서지분석기법을 이용한 신기술 도출 방법론

글 _ 이우형 · 미국 보스톤대학 경영학부 정보시스템연구실 · Leewh@bu.edu

1. 서론

오늘날 지적이거나 정책적인 이유로 과학분야나 사회과학분야에서의 개념, 아이디어, 그리고 문제들 간의 연계성을 도식화할 수 있도록 하는 것은 매우 중요하다. 이러한 도식화를 위해 몇 가지 방법이 시도되어졌다. 과학연구와 과학정책에서 사용되어진 전통적인 방법은, 상대적으로 소수 전문가들의 견해를 구하는 방법이었다(Law and Whittaker, 1992). 서지적 연구(다른 용어로는 bibliometrics 이라고도 불리운다)는 정량적 측면에서 이러한 작업을 수행하기 위한 또 다른 방법이다.

서지분석연구는 비교적 생소한 분야로, 과학문헌에 계량서지학을 적용시킴으로써 일정한 패턴을 찾아내고 설명하는데 중점을 두고 있으며 소스로는 서지학 데이터베이스를 주로 사용하였다. 풍부한 서지학 자료들은 서지분석연구의 몇몇 관점을 지식검색과정과 데이터 마이닝(data-mining)으로 변모시켰다. 또한 도메인 시각화(domain visualization)는 아직까지 거의 발전이 이루어지지 않은 연구분야로, 이는 전체 지식영역을 하나의 분석단위로 취급하던 전통적인 영역분석에서 파생된 분야이다.

도메인 시각화에서는 지식영역의 구조와 원동력을 탐색하여 연구하고 개발하는 과정에서 정보 시각화(information visualization)의 역할을 매우 강조하고 있다. 장래성 있고 주목을 끌고 있는 트랜드는 철학, 사회학, bibliometrics, 정보시각화, 영역분석 등의 여러 전문분야에 걸친 시너지를 통해 그 형태를 잡아가기 시작했다. 이러한 서지분석 방법론을 기반으로 과학지식의 거대한 그림을 그리는 것(본 논문에서는 이를 지식맵(Knowledge Map)이라고 정의한다)은 항상 여러 가지 이유에서 열망되어온 일이다. 전통적 접근법들은 본질적으로 주먹구구식(brute-force)인데, 이는 학자들은 그들의 연구를 수행하기 위하여 학문의 산을 처음부터 끝까지 분류하여야만 하기 때문이다. 분명 이 접근법은 시간 소모적이고, 반복하기 어려우며, 주관적인 방법이다. 게다가 그 복잡성으로 인해 업무량도 방대하다.

최근에 발표된 문헌들 중, 후에 아주 중요한 문헌으로 인식될 것들을 선별하는 것은 매우 노동 집약적인 작업이다. 전통적인 접근법은 정보가 증가하는 속도를 점점 더 따라잡기 힘들어지고 있다. 연구분야가 다학문화(multidisciplinary) 해짐에 따라, 현재 진행상태가 어떤지 개괄적으로 파악하는 것이 어려워졌다. 따라서, 과학기술정보를 분석하는데 있어서 다양한 형태의 분석대상 정보원에서 의미있는 정보 분석 결과를 도출하여 쉽게 이해할 수 있도록 정량적인 분석지표 개발을 위한 방법론 연구가 필요하다. 과학기술의 특징은 복잡하고 서로 다른 학문분야의 서로 다른 지식 도메인으로 구성되어 있고, 상호 관련된 측면이 존재한다는 것이다. 또한 오늘날 과학기술과 관련된 많은 양의 정보가 출판물과 특허에 체화(embedded)되어 끊임없이 증가하고 있다.

그러한 대규모 데이터로부터 구조화된(well-structured) 패턴의 정보를 추출하기 위한 기법을 개발하는 것은 상당

한 도전이다. 현재까지 그러한 패턴은 관련성을 인식하는 데 영향을 주는 숨어있는 특성이며 근원적인 것으로 드러났다. 진보된 서지분석 방법론, 특별히 지식맵(Knowledge Maps)은 많은 가능성을 제공하고 있다. 지식맵은 출판물, 특별히 키워드의 동시 출현(co-occurrence)에 숨어있는 데이터의 통계적 특성을 가시적으로 보여준다. 그러한 지도제작법 표현은 몇 가지 중요한 장점을 가지고 있으며, 거대하고 복잡한 데이터를 가시화(visualization)하는 것은 보다 짧은 시간 안에 좀 더 완성된 개관을 제공한다. 게다가, 그러한 가시화된 표현은 좀 더 쉽게 기억된다. 또한 지식맵의 시계열적 분석은 과학기술의 구조적 개발에 대한 역동성을 드러내준다. 예를 들어, 새로운 활동의 출현, 과학적 도구의 중요성 증가, 합성과 분열과 같은 과학기술분야의 시간흐름에 따른 중요한 변화를 확인할 수 있다.

2. 서지분석기법의 개념 및 종류

2.1 서지분석기법의 개념

Bibliometrics은 일반적으로 논문, 책, 보고서 등이 포함된 명시화된 지식을 대상으로 수학적, 통계학적 방법을 적용하는 것으로(Pritchard, 1969), 분석의 결과는 주로 맵의 형태를 취한다. 그리고 이러한 맵을 여러 학자들은 지식맵이라고 명명하고 사용하고 있다.

이 기법은 과학연구의 본질이 지식의 생산에 있고 과학문헌은 그러한 지식이 표현된 것이라는 관점에서 출발한다. 따라서 과학연구의 결과물인 출판문을 논리적으로 분석하면 연구의 흐름을 표현하는 하나의 지표가 되고, 그러한 지표들이 체계적으로 구성된 다양한 방법들로 나타나게 되는 것이다. 서지분석기법은 다음과 같은 범위에서 응용되고 있다(안규정 · 윤문섭, 2002).

첫째, 과학의 역사분야로 논문의 참고문헌을 이용하여 연구자에 의해 얻어진 결과들을 역으로 추적함으로 특정 과학학문의 발전을 명확히 할 수 있다.

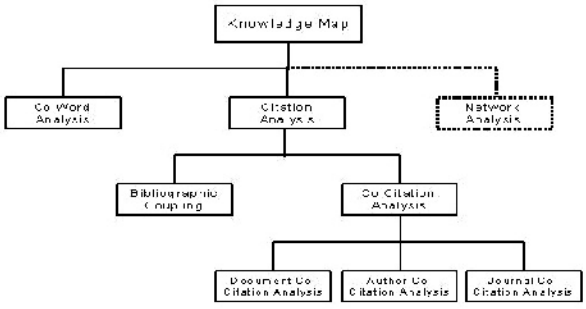
둘째, 사회과학분야에서는 과학문헌을 조사함으로써 연구자간의 네트워크뿐만 아니라 어떤 사회집단의 구조분석을 실증할 수 있다.

셋째, 문서학 분야로 학문분야별 핵심정보를 얻기 위해 필요한 저널의 양, 핵심저널, 정보의 2차 자료 및 학문주변을 구성하는 저널 등을 분석 · 확인할 수 있다.

넷째, 과학정책분야로 서지분석기법을 이용하여 생산성과 과학적 품질을 측정할 수 있는 지표를 제공하여 연구개발 평가 및 방향설정의 기본 자료로 이용할 수 있다.

2.2 서지분석기법의 종류

지금까지 연구된 결과를 바탕으로 지식맵의 종류를 보면 〈그림 1〉과 같다. 지식맵은 크게 동시단어 분석맵과 인용분석맵, 두 가지로 나눌 수 있다. 그러나 사회학에서 활발히 연구되고 있는 네트워크 분석의 결과도 맵의 형태를 취하고 있으며 여기에 사용되는 기본 알고리즘은 서지분석에서 사용하는 알고리즘과 유사하기 때문에 네트워크 분석도 광의의 개념으로 보면 지식맵의 한 종류라고 판단된다.



〈그림 1〉 서지분석기법의 종류

○ **인용분석(Citation Analysis)**

연구활동에서 선행연구는 인용되는 형식으로 제시되는데 인용분석법은 이렇게 인용을 통해 두 논문간의 주제가 동유되는 것으로 두 논문을 연결하는 방법이다. 이 방법은 학술논문에서 인용한 문헌과 인용된 문헌 간에 주제적으로 일정한 관계를 가질 것이라는 가설에 근거한 기법이다.

인용분석에는 서지결합법(bibliographic coupling)과 동시 인용법(co-citation)의 두 가지가 있다. 서지결합법은 두 편의 논문이 지존의 선행논문을 다 같이 인용한 경우, 이 두 논문은 서지적으로 결합되어 있다고 해석한다.

서지적으로 연결된 문헌간의 결합강도는 공통으로 인용된 문헌수로 측정된다. 동시인용법은 두 편의 문헌이 제3의 문헌에 동시에 인용 된 경우, 이 두 편의 문헌은 주제가 서로 관련되어 있다는 가설로서 문헌의 관계나 특정 주제의 구조를 규명하는 기법이다.

○ **동시단어분석(Co-word Analysis)**

동시단어분석은 문서로 제시된 특정 영역 안에서 논문의 주제 분야간의 관계를 확인하기 위해서 단어 또는 명사구 쌍이 동시에 출현하는 패턴을 파악하여 분석하는 방법이다. 단어가 동시에 출현하는 빈도에 기초하여 동시출현 메트릭스를 작성하여 단어간의 관계 강도가 측정된다. 이를 기초로 단어들을 클러스터링하여 맵으로 표현할 수 있다. 이는 해당 도메인의 구조 및 진화의 추적을 가능하게 해 지식발견으로의 중요한 접근수단이 된다.

○ **네트워크분석(Network Analysis)**

네트워크 분석의 목적은 구조나 연결망 형태의 특징을 도출하고, 관계성으로 체계의 특성을 설명하거나 체계를 구성하는 단위의 행위를 설명하는 것이다. 네트워크 분석은 그래프와 네트워크 연구하는 수학의 한 분과인 그래프 이론에서 출발하였다. 그래프는 꼭지점(vertices)과 모서리(edges)로 구성되고, 네트워크는 노드(nodes)와 노드를 잇는 선(links)으로 구성된다. 통신 네트워크, 클럽 회원 네트워크, 통합전기회로 등과 같이 수많은 현상들을 그래프로 표현할 수 있다. 예를 들어 사회적 네트워크를 그래프로 그리는 경우 사람들은 꼭지점으로, 사람들 간의 관계는 모서리로 표현된다.

3. 지식 도메인 시각화(Knowledge Domain Visualization)

지식 도메인 시각화는 빠르게 진보되어온 연구분야이다. 정보시각화에 관한 연구논문의 수는 지속적으로 증가하고 있다(Card, 1996; Hearst, 1999; Herman et al., 2000; Hollan et al., 1997; Mukherjea, 1999). 현재 정보시각화에 관한 여러 가지 책들이 출판되어 있으며(예를 들어, Card et al., 1999; Chen, 1999; Spence, 2001; Ware, 2000) 관련 서적은 그래프 시각화에 대한 알고리즘에 한해서 이용 가능하다(Battista et al., 1999). 또한 Palgrave Macmillan은 2002년에 전문가들이 평가하는 국제적인 저널, "Information Visualization"을 창간하였다. 정보시각화의 목표는 형태, 추세 그리고 다른 징조를 밝힘으로써 그 현상을 나타내는 것이며, 정보시각화는 추상적인 정보에 초점을 두고 있다. 정보시각화의 주요 과제는 비공간적이고 비수치적 정보를 효과적으로 시각화하는 것이다.

이러한 명확한 목표는 다음과 같은 정의로 표현될 수 있다(Card et al., 1999). "정보시각화란 인식의 정도를 확대하기 위해 컴퓨터에 의해 지원되고, 컴퓨터와 상호작용하여 추상적 데이터를 시각적으로 표현하는 것이다".정보시각화는 몇 가지 기본적인 주요 과제에 직면해있는데, 하나는 비공간적 · 비수치적 개념들을 시각적이고 유의한 것으로의 변환할 수 있는 설계에 대한 은유를 제안하는 것이고, 다른 하나는 특정한 은유에 근거하여 설계된 정보시각화의 기능이 실제로 작동하는지를 확인할 수 있는 방법을 찾는 것이다. 과학적 시각화는 정보시각화의 사촌이며, 과학적 시각화와 정보시각화 사이의 경계선은 자주 논쟁거리가 되곤 한다. 과학적 시각화는 사용자 인터페이스, 데이터 표현, 프로세싱 알고리즘, 시각적 표현 그리고 소리 혹은 접촉과 같은 감각적 표현을 모두 포함한다. McCormick et al.(1987)에 따르면 과학적 시각화의 핵심은 "상징주의적인 것을 기하학적으로 변화하는 것"이다. 정보시각화의 궁극적인 목적은 인식의 정도를 높이는 것이다. 정보시각화 과정에는 여러 단계, 즉 다시 말해 데이터를 시각적 형태로 매핑하는 단계, 시각적 구조를 설계하는 단계, 관점의 변환단계가 있다.

데이터를 시각적 형태로 매핑하는 것은 데이터 테이블, 변수의 형태, 그리고 메타데이터의 변환을 포함하고 있다. 정보시각화의 기원으로는 컴퓨터 그래픽, 과학적 시각화, 정보검색, 하이퍼텍스트, 지리학적 정보시스템, 소프트웨어 시각화, 다변량 분석, 인용 분석, 그리고 사회적 네트워크 분석과 같은 다른 사항들이 포함된다. 대량의 데이터를 처리하기 쉽고 유의한 데이터로 변환시키고 추상화시키고자 요구는 시각화 기법의 적용을 유발시키는 요소가 되었다. 다차원적 데이터의 분석은 정보시각화가 적용되었던 가장 초기의 영역 중의 하나이다.

예를 들어, Inselberg(1997)는 평행좌표라고 불리는 시각화 체계를 이용한 정보시각화가 다변량 분석을 어떻게 이

차원의 패턴-인식 문제로 변화시킬 수 있는가에 대하여 설명하였다.

최근까지는 과학문헌 지적구조의 모델링과 시각화는 정보시각화의 주류에 속하지 않았다. 전통적으로 과학매핑과 지적구조매핑에 관한 이슈를 잘 설명하여온 과학분야는 정보과학이다. 정보과학은 두 가지 하부 분야, 정보검색과 인용분석으로 구성된다. 그러나 정보검색과 인용분석은 문서의 해체된 부분에만 집중되어 있다. 정보검색은 제목과 키워드 리스트, 문서의 전체 텍스트와 같이 문서의 서지적 기록에 초점을 두고 있는 반면, 인용분석은 문서에 포함된 참고 링크 혹은 문서의 끝에 참고문헌들에 초점을 두고 있다. 정보시각화의 궁극적인 과제는 내재된 의미론을 전달할 수 있는 강력한 시각적 · 공간적 은유를 고안하고 적용하는 것이다. 일반적으로 정보시각화는 "강력한 응용프로그램(killer application)"을 기다리는 교차로에 서있다고 할 수 있다. 적정한 시각적 은유를 창안하고 선택하는 것은 간단하지가 않다. 문제점을 적정한 설계 은유와 항상 매치시킬 수 있는 분류법은 여전히 사람들의 희망사항이다. 정보검색은 무수한 영감과 과제들을 정보시각화 영역으로 끌어왔다.

또한 이는 분야를 형성하는데 상당한 역할을 수행하였다. 서지분석기법에서의 탐색 목표는 정보를 검색하는 것 이상이며, 과학 지식의 성장과 풀어야 할 주요 문제점들과 지원해야할 중심 임무에 초점을 두고 있다. 과학분야에서 특정 아이템의 발견에 초점을 두는 대신에 서지분석기법에서는 과학적 패러다임과 과학영역에서의 그 움직임이라는 더 높은 수준에 초점을 두었다.

4. 결론

지식맵이라는 용어는 지식을 도식화하고, 마이닝, 분석, 분류하며, 네비게이션과 디스플레이 할 수 있도록 하는 과정을 겨냥해 새롭게 진화하고 있는 과학의 다 학문화 영역을 설명할 때 선호되는 용어이다. 이 분야는 정보접근을 용이하게 하고, 지식구조를 분명하게 하며, 지식탐구자의 탐구 노력이 성공할 수 있도록 하는데 목표를 두고 있다. 비록 이 영역의 존재한지는 수천 년이 되었지만, 최근 15년 동안에 거대한 변화를 겪어왔으며, 이러한 변화는 이용 가능한 정보의 폭발적인 증가와 컴퓨터 저장용량, 속도, 힘의 증가로 인해 가능해진 새로운 분석, 검색, 시각화 기법을 이용해 더욱 쉽게 정보에 접근할 수 있는 접근성에 기인하고 있다. 우리가 위에서 살펴본 연구들로부터 이끌어낼 수 있는 점은 이미 도메인 맵 혹은 시각화라는 것이 널리 사용되고 있으며, 더 나아가 이를 통해 문헌간의 관계를 파악하고, 주어진 학문분야에서 가장 중요한 저자들을 밝혀낼 수 있으며, 지식 영역의 구조와 그 발전단계를 분석할 수 있다는 점이다. 그 방법론으로는 클러스터링, MDS, 요인분석, 그래프 모형에 근거한 사회적 네트워크, 그리고 이 모든 것들의 조합을 들 수 있다. 오늘날 많은 연구들의 공통적인 목적은 비전문가들에게도 충분히 개괄적이고 유익하면서도 다변량 분석기법이나 네트워크 분석을 통해 그 학문분야의 수준별로 자세히 조명해볼 수 있는 도메인 맵을 작성하는 것이다.

참고문헌

안규정 · 윤문섭(2002), 우리나라의 과학수준 및 구조의 특징: SCI 논문 분석을 중심으로, 과학기술정책연구원

Card, S.(1996), Visualizing retrived information: A survey. *IEEE Computer Graphics and Applications*, Vol. 16, No. 2, pp. 63-67

Card, S., Mackinlay, J. and Shneiderman, B.(Eds.).(1999), *Readings in Information Visualization: Using Vision to Think*: Morgan Kaufmann

Chen, C.(1999), Visualizing semantic spaces and author co-citation networks in digital libraries, *Information Processing and Management*, Vol. 35, No. 2, pp. 401-420

Hearst, M. A.(1999), *User interfaces and visualization. In R. Baeza-Yates & B. Ribeiro-Neto(Eds.), Modern Information Retrieval*: Addison-Wesley

Herman, I., Melancon, G. and Marshall, M. S.(2000), Graph visualization and navigation in information visualization: a survey, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 1, pp. 24-44

Hollan, J. D., Bederson, B. B. and Helfman, J.(1997), *Information Visualization. In M. G. Helenader and T. K. Landauer and P. Prabhu(Eds.)*, The Handbook of Human Computer Interaction(pp. 33-48), The Netherlands: Elsevier Science

Inselberg, A.(1997), Multidimensional detective, *Proceedings of IEEE InfoVis'97, October 1997, Phoenix, AZ, USA, IEEE Computer Society*, pp. 100-107

Mukherjea, S.(1999), Information visualization for hypermedia systems, *ACM Computing Surveys*, Vol. 31, No. 6

McCormick, B. H., DeFanti, T. A. and Brown, M. D.(1987), Visualization in scientific computing, *Report of the NSF Advisory panel on Graphics*, Image Processing and Workstations

Pritchard, A.(1969), Statistical bibliography or bibliometrics, *Journal of Documentation*, Vol. 24, pp. 348-349

Spence, B.(2001), *Information Visualization*: Addison-Wesley

Ware, C.(2000), *Information Visualization: Perception for Design*: Morgan Kaufmann