문헌계량분석기법을 이용한 효율적 벤치마킹 대상 분석 연구

Developing a way of searching bench-marking objects through bibliometric analysis technique

이준영*·박진서**·박선영***·고병열****
June Young Lee · Jinseo Park · SunYoung Park · Byoung-Youl Coh

Ⅰ. 서론

연구기관에서 R&D 전략기획(strategic planning)을 수립하는 과정은 연구개발의 목표, 내용, 또는 연구개발의 직접적인 결과물(output)의 활용에 영향을 끼치는 관련 요소들을 찾아낸 뒤, 최적의 목표수준을 설정하고 효과적으로 목표를 달성하기 위해 각 요소들을 전략적으로 연결하여 조정하는 작업이라고 할 수 있다(Day et al., 2000). 따라서 전략기획이 성공하기 위해서는 관련 요소들을 파악하는 기능, 즉 기술동향을 비롯한 시장, 산업의 환경 변화 및 수요를 분석하는 모니터링이 제대로 가동되어야 한다. 특히 목표수준 결정에는 내부 역량에 대한 분석, 수요자들의 요구 따위도 상당한 영향을 끼치게 되지만, 무엇보다도 자신보다 앞서 있는 관련 경쟁자들의 수준 자체가 목표에서 일정한 기준을 제시하고 전략의 내용을 좌우하게 된다(Debackere and Luwel, 2004).

이러한 "벤치마킹" 작업은 경쟁자가 명확한 상태에서는 경쟁자의 현 수준을 정확히 파악하는 것이 주를 이루며 이를 위한 동향 모니터링도 경쟁자의 수준과 이를 달성하는데 가동되는 "best practices"를 분석하는 데에 초점을 맞추게 된다. 하지만 경쟁자가 명확하지 않은 상태, 즉 신규 연구분야를 개척하거나 막연하게 선정된 경쟁자가 정작 무엇을 하는지 잘 알지 못할 때에는 먼저 "누가 무엇을 하는지(who's doing what)"를 파악·분석하는 작업이 필요하다(Porter and Cunningham, 2005). 이러한 성격의 모니터링은 일종의 '경쟁정보분석(CTI, Competitive Technological Intelligence)'이라고 할 수 있으며 최근 연구개발 전략기획에서 점차 강조되고 있는 영역이기도 하다.

그런데 "누가" 그리고 "무엇을" 하는 지를 파악하는 작업은 아직까지 전략기획 담당자들 또는 참여 전문가의 정성적(qualitative)인 판단과 분석에만 의존하는 경우가 많다 (공공연구기관의 중장기 연구계획이나 TRM을 무작위로 들춰보라. 벤치마킹 대상기관은 별다른 경험적 근거 없이 이미 "선험적·규범적"으로 제시되어 있음을 확인할 수 있다). 이들은 대개 주요 관련 경쟁자들을 열거하고 각 경쟁자의 개략적인 수준과 세부 내역을 서로 비교할 수 있도록 도식화하여 표현하게 된다. 그러나 이러한 과정에는 분석자의 주관적인 선입관이나 제한된 지식과 경험으로 인해 경쟁자를 찾고 파악하는 과정이 충분치 못할 가능성이 뒤따를 수밖에 없다. 물론 이러한 지적이 경쟁정보분석은 정성적으로 진행될 경우 흠이 있을 수밖에 없음을 뜻하지는 않는다. 다만 정성적인 분석을 보완해 줄 수 있는 것으로서 정량적인 분석방법을 논의하자는 것이다.

본 연구에서는 한 해에 1백만 건 이상이 누적되면서 과학기술활동과 관련한 항목에 대한 각종 정보를 담고 있는 "기록물"(Van Raan, 2004)로서 과학기술문헌 데이터베이스를 활용하여 벤치마킹 대상을 찾고 파악하는 작업을 실험적으로 시도해보고자 한다. 본 연구는 유사도 계산을 통해 단순히 자신과 비슷한 경쟁자를 검색하려는 시도(Breitzman, 2005)를 넘어서서, 해당 분야의 활동을 매핑하고 이러한 전체 공간에서 경쟁자들이 차지하는 위치를 파악(positioning)하려는 것이다. 이를

* 한국과학기술정보연구원 미래전략팀, 선임연구원, 02-3299-6044, road2you@kisti.re.kr
** 한국과학기술정보연구원 미래전략팀, 선임연구원, 02-3299-6047, jayoujin@kisti.re.kr
*** 한국과학기술정보연구원 미래전략팀, 연구원, 02-3299-6024, sypark@kisti.re.kr
**** 한국과학기술정보연구원 미래전략팀, 선임연구원, 02-3299-6039, cohby@kisti.re.kr

위해 학문분야의 지적 구조(intellectual map)를 파악하기 위해 문헌계량분석에서 자주 활용되는 동시출현단어분석(co-word analysis) 연구를 응용하여 경쟁자들에 대한 분석이 결합될 수 있는 방법을 모색해 볼 것이다.

II. Co-Word 분석의 개관

1. Co-Word 분석 방법의 핵심 요소

Co-Word분석은 단어들이 문헌에서 동시출현하는 빈도(frequency)를 근거로 하여 단어들 간의 유사도를 측정하고, 이를 다시 군집분석(clustering)을 거쳐 서로 관련 있는 단어들이 "뭉쳐진" 군집들, 즉 세부 영역간의 관계를 파악함으로써 특정 학문 분야의 전체적인 지적 구조를 매핑하는 방법이다. Co-Word분석에서 핵심적인 가정은 단어들은 연구개발자가 연구를 수행하면서 사용한 개념(concept)을 나타내고, 연구개발자가 단어들을 한 문헌에서 동시에 연결하여 사용한다는 것은 곧 그러한 개념들이 연관되어 있음을 뜻한다는 것이다(Noyons, 1999). 따라서 Co-word분석에서는 분석의 대상으로 삼는 단어들이 해당 연구개발의 개념을 표상할 수 있도록 정제(cleaning)하는 것이 매우 중요하다고 할 수 있다. 이러한 정제 과정에서 발생할 수 있는 문제점은 계속 언급할 것이다. 여기서는 Co-Word 분석시 따르게 되는 절차들을 관련 문헌들을 종합하여 요약한다.

(1) 분석될 단어를 추출할 항목 결정

과학기술 문헌데이터베이스는 해당 문헌에 대한 수많은 정보요소(information elements)들로 구성된 메타정보를 담고 있다. 문헌 제목, 저자, 저자 소속기관, 초록, 키워드, 인용 정보, 분류 기호 따위가 그것들이다. Co-Word분석에서는 동시출현하는 단어를 문헌 제목에서만 추출할 수도 있고, 키워드만을 대상으로 할 수도 있고, 초록을 포함할 수도 있으며, 심지어는 문헌 전체 내용을 대상으로 할 수도 있다. 만약 문헌 전체 내용을 대상으로 단어를 추출한다면 분석에 포함될 단어의 수는 매우 많아지게 되어 분석 시간이 오래 걸릴 뿐만 아니라 해당 연구를 대표하는 개념(단어)외에 불필요한 것들이 많이 섞이게 될 수 있다. 또한 키워드만을 대상으로 하게 되면 해당 메타정보 작성자(indexer)의 선입관이 개입될 여지가 발생한다(Leydesdorff, 1987). 그렇다고 제목만 선택하게 된다면 이번에는 저자의 의도성, 비표준화 문제가 발생할 수도 있다(Whittaker, 1989). 결국 단어 추출 범위를 결정하는 문제는 여러가지 상황을 고려하여 절충적으로 흐를 수밖에 없게 된다.

(2) 추출된 단어의 정제

단어를 추출하여 빈도별 리스트를 작성하게 되면, 의미 없는 불필요한 단어들을 제거하거나 유사한 단어들을 하나로 묶는 작업 따위가 반드시 필요하다. 일종의 분석될 단어의 표준화 과정으로, 이 과정은 시소러스나 사전에 정의된 규칙에 의거할 수도 있지만 결국 전문가의 판단이 가장 중요한 요소로 작용하게 된다. 또한 적절한 역치(threshold)를 기준으로 일정 빈도 이하의 단어들을 버리는 작업도 중요하다.

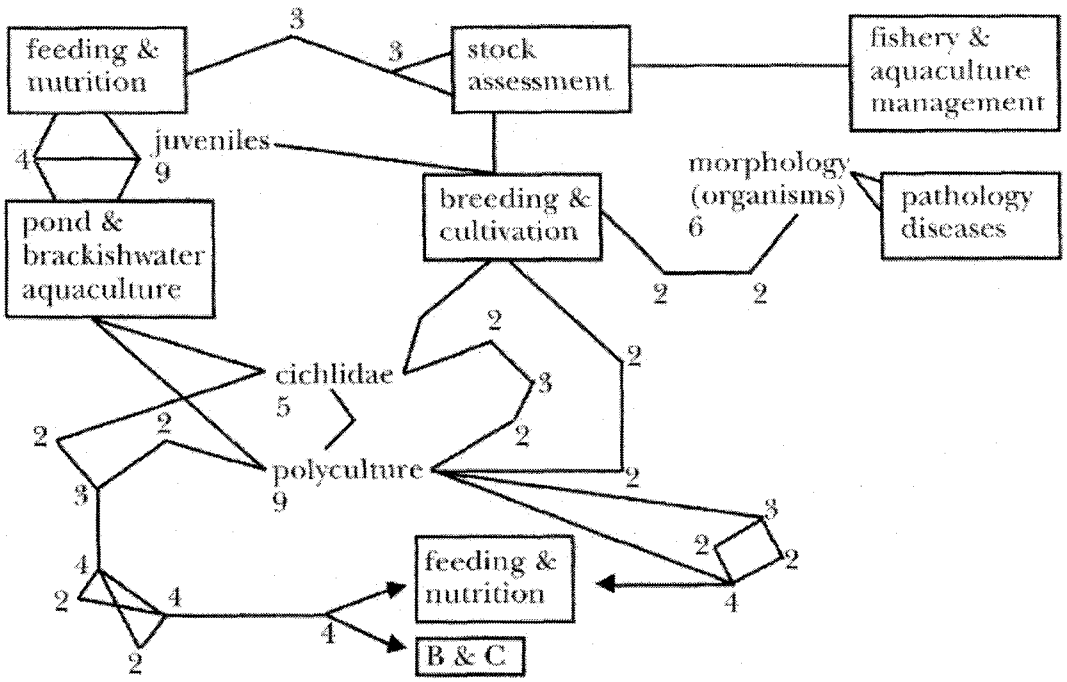
(3) 정제한 단어들의 군집화

정제된 단어에 대해 동시출현행렬(co-occurrence matrix)을 작성하고, 동시출현 빈도값인 셀값을 표

준화하고 유사도[1]로 표시한다. 다음에는 유사도 행렬을 대상으로 군집 분석을 수행한다. 군집분석을 거쳐 나오는 군집들은 세부 영역을 포함하는 것으로 가정하며, 군집에 포함된 단어를 기반으로 적절하게 군집의 이름을 붙인다.

(4) 군집의 시각화 표현 및 해석

각 군집들을 노드와 링크로 구성된 네트워크로 표현하거나 다차원축척(MDS) 등을 이용해 시각화한 뒤, 군집들의 분포와 연관 관계를 해석한다. Callon, Bauin, Coutial 등과 같은 Actor-Network 또는 Techno-Economic Network 이론에 기초를 둔 프랑스 학자들은 1980년대 Co-word분석을 주창하면서 LEXIMAPPE라는 시각화 도구를 개발하였는데, 아래 <그림 1>이 그 예이다. 이 그림은 강하게 연계되는 단어군집을 박스로 표현하고, 군집 간 연계 강도는 각 군집에 동시에 속하는 단어의 수로 표시하여 특정 학문분야의 세부 영역간 매핑을 표시한 것이다.



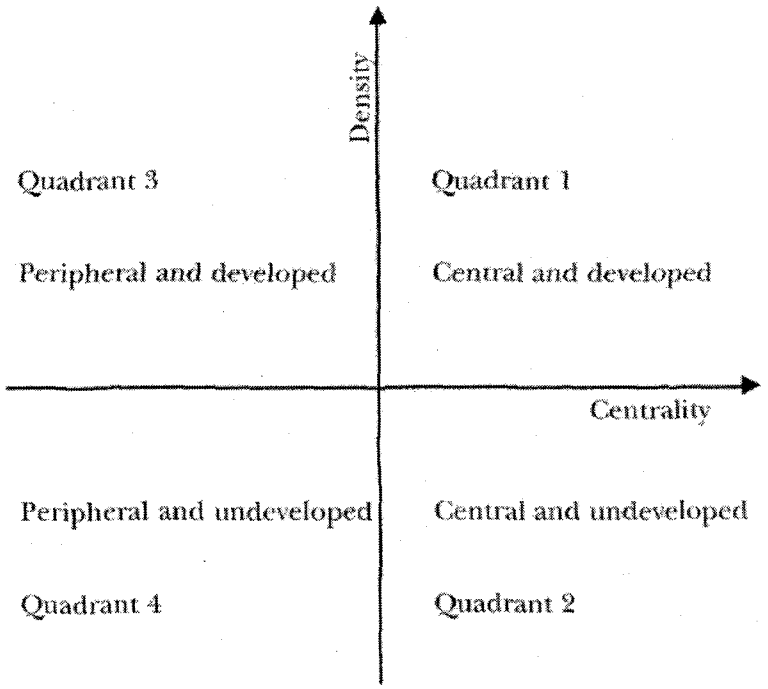
<그림 1> Co-Word 분석의 전형적 시각화 (1981년 프랑스 해양연구 분야) - Bauin (1986)

2. Co-Word 분석의 활용

Co-word분석은 앞에서도 언급했듯이 특정 분야의 지적 구조를 매핑하는 작업에 주로 활용되었다. Noyons & Van Raan (1998)은 인공지능 분야에 대해, Coulter 등(1996)은 소프트웨어 엔지니어링 분에 대해, Courtial & law (1989)도 인공지능 분야에 대해, Ding (2001)은 정보검색 연구 분야에 대해 Co-Word분석을 활용하여 매핑을 하였다. 한편 Co-Word분석에서는 Callon(1991) 이후에 전략다이어그램(Strategic Diagram)을 활용하여 과학기술발전단계에서 해당 분야가 어느 단계에 있는지 판단하게 해주고 있다 (아래 <그림 2> 참조).

전략다이어그램의 기본 원리는 군집 내 단어들 간 연계강도의 평균을 Density로, 군집간 연계강도는 Centrality로 나타내어, 이를 두 축으로 하여 군집의 위치를 매핑하는 것이다. 이에 대한 표준적인 해석은 밀도와 중심성이 낮은 4사분면에 있는 군집은 주변부이고 미성숙한 분야이며, 밀도는 낮으나 중심성이 높은 2사분면에 있는 군집은 주목을 끄는 분야로 활발히 논의되고 있는 중심적

인 위치를 차지하고는 있으나 아직 내부적 성숙도는 덜한 분야이며, 1사분면에 있는 군집은 중심성과 밀도 모두 높아 가장 주류의 연구분야이며, 3사분면에 있는 군집은 내부 성숙도는 높으나 중심성이 떨어져 점점 자체적인 독자적 완성도를 갖추면서 주목을 덜 받게 되는 분야를 뜻한다.



<그림 2> Co-Word 분석의 전략다이어그램 예시

Callon & Courtial (1997)은 Co-Word분석에 대해 해당 분야를 풍부하게 기술하는 단어와 이 단어들 간의 연관관계를 기반으로 하고 있어 분류기호 중심의 분석, 단순 빈도 중심의 분석이 갖는 한계를 극복할 수 있다고 주장한다. 해당 분야를 전체적으로 조망할 수 있을 뿐만 아니라 전략 다이어그램 등을 통해 역동적인 발전 과정을 파악할 수 있는 것이 Co-Word분석의 장점이라는 것이다.

3. Co-Word 분석의 문제점

Co-Word분석을 문헌동시인용분석(Document Co-Citation Analysis, Small, 1973 최초 제안), 저자동시인용분석(Author-Cocitation Analysis, White and Griffith, 1981 최초 제안)과 같이 문헌 또는 저자들이 동시에 인용되는 빈도수를 기반으로 학문 분야를 매핑하는 방법들과 견줄 때 나타나는 가장 큰 차이점은 당연한 말이지만 인용 데이터에 영향을 받지 않는다는 것이다. 하나의 문서에서 다른 두 개의 문서를 동시에 인용했다는 인용관계에 기반하여 문헌간 유사도를 측정하게 되면 인용의 "고전적인 문제", 즉 인용의 동기는 인용자마다 크게 다를 수 있기 때문에 반드시 유사성을 의미하지 않는다는 문제가 발생한다. 다시 말해 논문의 각 저자들은 문헌간 유사도를 측정하기 위한 목표를 갖고 자신들의 글에서 동시인용을 통해 일종의 문헌간 유사도에 대한 "투표"를 한 것은 아니라는 것이다. 어쨌든 인용데이터에 기반한 유사도는 문헌의 저자가 직접 부여하는 것이 아니라 그 문헌을 활용하여 인용하는 다른 저자들이 부여하는 속성을 갖는다.

이에 반해 Co-Word분석은 저자들이 "직접" 작성한 제목, 초록, 키워드 등에서 동시 출현하는 단어들의 빈도를 기반으로 분석하므로 동시인용분석에서 나타나는 문제는 없는 것으로 일견 보일 수 있다. 더구나 앞서 언급했듯이 (Callon & Courtial, 1997) 인용관계가 단순히 관계성만을 나타내

는 데 비해 Co-Word분석은 관계성 외에도 분석대상인 단어들 자체로 주제에 대해 풍부한 정보를 제공해 준다고 주장된다. 하지만 여전히 Co-Word분석도 여러 문제가 나타난다. Leydesdorff(1987)가 지적하듯이 동시 출현하는 단어의 연결관계는 문장, 단락 등의 맥락에 따라 현저히 다를 수 있고, Callon(1986) 스스로 인정하듯이 메타정보인 키워드나 초록을 작성한 색인자(indexer)의 영향을 완전히 제거할 수도 없는 것이다. 그러나 무엇보다도 Co-Word분석의 가장 큰 문제는 분석 대상으로 삼는 단어들을 추출하여 정제하는 과정에서 임의성이 개입될 수 있다는 점이다. 단어들을 비슷한 것으로 판단하여 묶거나, 상관없는 것으로 여겨 제거하는 과정에서 해당 분야에 깊은 이해를 갖고 있지 못한 분석가들이 실수를 할 가능성은 매우 큰 것이다.

III. Co-Word 분석을 활용한 벤치마킹 대상 선정

1. Co-Word 분석과 경쟁자 분석의 결합

지금까지 Co-Word분석이 갖고 있는 장점과 한계를 간단히 살펴보았다. 분석자의 임의성이 개입될 수 있는 소지가 있으나 그럼에도 Co-Word분석은 대용량의 문헌계량분석에서 상당한 중요성을 차지할 수밖에 없다. 왜냐하면 현실적으로 인용분석은 Web of Science라는 인용관계 정보를 거의 독점하고 있는 특정 데이터베이스에 종속될 수밖에 없으나, Co-Word분석은 비교적 간단하게 분석대상이 되는 단어를 추출하고 해당 분야의 전문가와 결합하였을 때 상당히 풍부한 설명력을 갖는 매핑을 그릴 수 있게 해주기 때문이다. 아래에서는 Co-Word 분석을 통해 얻게 되는 해당 분야의 매핑 공간에서 경쟁자들, 특히 경쟁자가 기관인 경우 해당 기관들의 위상을 포지셔닝하는 방법을 두 가지로 제시한다.

(1) 기관별 DNA-encoding 분석방법

본 분석방법은 사실상 검색연구에서 활발히 이용되는 벡터공간 기법과 유사한 것으로, "DNA encoding"이라는 표현은 Noyons (1999), Van Raan & Noyons (2002)에서 표현한 “genetic code of publication"에서 따왔다. 이 기법의 적용은 분석대상이 대는 문헌집합에서 잘 정제된 N개의 개념(단어)를 추출하여 N개의 속성(attribute) 공간을 만드는 것에서 출발한다. 이를 통해 각 문헌들에 대해 N개의 개념들의 포함여부(포함은 1, 비포함은 0)를 계산하여 일종의 DNA 코드와 같은 것을 만들 수 있다. 이를테면 10개의 개념이 있을 때 어떤 문헌 D1의 “DNA 코드”는 0010011110이 될 수 있는 것이다.

이렇게 해당 분야의 문헌들에 대해 DNA코드를 부여한 뒤, 이 코드들의 유사도를 계산하여 문헌×문헌인 동시출현 유사도행렬을 만들 수 있다. 이 행렬을 바탕으로 MDS등을 통해 문헌들의 매핑을 수행할 수 있다. 한편 초기의 행렬을 개념×개념인 동시출현행렬로도 변형할 수 있는데, 이때에는 앞 단원에서 설명한 전통적인 개념 간 군집을 통한 매핑을 수행할 수 있다.

그렇다면 위 절차에서 문헌을 기관으로 바꾼다면 해당 분야에서 개념 간 유사도에 기반을 둔 기관들 간의 매핑을 수행할 수 있을 것이다. 그러나 문제는 문헌인 경우에는 초록, 키워드 등과 같은 정식화된 개념(단어)의 추출 대상이 존재한다. 그러나 기관을 매핑하기 위해 초기에 개념공간을 만들기 위한 개념(단어)들은 어디서 추출할 수 있을 것인가? 본 연구에서는 이를 위해 각 기관의 “미션과 사업소개 자료들”을 개념들을 추출할 정보원(information resource)으로 제시하였다. 이 방법을 통해 새로운 연구영역을 개척하는 작업 초기에 관련 기관들에 대한 분류가 명확하지 않고 위치설정에 곤란을 겪을 때 비교적 손쉽게 이를 해결할 수 있을 것이다. 아래에서는 그 사례로 정보분석 관련 기관을 대상으로 한 분석을 소개한다.

(2) 단어군집에 포함되는 연구논문 비중분포 분석방법

표준적인 Co-Word분석 방법은 특정 학문 분야에서 세부 전공, 세부 분야 간의 거시적인 관계도를 확인할 수는 있었으나 그러한 매핑 공간에서 연구주체들이 어떻게 위치하고 있는지를 보여주지는 못했다. 즉 나와 경쟁하고 있는 기관은 매핑 공간에서 나와 어떠한 차이를 갖고 있는지 확인할 수 없는 것이다.

따라서 본 연구에서는 Co-Word분석 방법을 통해 그려지는 매핑된 공간에 속한 군집을 대상으로 해당 군집에 속하는 단어들을 분석대상 기관의 연구 비중/분포와 연결하였다. 즉, Cluster A라는 군집에 {a1, a2, a3, a4, a5} 5개의 키워드가 묶였다면, X라는 기관이 분석대상기간 동안 작성한 전체 논문들 가운데 얼마나 A에 속하는 논문이 나왔는지를 확인할 수 있다. X가 해당 기간동안 N_x 개의 논문을 발표했고, Cluster A에 해당하는 논문이 N_a_x 개라면 Cluster A에 해당하는 X기관의 상대적 비중은 당연히 N_a_x ÷ N_x가 될 것이다.

그러나 이러한 기관별 상대적 비중은 전체의 분포에 대하여 정규화할 필요가 있고, 이를 위해서는 국가별 논문 비중에 대한 지표(Schubert and Telcs, 1989)인 활동지수(Activity Index)를 응용할 수 있을 것이다. 즉, 특정 군집에서 해당 기관의 활동 지수 AI는 아래와 같이 구할 수 있다.

$$AI = \frac{\text{해당기관의 대상 기간동안 전체 발표 논문에서 특정 군집의 비중}}{\text{대상 기간 동안 해당 분야의 전세계 발표 논문에서 특정 군집의 비중}}$$

AI지수는 1 값을 기준으로 1보다 높으면 해당 세부 군집(분야)가 세계 평균 비중보다 더욱 치중하고 있음을, 1보다 낮으면 세계 평균비중보다 낮은 비중으로 연구를 투입하고 있음을 뜻한다. 이를 통해 Co-Word분석을 통한 세부 연구분야 매핑에서 경쟁 기관간의 연구비중을 손쉽게 파악할 수 있을 것이다. 아래에서는 천문학 연구기관을 대상으로 간단한 분석을 제시하였다.

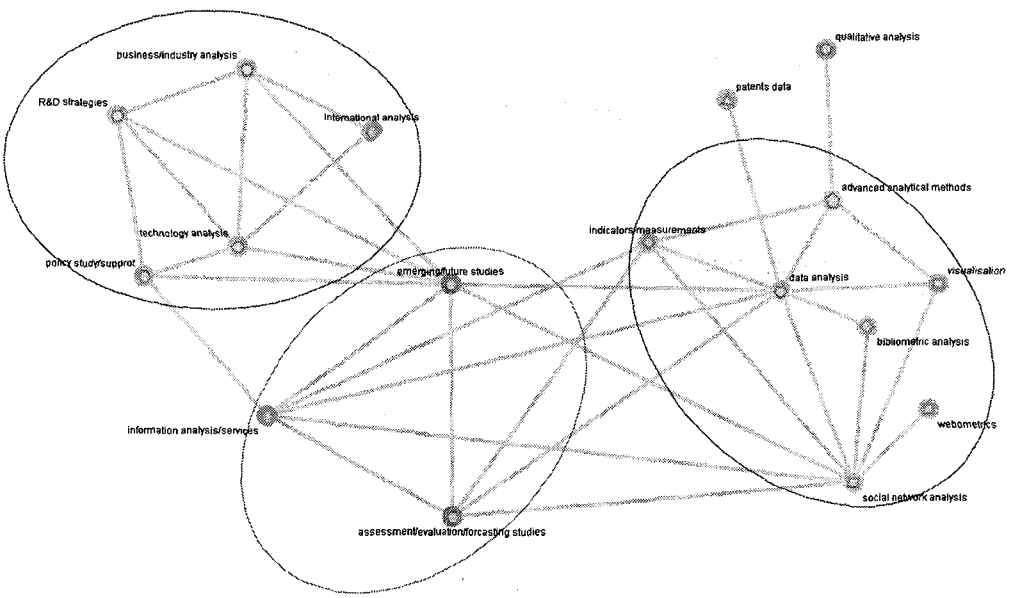
2. 사례 (1) : 정보분석기관의 매핑

정보통신연구진흥원 (IITA)	Institute for the Future (IFTF)	영국 총리실 산하 미래전략청 (STRUN)	영국 DTI의 기술예측사업 (FOREUK)
정보통신정책연구원 (KISDI)	RAND (RAND)	오스트리아 ARC System Research GmbH (ARC)	영국 DTI의 Global Watch Service 사업 (GWS)
한국특허정보원 (KIPI)	세계미래학회 (WFS)	네덜란드 라이덴대학교 CWTS (CWTS)	EU NEST (NEST)
LG경제연구원 (LG-ERI)	Frost & Sullivan (FROST)	독일 FIZ Karlsruhe (FIZ)	OECD IFP (IFP)
SRIC-BI (SRIC-BI)	노무라종합연구소 (NRI)	인도 TIFAC (TIFAC)	The role of Europe in world-wide science and technology monitoring and evaluation in a context of global competition (EWSTME)
Georgia Tech TPAC (TPAC)	미쯔비시종합연구소 (MRI)	대만 STIC (STIC)	Evaluation of Scientific & Technological Innovation and Progress in Europe, through Patents (ESTIPEP)
ipIQ (CHI Research 후신) (IPIQ)	일본과학기술정책연구소 (NISTEP)	Georgia Tech TPAC의 TOA 연구 (TOA)	European Indicators, Cyberspace And The Science-Technology-Economy System (EICSTES)
미국 상무부 기술관리국 (TA)	일본 JST R&D 전략센터 (CRDS)	미국 AAAS의 EurekAlert 사업 (EurekAlert)	Web indicators for scientific, technological and innovation research (WISTIR)
TechCast (TECHCAST)	신에너지산업기술종합개발기구 (NEDO)	SRIC-BI의 Scan™Monthly 사업 (SCAN)	Critical events in evolving networks (CEEN)
MIT Technology Review (TechReview)	캐나다 CISTI (CISTI)	미국 NIST의 SBIR 사업 (SBIR)	Identification and Assessment of Promising Emerging Technological Fields in Europe (IAPETFE)

<그림 3> 정보분석센터 관련 기관 매핑 대상 목록

이 사례는 본 연구자의 소속 조직인 정보분석센터를 대상으로 분석을 수행한 것이다. KISTI 정보분석센터는 급변하는 환경에 대응하여 최근 기존 업무를 혁신하고 새로운 타겟 영역을 "동향모니터링 및 계량분석에 기반한 미래유망기술 발굴"로 설정한 바 있다. 또한 이러한 영역과 관련된 기관들에는 어떠한 것들이 있고, 기관들 간의 연계는 어떻게 되는지를 분석하여 향후 전략수립에 반영코자 하였다. 이를 위해 각 내부 직원들이 자신들의 지식, 또는 외부 전문가의 도움을 얻어 해당영역과 관련있다고 판단되는 국내외 기관 40개를 추출하였다 (<그림 3> 참조).

그러나, 위 40개 목록은 말 그대로 목록일 뿐이며, 각 기관들이 어떠한 성격을 갖고 어떤 영역에 위치하는지 정확히 파악할 수는 없었다. 이를 위해 각 기관들의 홈페이지, 소개자료를 수집한 뒤 여기서 256개의 키워드를 추출하였다. 이 256개의 키워드는 다시 너무 보편적인 단어 또는 정보분석과 관련성이 낮다고 판단되는 단어를 제외하고, 유사한 키워드는 묶어 최종 23개의 키워드를 추출하였다.



<그림 4> 정보분석 영역의 Co-word분석을 통한 매핑

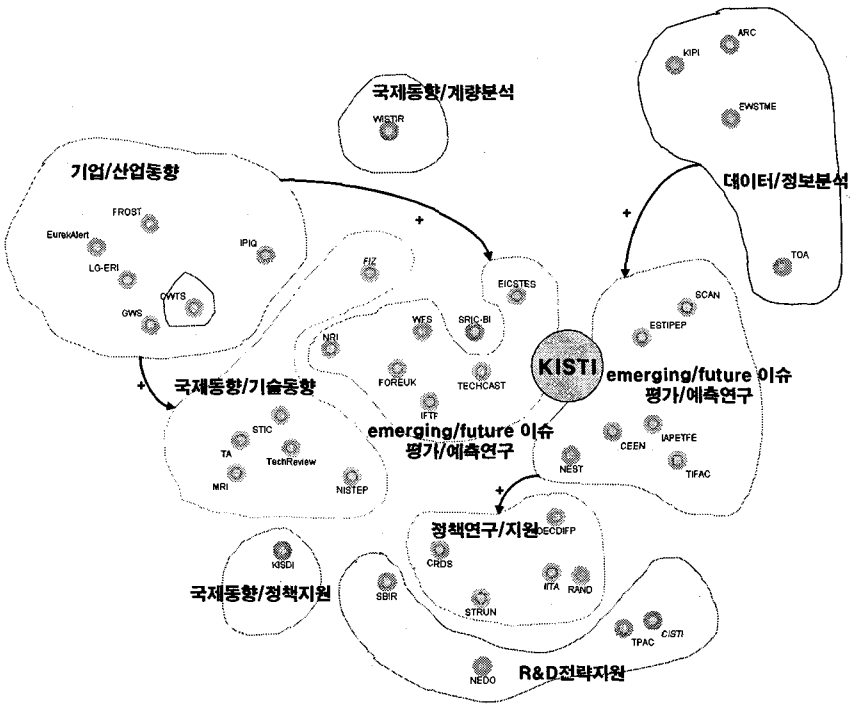
<그림 4>는 단어×단어 동시출현행렬을 Cosine 정규화방법을 통해 유사도를 계산하고, Ward 군집분석을 수행한 결과이다. 결과의 시각화는 MDS를 이용하지 않고, 노드와 링크의 분포를 최적화하는 네트워크 시각화의 일종인 Force Directed 레이아웃을 적용하였다[2]. 이 그림을 통해 대략 크게 3가지 영역, 즉 추출된 키워드를 통해 설정되는 정보분석 영역은 전략/산업/정책분석, 유망영역발굴/예측분석, 정보계량분석으로 매핑되는 것을 확인할 수 있다.

이어서, DNA-encoding 방식으로 앞서 언급한 방법을 적용한다. 27개의 키워드가 40개의 기관들에 어떻게 "encoding"되는지를 확인하고, 이를 통해 얻은 기관×키워드 출현행렬에서 유사도를 계산 한 뒤, 위와 같이 Cosine 정규화, Ward 군집분석을 수행한 뒤, MDS 방식으로 시각화 시켰다. 이 결과는 <그림 5>이고, 이것을 간단히 도식화 한 것이 <그림 6>이다.

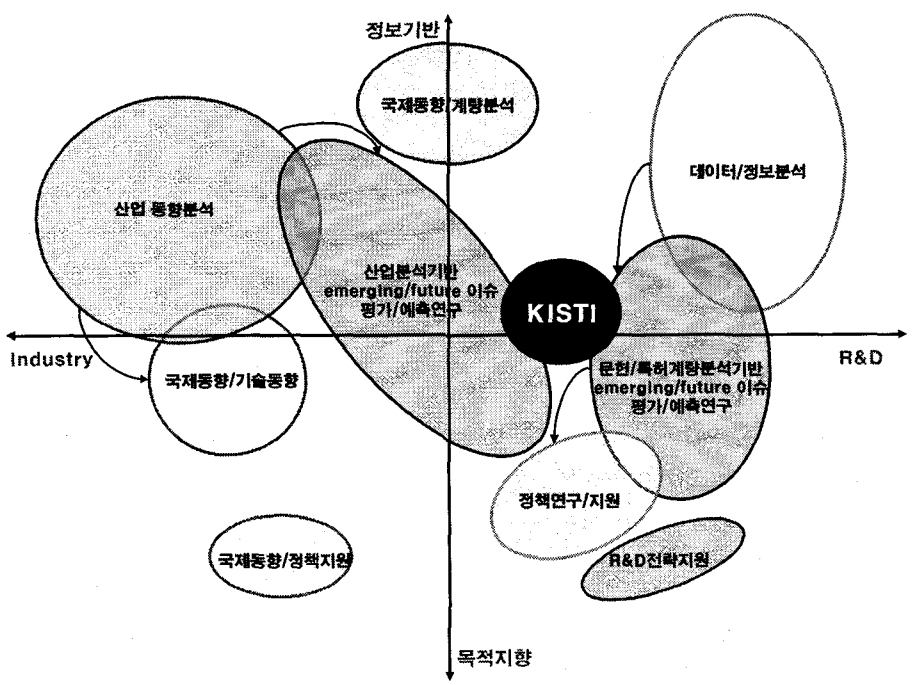
매핑된 기관의 성격을 검토하여 MDS의 가로축과 세로축을 각각 "산업⟷R&D", "정보기반⟷목적지향"으로 구분할 수 있었다. 결과로 나타난 매핑에서는 정보분석의 영역, 특히 미래유망연구영역 발굴과 같은 영역은 크게 2가지 영역, 즉 데이터 기반연구와 문헌계량분석 연구 중심의 영역과 산업/시장 동향분석에 기반한 기업전략수립을 중심으로 하는 영역이 있음을 확인할 수 있었다.

KISTI 내부직원과 전문가를 동원해 추출한 기관들이 이와 같이 다소 구분되는 두개의 중심영역

을 갖는 다는 것은 무엇을 뜻하는 것일까? KISTI 내부 구성원 또는 외부 전문가가 단일한 배경을 갖지 못하고 두개의 대립되는 배경을 가지고 있다는 것인가? 아니면 진정 미래유망연구영역이라는 새로운 영역은 대립되는 듯이 보이는 두개의 영역을 "융합"해야만 가동될 수 있다는 것인가? R&D와 사업화 기획이 하나로 통합되는 이른바 "R&BD"라는 최근의 흐름은 후자의 해석을 지지하는 것으로 조심스럽게 판단을 내릴 수 있을 것이다.



<그림 5> 정보분석 관련 기관 매핑

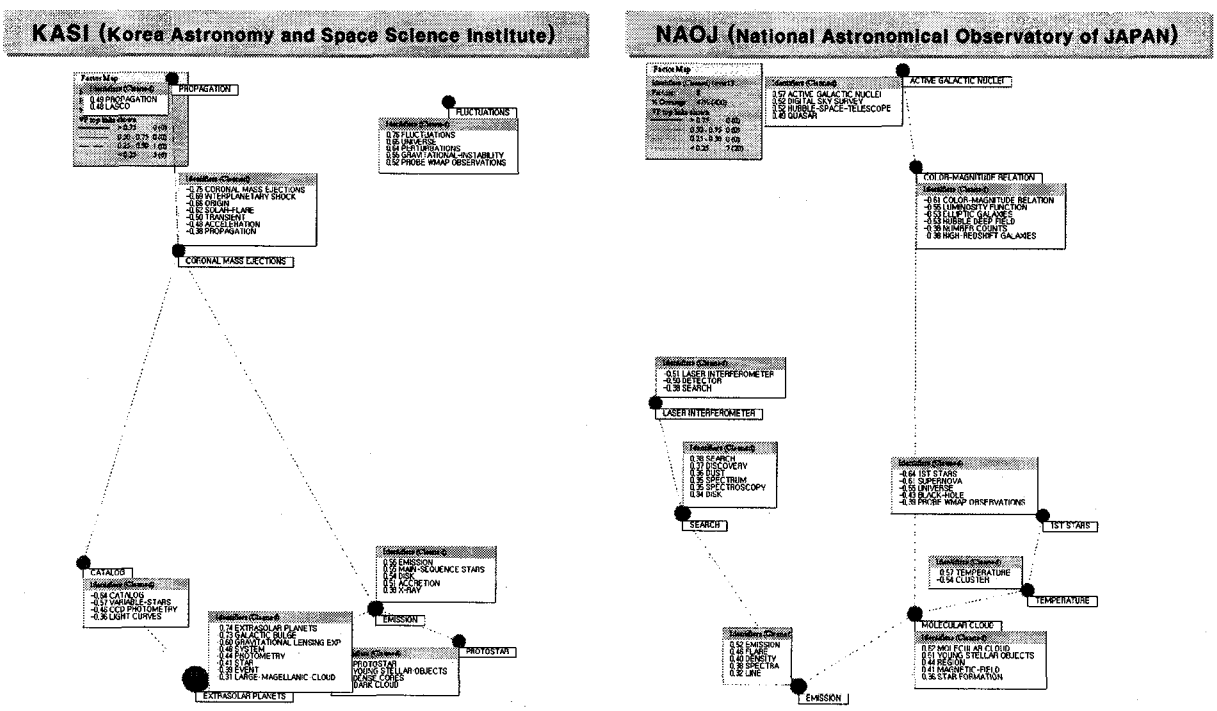


<그림 6> 정보분석 관련기관 매핑 도식화

3. 사례 (2) : 천문학 경쟁연구기관의 상대적 연구비중 파악

본 사례는 우리나라의 유일한 천문학 관련 정부출연연구기관인 천문연구원(KASI, Korea Korea Astronomy and Space Science Institute)과 이와 유사한 일본의 국립천문연구원(NAOJ, National Astronomical Observatory of JAPAN)을 대상으로, 세계적 수준에서 천문학분야의 세부연구분야 간 매핑을 고려했을 때, 두 기관의 포지셔닝을 비교 분석하고자 한 것이다.

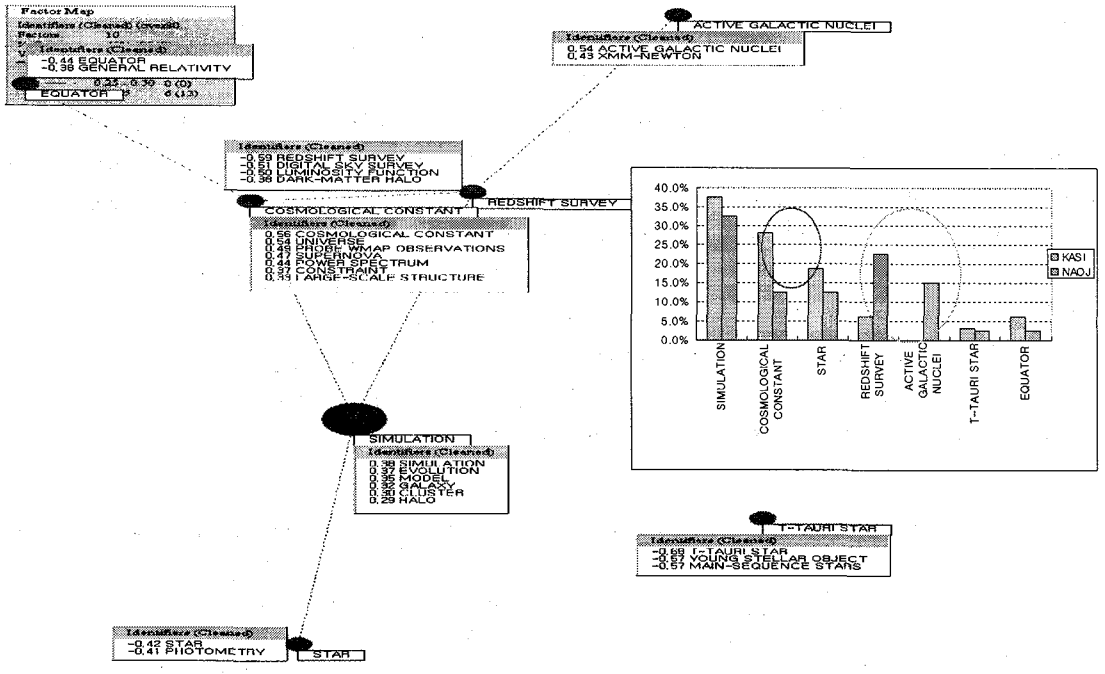
먼저, 표준적인 Co-Word분석을 통해 각 기관별로 중점연구영역을 매핑할 수 있다. <그림 7>은 2003~2005년 3년간 KASI와 NAOJ 각 기관이 발표한 논문(Web of Science DB수록)에서 키워드를 추출하고, 이를 정제하여 군집분석하여 나타낸 것이다. 이러한 Co-Word분석은 특정 기관에 대하여 세부 연구영역의 분포와 관계를 확인할 수는 있으나, 우리가 원하는 "세계적 영역을 포괄하는 연구분야 매핑 속에서 두 기관 간 연구비중"을 확인할 수는 없다.



<그림 7> KASI와 NAOJ의 Co-word분석을 통한 연구분야 매핑 비교

따라서, 먼저 필요한 데이터는 특정 기관에 대한 논문이 아니라 분석대상 기간 동안 천문학 분야에서 발행된 주요 데이터를 확보하는 것이 필요하다. 이를 위해 천문학 분야에서 Impact Factor가 일정 수준 이상의 주요 저널들을 대상으로 논문데이터를 수집한다. 그리고, 이를 기반으로 Co-Word분석의 절차를 밟아 전세계를 포괄한 천문학 연구분야를 매핑한다. 다음은 앞서 언급했듯이 두 기관이 전체 천문학연구 매핑시 발생한 군집에서 차지하는 비중을 계산한다. 이때 두 기관의 비중은 세계 수준에서 특정 군집의 비중을 기준으로 정규화하는 것이 필요하다[3]. 이를 통해 최종 <그림 8>을 얻게 되었다. 이 그림에서는 세계 수준보다 더 높은 (값이 1 이상) 비중으로 연구하는 분야, 경쟁기관에 비해 비중이 적은 분야 등을 손쉽게 확인할 수 있다.

덧붙여, 전 세계 수준의 매핑에서 군집의 갯수를 조절하면 좀 더 세부적인 논의가 가능해 질 것으로 기대된다. 또한 Ding(1999)과 같은 연구자처럼 먼저 군집 간의 거시적 관계를 확인한 후 특정 군집을 대상으로 군집을 구성하는 단어들을 가지고 다시 MDS를 그릴 수도 있을 것이다. 이 경우에도 기관관 논문 비중을 나타내는 AI를 추출하는 것은 가능할 것이다.



<그림 8> 전세계 천문학 연구분야 매핑시 각 군집에서 두 기관의 상대적 비중 분포

Ⅳ. 마치며

본 연구는 연구개발의 전략기획시 벤치마킹 대상이 불명확할 때, 또는 벤치마킹 대상 간 관계에 대한 지식이 불충분할 때, 문헌데이터베이스를 기반으로 Co-Word분석을 변형 응용하여 벤치마킹 대상에 대해 개략적인 연관관계를 나타내는 연구분야 매핑을 작성하고 대상 기관의 상대적 연구 비중과 분포 등을 파악할 수 있는 방법을 시험적으로 제시하였다. 당연한 언급이겠지만, 이러한 작업으로 벤치마킹이 끝나는 것은 아니다. 정밀한 벤치마킹을 통한 "best practice"의 습득을 위해서는 정성적인 정보수집·분석, 전문가의 자문이 필수적이다. 또한 Co-Word분석과 같이 데이터베이스의 정보항목들 간의 관계적 속성에 기반한 분석 외에도 정규화된 논문발표수/ Impact Factor 등의 보완적인 계량분석 또한 필요할 것이다.

이미 여러 차례 언급했듯이 Co-Word분석의 한계는 분명하다. 그러나, 그러한 한계가 분석자체의 유용성 자체를 훼손하지는 않는다. 단어 추출과 정제, 유사도 변환 및 군집분석 등의 통계분석 등을 소홀하게 생각하지 않고 꼼꼼하게 확인하고 분석한다면 Co-Word분석은 전략기획을 보조하는 상당히 매력적인 장치로 기능할 수 있을 것이다.

<참고문헌>

Ahlgren, P., et. al. "Requirements for a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient", *JASIST*, 54(6), pp.550-560

Bauin, S. (1986), "Aquaculture: A Field by Bureaucratic Fiat", in Callon, M., Law, J., and Rip, A. eds. *Mapping the Dynamics of Science and Technology*, London: Macmillan Press, pp.124-141

Breitzman, A. (2005), "Automated Identification of Technologically Similliar Organizations", *JASIST*,

56(10), pp.1015-1023

Callon, M., Law, J., and Rip, A. (1986), "Qualitative Scientometrics", in Callon, M., Law, J., and Rip, A. eds. *Mapping the Dynamics of Science and Technology*, London: Macmillan Press, pp.124-141

Callon, M., Courtial, J. P. and Laville, F. (1991), "Co-Word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemistry", *Scientometrics*, 22(1), pp.155-205

Callon, M. and Courtial, J. P. (1997), "Using Scienometrics for Evaluation", in Callon, M. Larédo, P. and Mustar, P. eds., *The Strategic Management of Research and Technology*, Paris: Economica, pp.165-219

Coulter, N. Monarch, I. Konda, S. Carr, M. (1998), "Software Engineering as Seen through Its research Literature: A Study in Co-Word Analysis, *JASIS*, 49(13), pp.1206-1223

Courtial, J. P. and Law, J. (1989), "Co-Word Study of Artificial Intelligence", *Social Studies of Science*, 19, pp.301-311

Day G. S., Schoemaker, P. J. H., and Gunther, R. E. (2000), *Managing Emerging Technologies*, New York: John Wiley and Sons

DeBackere, K. and Luwel, M. (2004), "Patent Data for Monitoring S& Portfolios", in Moed, H. F. , Glänzel, W. and Scmoch, U. eds., *Handbook of Quantitative Science and Technology Research: The Use of Publication and patent Statistics in Studies of S&T Systems*, Dordrecht: Kluwer Academic Publishers, pp.569-586

Ding, Y. and Chowdhury, G. G., Foo, S. (2001), "Bibliometric Cartography of Information Retrieval Research by Using Co-Word Analysis", *Information Processing and Management*, 37, pp.817-842

He, Q. (1999), "Knowledge Discovery through Co-Word Analysis", *Library Trends*, 48(1), pp.133-159

Leydesdorff, L. (1997), "Why Words and Co-Words Cannot Mapt the Development of the Science", *JASIS*, 48(5), pp.418-427

Noyons, E. C. M. (1999), *Bibliometric Mapping as a Science Policy and Research Management Tool*, Ph.D. Thesis Leiden University, Leiden: DSWO Press

Noyons, E. C. M. and van Raan, A. F. J. (1998), "Monitoring Scientific developments from a Dynamic Perspective: Self-organized Structuring to Map Neural Network Research", *JASIST, 49,* pp.68-81

Porter, A. and Cunningham, S. W. (2005), *Tech Mining,* New Jersey: John Wiley & Sons, Inc.

Schubert, A., Telcs, A. (1989), "Estimation of the Publication Potential in 50 U.S. states and in the District of Columbia based on the Frequency Distribution of Scientific Productivity", *JASIS*, 40(4), pp.291-297

Small, H. (1973), "Co-citation in the Scientific Literature: A New Measure of the Relationship Between Publications", *JASIS*, 24, pp.265-269

Van Raan, A. F. J. and Noyons, Ed C. M. (2002), "Discovery of Patterns of Scientific and Technological Development and Knowledge Transfer", W. Adamczak and A. Nase (eds.) *Gaining Insight from Research Information*, Proceedings of the 6th International Conference on Current Research Information Systems, University of Kassel, August 29-31, 2002, Kassel: University Press, pp.105-112

Van Raan, A. F. J. (2004), "Measuring Science: Capita Selecta of Current Main Issues", in Moed, H. F. , Glänzel, W. and Scmoch, U. eds., *Handbook of Quantitative Science and Technology Research: The Use of Publication and patent Statistics in Studies of S&T Systems*, Dordrecht: Kluwer Academic Publishers, pp.19-50

White, H. D., Griffith, B. C., "Author cocitation: A Literature Measure of Intellectual Structure", *JASIS*, 32, pp.163-172

Whittaker, J.. (1989), "Creativity and Conformity in Science: Titles, Keywords and Co-Word Analysis",

1) Ahlgren 등 (2003)은 동시인용행렬(co-citation matrix)의 셀값을 유사도 계수로 변환하는 방법으로 행렬의 열(row)을 벡터로 상정하여 다른 단어들과 맺고 있는 관계패턴의 유사성을 측정하는 (피어슨 상관계수(pearson correlation) 등을 이용) "global approach"와, 각 항목 간 쌍비교(A, B 각각의 출현빈도와 A와 B의 동시출현빈도를 이용하는)를 유사도 계수로 전환하는 "local approach"를 대비시키고 있다. Co-Word분석 또한 Co-citation과 마찬가지로 두 가지 접근법을 취할 수 있다. Co-Word분석을 최초에 제시한 Callon 등(1991)의 주류적 방법은 Equivalence Index(A와 B의 동시출현빈도의 제곱값을 A의 출현빈도와 B의 출현빈도로 나눈 값)와 같은 local approach를 취하며, 간혹 Ding(2001), Noyons(1999)과 같은 연구자는 피어슨 상관계수 또는 코사인계수를 쓰는 global approach를 취한다.

2) 사례 1에 사용된 분석프로그램은 데이터 정제는 상용프로그램인 VantagePoint, 정규화/군집화/시각화는 KISTI에서 프로토타입으로 개발한 STI Map을 이용하였다. 한편, 사례 2는 모두 VantagePoint를 이용하여 작성한 것이다.

3) <그림 8>에서는 분석데이터의 검증을 거치지 못하여 Impact Factor가 높은 저널 5개, 분석대상기간은 2005년, 상대적 비중의 정규화 없이 AI의 분자에 해당하는 값들을 서로 비교하였다.