서지데이터 분석 툴에 대한 특성 및 편의성 비교분석

Comparative analysis on the distinctive functions and usability of bibliographic data analysis softwares

이방래, 이준영, 여운동, 이창환, 문영호, 권오진
한국과학기술정보연구원

Lee bang-rae, Lee June Young, Yeo Woon-dong,
Lee Chang-Hoan, Moon Young-Ho, Kwon Oh-jin
Korea Institute of Science and Technology Information

요약

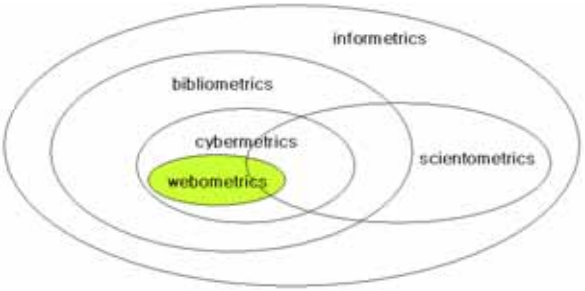
최근에 한국과학기술정보연구원은 계량서지분석에 활용하기 위한 독립형 데이터 분석 시스템 Knowledge Matrix를 개발하였다. 본 논문에서는 개발된 시스템의 성능 수준을 이 분야에서 잘 알려진 분석툴인 VantagePoint와 BibTechMon과 비교분석 하였다.기능 비교는 데이터, 행렬, 분석, 시각화, 데이터 전처리 부문에서 수행 하였다. 분석결과 각 분석툴의 특장점이 서로 다르지만 전반적으로 KnowledgeMatrix가 좀 더 우수한 기능을 보였다.

Abstract

Recently KISTI has developed the KnowlegeMatrix which is a stand-alone type bibliographic data analysis software. In this paper, we try to benchmark test on the performance level of KnowledgeMatrix with well-known S/Ws such as VantagePoint and BibTechMon. We compare distinctive functions and usability of each S/Ws on comparative categories including Data, Matrix, Analysis, Visualization and Preprocessing. Test results show that all S/Ws have differentiated specific feature, but there is some performance gaps. KnowledgeMatrix overally shows better performance than others.

I. 서 론

계량정보분석은 과학기술 성과의 대표적 형태인 논문과 특허 데이터에 내재된 수많은 서지 정보를 다양한 계량분석 기법을 활용하여 새로운 지식을 도출해내는 연구분야이다[1]. 계량정보분석은 특히 SCI(Science Citation Index) DB가 구축된 이후에 더욱 발전하였고 Bibliometrics, Informetrics, Cybermetrics, Webometrics 등의 분야로 세분화 되었다. 각 연구 분야의 포함관계는 다음과 같다[2].



▶▶ 그림 1. 계량정보학 및 하위 범주간의 관계

한국과학기술정보연구원에서는 해외의 유명한 계량분석 관련 분석 툴을 벤치마킹하여 다양한 기능을 제공하는 KnowledgeMatrix를 개발하였다. 벤치마킹한 두 개의 분석툴은 미국 Search Technology사의 VantagePoint[4]와 오스트리아 연구회(ARC)의 BibTechMon[5]인데 이 둘은 계량정보분석 분야에서 널리 이용되는 분석툴이며 특히 Vantage Point는 최근 Thomson사에서 부가 서비스로 제공하고 있다. 본 논문에서는 한국과학기술정보연구원이 개발한 Knowledge Matrix[6]와 벤치마킹 대상인 두 개의 분석툴을 기능면에서 세부적으로 비교분석하였다.

II. 분석 툴 비교분석

[표 1] 분석 툴 종합비교분석표

Function	Detail	VantagePoint (ver. 5.0)	BibTechMon (ver. 4.4.3)	KnowledgeMatrix (ver. 0.9)
Data	Import	RT, RE, X	RT, TM	RT, EM, TM, DM
	Management	P	P, RDB	P
Matrix	Type	O, C, S	o, c, s	O, C, S, D
	Data formation	R	R	R, ER
	Operation	P	P	O, P
	Based on	R, I	R	R, I

	Others	L, s, E	-	L, S, E
Analysis	Clustering	PCA	HC	HC, NHC
Visualization	Chart	LC, OC	-	LC, OC
	Map	FM	FDP, DM	FDP, SD, PFNet
	Options	E, Z, S, N,	M, E, Z, S, N, C, L, CI, CM	M, E, Z, S, R, N, L, C, CI, CM, ED
Preprocessing	Subdata set	DF, RF, SG	-	DF, RF, SG
	Field	MF, FG, CFT, CFS, FR	-	MF, FG, CFT, CFS, FR, FC
	Grouping	NG, LC, GT, GS, TG	-	NG, LC, GT, GS, TG
	Editor	TE, SE, IE	IE	TE, SE, IE
Others		S, M, S	-	M, HK, K

1. 데이터

1.1 반입

분석을 위해 입력하는 데이터의 형태는 보통 DB 검색결과의 다운로드 파일과 사용자가 직접 정보항목간의 관계를 나타낸 행렬 형태의 파일을 고려할 수 있다. VantagePoint는 텍스트(Raw Text; RT) 형태와 엑셀 (Raw Excel; RE) 및 XML(X) 형식의 데이터를 입력받을 수 있다. BibTechMon은 텍스트(Raw Text; RT)와 동시발생행렬 유형의 텍스트(Text Matrix; TM)를 입력 받는다. KnowledgeMatrix는 텍스트(RT) 형태, 엑셀 행렬(EM), 텍스트 행렬(TM), 직접입력방식의 행렬(DM) 형태로 데이터를 입력 받을 수 있다.

1.2 관리

소프트웨어 내부에서 데이터파일을 관리하는 방식과 관련된 부분이다. VantagePoint는 내부적으로 고유한 프로젝트(P) 파일인 *.vpt 형태로 관리된다. BibTechMon은 *.prj 프로젝트 파일로 관리하고 여기에 관계형 데이터베이스(RDB)를 제공하여 다른 형태로 변환하여 이용하기가 편리하다. KnowledgeMatrix도 프로젝트 파일형태로 관리한다.

2. 행렬

2.1 형태

행렬 형태는 출현빈도(Occurrence(O)), 동시출현빈도(Co-occurrence(C)), 유사도(Similarity(S)), 비유사도(Dissimilarity(D)) 형태를 고려할 수 있다. VantagePoint는 Co-occurrence 메뉴에서 출현빈도 추출기능을 포함하고 있으며 TF-IDF를 지원한다. 한편 유사도에 대해서는 동일한 필드를 사용하는 자동-상관(Auto-corrrelation)과 서로 다른 필드를 사용하는 교차상관(Cross-correlation)을 지원하고 있는데 상관 함수로는 피어슨 r과 코사인(cosine), 최대값-비율(Max Proportional)을 지원하고 있다. BibTechMon은 출현빈도, 동시출현빈도, 유사도를 분석에 활용하고는 있으나 사용자가 변환된 값을 확인할 수는 없다. KnowledgeMatrix는 출현빈도, 동시출현빈도, 유사도, 비유사도 형태를 모두 지원하고 있고 직접 그 값을 확인할 수 있다. 또한 모든 출현빈도에 대해서 전치행렬과의 연산을 통해 동시출현빈도 행렬을 생성하는 기능을 제공한다. 한편 KnowledgeMatrix는 유사도와 비유사도를 동시에 지원하는데 유사도지수로 지원하는 것은 코사인, 자카드(jaccard), 다이스(dice), equivalence index, 피어슨, 유클리드(euclidean), squared euclidean, minkowski p-metric 등이다.

2.2 행렬 생성 방식

출현빈도 행렬 데이터 값은 두 가지 관계구조에 근거해서 형성할 수 있다. 첫째는 레코드(Record(R)에 기반해서 출현빈도를 추출하는 것이고 두 번째는 개체관계(Entity-Relationship)(ER)[3] 에 근거해서 출현빈도를 생성하는 것이다. VantagePoint와 BibTechMon은 레코드에 기반해서 출현빈도 행렬을 생성한다. KnowledgeMatrix는 두 가지 방식을 모두 지원하고 있다.

2.3 연산

출현빈도 행렬은 두 가지 형태를 고려할 수 있다. 첫째는 하나의 필드와 레코드와의 관계를 살펴보는 것이다. 둘째는 서로 다른 두 필드가 레코드를 매개로 하여 얼마나 연계되었나를 살펴보는 것이다. 두 번째 형태의 행렬을 만들기 위해서는 행렬연산이 필요하며 이를 위해서 행렬곱(matrix product)과 중첩함수(overlap function)를 고려할 수 있다. VantagePoint는 행렬곱 형태를 지원하고 있고, BibTechMon 은 이진 형태의 데이터에 대해서 행렬곱을 지원하고 있으나 행렬값을 확인할 수는 없다. KnowledgeMatrix는 행렬곱과 중첩함수 모두를 지원하고 있다.

2.4 이진데이터와 계량값

행렬값은 어떤 데이터를 기준으로 추출하느냐에 따라서 2가지 형태로 추출할 수 있다. 먼저 문서에서 값의 존재여부(dichotomous)기준으로 이진데이터를 추출할 수 있다. 이러한 형태는 레코드에 기반해서 데이터를 추출한다. 두 번째는 문서 내 중복 여부에 관계없이 값이 발생한 빈도수를 기반으로 추출할 수 있다. 이러한 형태를 인스턴스(Instance(I))에 기반해서 데이터를 추출했다고 한다. VantagePoint와 KnowledgeMatrix는 레코드와 인스턴스에 대해서 지원하고 있고 BibTechMon은 레코드에 대해서만 지원하고 있다.

2.5 기타

리스트(List(L))기능은 하나의 필드에 대해서 각 항목들이 몇 건의 발생건수가 있는지를 목록 형태로 생성하는 것이다. VantagePoint와 KnowledgeMatrix는 리스트 기능을 지원하고 있다. 한편 행렬에 대한 요약 통계값(S)이 데이터분석에서 중요한 판단변수로 활용될 수 있는데, VantagePoint는 간단한 형태의 통계값만을 제공한다. 반면, KnowledgeMatrix는 다양한 요약통계값을 지원한다. 마지막으로 행렬값에 대한 반출(Export(E))기능을 들 수 있다. Vantagepoint와 KnowledgeMatrix는 행렬값을 파일로 출력할 수 있다.

3. 클러스터링 방법

계량정보분석툴에서는 정보 항목간의 관계를 이용해 그룹핑을 하는 작업이 중요하다. VantagePoint는 주성분 분석(Principal Component Analysis) 기법을, BibTechMon은 계층적 클러스터링(Hierarchical Clustering(HC))을 지원하고 있다. KnowledgeMatrix는 계층적 클러스터링 기법과 비계층적 클러스터링 기법인 K-Means 클러스터링을 동시에 지원한다. 한편 KnowledgeMatrix는 클러스터링 된 결과를 디렉토리 구조로 가시화하여 보여줄 수 있다.

▶▶ 그림 2. KnowledgeMatrix의 클러스터링 결과

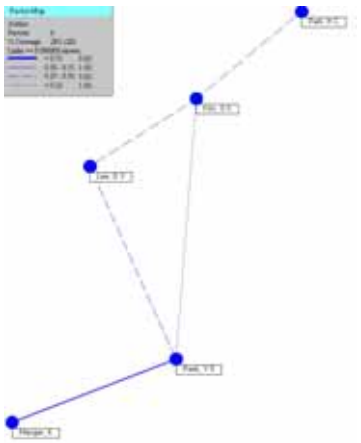
4. 시각화

4.1 차트

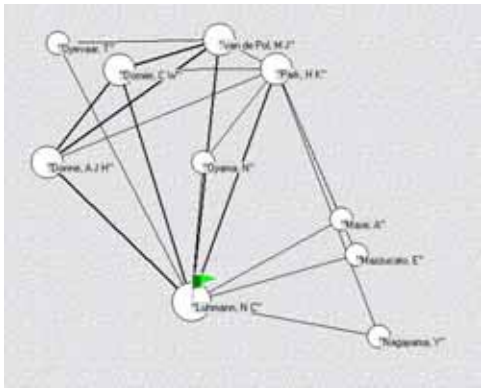
차트 기능에서는 두 가지로 분류해서 볼 수 있는데 필드 하나를 선택해서 보여주는 리스트 차트(LC)와 두 개의 필드를 결합하여 생성한 출현빈도 행렬에 기반한 차트(OC)이다. VantagePoint와 KnowledgeMatrix는 두 가지 형태를 모두 지원한다.

4.2 Map

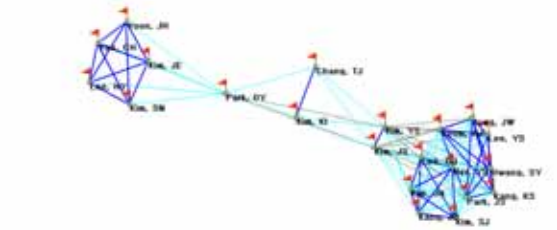
시각화의 가장 대표적인 형태는 FDP(Force-directed Placement)인데 유사도가 높은 항목끼리 가까이 위치시키고 먼 항목은 멀리 위치시켜 매핑을 하는 것이다. VantagePoint는 요인맵(Factor Map(FM))을 지원하고 있는데 FDP의 일종이다. BibTechMon은 FDP와 원형 형태의 덴드로그램을 지원하고 있다. KnowledgeMatrix는 FDP 이외에 전략 다이어그램(Strategic Diagram(SD))과 패스파인더네트웍(PFNet)을 지원하고 있다. 전략맵은 동시단어분석의 결과로 생성된 클러스터들을 내부 연결정도(density)와 외부 연결정도(centrality) 지표를 축으로 2차원 도면에 매핑하여 상대적인 발전 단계를 파악하는 기법이다.



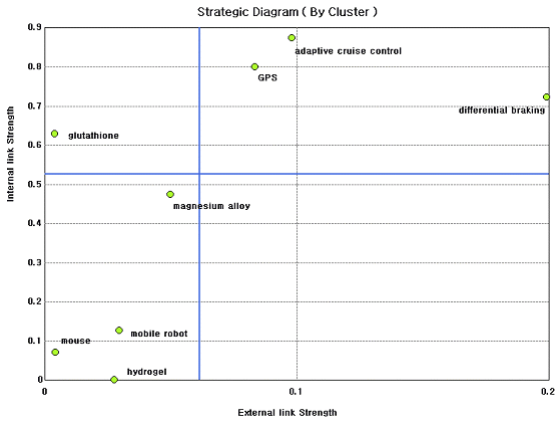
▶▶ 그림 3. VantagePoint Factor Map



▶▶ 그림 4. BibTechMon Map



▶▶ 그림 5. KnowledgeMatrix FDP



▶▶ 그림 6. KnowledgeMatrix Strategic Diagram

4.3 옵션

맵 상에서 추가되는 옵션기능으로 이동기능(M), 저장 및 출력기능(E), 줌 기능(Z), 노드나 라인 선택기능(S), 특정한 노드나 라인만을 선택하여 다시 그리는 기능(R), 노드에 대한 속성 변경(N), 라인에 대한 속성변경(L), 클러스터에 대한 속성 변경(C), 클러스터 구분 기능(CI), 클러스터 간 연결맵(CM), 외부 데이터와의 결합(ED) 등으로 분류할 수 있다. VantagePoint는 시각화 부분에서 옵션기능이 많지 않고 BibTechMon은 시각화옵션 기능이 매우 우수하다. KnowledgeMatrix는 이동기능(M), 저장 및 출력기능(E), 줌 기능(Z), 노드나 라인 선택기능(S), 특정한 노드나 라인만을 선택하여 다시 그리는 기능(R), 노드에 대한 속성 변경(N), 라인에 대한 속성변경(L), 클러스터에 대한 속성 변경(C), 클러스터 구분 기능(CI), 클러스터간 연결맵(CM), 외부 데이터와의 결합(ED) 등의 기능을 지원하고 있다.

5 전처리 기능

5.1 부분 데이터 집합 추출

원래의 데이터 집합으로부터 핵심이 되는 데이터집합만을 별도로 추출하여 정보분석을 하는 경우가 자주 발생한다. 서로 다른 데이터집합으로부터 필드정보를 비교함으로써 특정 필드 정보만 추출하는 기능(Data Fusion; DF)과 레코드를 추출하는 기능(Record Fusion; RF), 그룹정보를 이용하여 부분데이터를 생성하는 기능(Creating Subdata set using Group; SG) 등으로 살펴볼 수 있다. VantagePoint와 KnowledgeMatrix가 이러한 방법을 지원하고 있다.

5.2 필드처리

필드처리에 관한 기능으로는 몇 개의 필드를 결합해서 새로운 필드를 만드는 기능(Merging Fields; MF), 그룹핑된 항목들을 필드로 만드는 기능(Creating Field from Group; FG), 시소러스를 이용한 필드 정제기능(Cleanup Field using Thesaurus; CFT), 스테밍기법을 이용한 리스트 정제기능(Cleanup Field using Stemming; CFS), 검색 및 변경 기능(FR), 클러스터된 결과를 필드로 만드는 기능(Creating Field from Cluster; FC) 등이 주로 이용된다. VantagePoint와 KnowledgeMatrix가 이 기능들을 지원한다. 한편 KnowledgeMatrix는 클러스터링한 결과를 필드로 생성하는 기능(Creating Field from Cluster; FC)을 차별적으로 지원하고 있다.

5.3 그룹핑

그룹핑 기능으로는 항목들을 선택하여 새로운 그룹을 생성하는 기능(New Group; NG), 두 개의 리스트 비교를 통한 그룹핑(Grouping from List Comparison; LC), 시소러스를 이용한 그룹핑(Grouping using Thesaurus; GT), 스테밍 기법을 이용한 그룹핑(Grouping using Stemming; GS), 그룹핑을 이용한 시소러스 생성(Creating Thesaurus using Group; TG) 등이 주로 이용된다. VantagePoint와 Knowledge Matrix가 이러한 기능을 지원한다.

5.4 편집

편집기 기능으로는 시소러스 에디터(Thesaurus Editor; TE), 스트링 편집기(String Editor(SE)), 반입 편집기(Import Editor(IE)) 등이 있다. VantagePoint와 KnowledgeMatrix는 세 가지 기능을 모두 지원하며 BibTechMon은 반입 편집기(Import Editor)만 지원된다.

6. 기타

기타기능으로 도움말(Help; H), 매뉴얼(Manual; M), 스크립트(Script; S), 단축키(Hot Key; HK), 한글처리(Korean language; K) 등을 살펴볼 수 있다. VantagePoint는 도움말, 매뉴얼 기능과 강력한 스크립트(Script; S) 기능을 지원한다. KnowledgeMatrix는 단축키(Hot Key; HK), 매뉴얼을 지원하며 특히 한글 문서에 대한 형태소 분석을 통해 단어나 어절을 추출할 수 있다.

Ⅲ. 추천사항

이 논문에서는 세 가지 정보분석 소프트웨어들의 주요기능들을 비교분석하였다. 각 툴들은 개발된 목적이 조금씩 다르며, 저마다 고유한 특성을 갖고 있기 때문에 절대적인 우수성을 평가하기가 힘들다. 그러나 툴을 직접 활용하는 이용자 측면에서 시스템의 활용 편리성, 지원 기능의 다양성 등을 고려

해 다음과 같이 대략적인 상대 비교를 시도하여 다음과 같이 제시한다.

[표 2] 분석 툴 추천사항

Function	VantagePoint	BibTechMon	KnowledgeMatrix
Data	◎	◎	◎
Matrix	○	x	◎
Analysis	○	△	○
Visualization	△	◎	◎
Preprocessing	◎	x	○
Others	◎	△	○

위의 표 2에서 상대적 등급을 ◎(매우우수), ○(우수), △(부족), x(불량) 등으로 구분하였다. 데이터의 입력과 관리 측면에서는 세 가지 툴이 모두 독특한 장점을 보유하고 있다. 행렬처리는 KnowledgeMatrix가 가장 우수하다. 분석 측면에서는 각 툴이 고유한 목적을 가지고 서로 다른 방법을 채택했기 때문에 어느 것이 가장 우수하다고 단정하기는 힘들다. 시각화 측면에서는 BibTechMon이 매우 우수하다고 할 수 있지만 KnowledgeMatrix는 FDP 이외에 전략 다이어그램 과 패스파인더 네트웍을 지원하고 있어서 대등하다고 할 수 있다. 데이터 전처리 측면에서는 VantagePoint가 가장 우수한 기능을 보유하고 있다. 전체 결과를 종합하면 KnowledgeMatrix가 다른 두 분석툴에 비해 이용자 측면에서 활용도가 더 높을 것으로 판단된다.

▌참 고 문 헌▐

[1] 이우형, IT 중장기 분석, 주간기술동향 통권 1296호 2007.5.16
[2] Thelwall. M. 2003. "A layered approach for investigating the topological structure of communities in the Web." Journal of Documentation, 59(4): 410-429
[3] Morris, S. A., "Unified Mathmatical Treatment of Complex Casdaded Bipartite Networks : The Case of Collections of Journal papers", Oklahoma State University, 2005
[4] http://www.thevantagepoint.com
[5] http://www.systemsresearch.ac.at
[6] http://miso.yeskisti.net