지』, 23(3): 69-90.

이재윤. 2005. 계량서지적 동시출현 분석에 관한 방법론적 고찰. 한국과학기술정보연구원 내부세미나 발표자료.

이재윤. 2007. 클러스터링 기반 네트워크 생성 알고리즘. 『제14회 한국정보관리학회 학술대회 논문집』, 147-154.

Bhattacharya, Sujit, and Prajit K. Basu. 1998. "Mapping a research area at the micro level using co-word analysis." *Scientometrics*, 43(3): 359-372.

Buter, R.K., and E.C.M. Noyons. 2002. "Using bibliometric maps to visualize term distribution in scientific papers." In *IEEE Proceedings of the 6th International Conference on Information Visualisation*, London, July 2002, pp. 697-705.

Callon, M., J. Law, and A. Rip. 1986. *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*. London: The Macmillan Press Ltd.

Callon, M., Jean-Pierre Courtial, W. A. Turner, and S. Bauin. 1983. "From translations to problematic networks: an introduction to co-word analysis." *Social Science Information*, 22(2): 191-235.

Courtial, J.-P., M. Callon, and M. Sigogneau. 1984. "Is indexing trustworthy? Classification of articles through co-word analysis." *Journal if Information Science*, 9: 47-56.

de Nooy, W., A. Mrvar, and V. Batagelj. 2005. Exploratory Social Network Analysis with Pajek. New York: Cambridge University Press.

Ding, Y., Gobinda G. Chowdhury, and Schubert Foo. 2001. "Bibliometric cartography of information retrieval research by using co-word analysis." *Information Processing & Management*, 37(6): 817-842.

He, Qin. 1999. "Knowledge discovery through co-word analysis." *Library Trends*, 48(1): 133-159.

Jacobs, N. 2002. "Co-term network analysis as a means of describing the information landscapes of knowledge communities across sectors." *Journal of Documentation*, 58(5): 548-562.

Jassens, F., J. Leta, W. Glänzel, and B. De Moor. 2006. "Towards mapping library and information science." *Information Processing & Management*, 42(6): 1614-1642.

Kostoff, R. N. 1993. "Database tomography for technical intelligence." *Competitive Intelligence Review*, 4(1): 38-43.

Kostoff, R. N., R. Tshiteya, K. M. Pfeil, J. Humenik, and G. Karipis. 2005. "Power source roadmaps using bibliometrics and database tomography." *Energy*, 30(5): 709-730.

McCain, Katherine W. 1995. "R&D themes in information science: A preliminary co-descriptor analysis." In M. Koenig and A. Bookstein (eds.), *Proceedings of the Fifth Biennial Conference of the International Society for Scientometrics and Informetrics*, Rosary College, Pine Forest, IL, June 7-10, 1995, (Medford, N.J.: Learned Information, 1995), pp. 275-282.

Noyons, E.C.M., and A.F.J. van Raan. 1998. "Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research." *JASIS*, 49(1): 68-81.

Porter, Alan L., and Scott W. Cunningham. 2005. *Tech Mining: Exploiting New Technologies for Competitive Advantage*. Hoboken, NJ: John Wiley & Sons, Inc.

Spasser, M. A. 1997. "Mapping the terrain of pharmacy: co-classification analysis of the international pharmaceutical abstracts database." *Scientometrics*, 39(1): 77-97.

Tijssen, R.J.W. 1992. "A Quantitative assessment of interdisciplinary structures in science and technology: co-classification analysis of energy research." *Research Policy*, 21: 27-44.

Todorov, R. 1989. "Co-classification analysis for science mapping: an example from superconductivity." In A. F. J. van Raan, A. J. Nederhof, and H. F. Moed. (eds.) *Science and Technology Indicators*. Leiden: DSOW Pr., University of Leiden, pp. 263-270

Todorov, R., and M. Winterhager. 1990. "Mapping Australian geophysics: a co-heading analysis." *Scientometrics*, 19(1-2): 35-56.

Voutilainen, A. 1993. "NPtool. a detector of English noun phrases." In *Proceedings of the workshop on very large corpora*, Columbus Ohio, Ohio University, June 1993.

Watts, R. J., A, Porter, and B. Minsk. 2004. "Automated text mining comparison of Japanese and USA multi-robot research." Ed. by A. Zanasi, N. Ebcken & C. Brebbia. *Data Mining*. WIT Press, pp. 61-74.

Widhalm, Clemens, Ute Gigler, Alexander Kopcsa, and Edgar Schiebel. 2001. "Co-occurrence and knowledge mapping to identify hot topics and key players in the field of mobility and transport." In *Proceedings of the 8th Conference of the International Society for Scientometrics and Informetrics*, Sydney, July 2001, pp. 751-758.