계량정보분석시스템으로서의 KnowledgeMatrix 개발

Development of the KnowledgeMatrix as an Informetric Analysis System

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

Bangrae Lee(brlee@kisti.re.kr), Woon-Dong Yeo(wdyeo@kisti.re.kr),
June-Young Lee(road2you@kisti.re.kr), Chang-Hoan Lee(cheree1@kisti.re.kr),
Oh-Jin Kwon(dbajin@kisti.re.kr), Yeong-Ho Moon(yhmoon@kisti.re.kr)

요약

데이터베이스로부터 지식을 발견하고 이를 연구기획자, 정책의사결정자들이 활용하는 움직임이 전세계적으로 활발해지고 있다. 이러한 연구분야 중 대표적인 것이 계량정보학이고 이 분야를 지원하기 위해서 주로 선진국을 중심으로 분석시스템이 개발되고 있다. 그러나 외국의 분석시스템은 실제 수요자의 요구를 충분히 반영하지 못하고 있고, 고가이면서 한글이 지원되지 않아 국내 연구기획자가 사용하기에 어려운 점이 있다. 따라서 한국과학기술정보연구원에서는 이러한 단점을 극복하기 위해서 계량정보분석시스템 KnowledgeMatrix를 개발하였다. KnowledgeMatrix는 논문 및 특허의 서지정보를 분석하여 지식을 발견하기 위한 목적으로 설계된 독립형(stand-alone) 시스템이다. KnowledgeMatrix의 주요 구성을 살펴보면 행렬 생성, 클러스터링, 시각화, 데이터 전처리로 요약된다. 본 논문에서 소개하고 있는 KnowledgeMatrix는 외국의 대표적인 정보분석시스템과 비교했을 때 다양한 기능을 제공하고 있고 특히 영문데이터 처리 이외에 한글데이터 처리가 가능하다는 장점을 갖고 있다.

■ 중심어 : | 정보분석시스템 | 클러스터링 | 텍스트마이밍 | 데이터마이닝 | 정보계량학 |

Abstract

Application areas of Knowledge Discovery in Database(KDD) have been expanded to many R&D management processes including technology trends analysis, forecasting and evaluation etc. Established research field such as informetrics (or scientometrics) has utilized techniques or methods of KDD. Various systems have been developed to support works of analyzing large-scale R&D related databases such as patent DB or bibliographic DB by a few researchers or institutions. But extant systems have some problems for korean users to use. Their prices is not moderate, korean language processing is impossible, and user's demands not reflected. To solve these problems, Korea Institute of Science and Technology Information(KISTI) developed stand-alone type information analysis system named as KnowledgeMatrix. KnowledgeMatrix system offer various functions to analyze retrieved data set from databases. KnowledgeMatrix's main operation unit is composed of user-defined lists and matrix generation, cluster analysis, visualization, data pre-processing. Matrix generation unit help extract information items which will be analyzed, and calculate occurrence, co-occurrence, proximity of the items. Cluster analysis unit enable matrix data to be clustered by hierarchical or non-hierarchical clustering methods and present tree-type structure of clustered data. Visualization unit offer various methods such as chart, FDP, strategic diagram and PFNet. Data pre-processing unit consists of data import editor, string editor, thesaurus editor, grouping method, field-refining methods and sub-dataset generation methods. KnowledgeMatrix show better performances and offer more various functions than extant systems.

■ keyword : | Informetric Analysis System | Clustering | Textmining | Datamining | Scientometrics |

접수번호 : #071214-003
접수일자 : 2007년 12월 14일
심사완료일 : 2007년 12월 31일
교신저자 : 문영호, e-mail : yhmoon@kisti.re.kr

I. 서 론

정보자원의 기하급수적인 증가에 따라 필요한 정보만을 선별하여 분석하고 전략적으로 활용하는 것이 점점 더 중요해지고 있다. 또한 연구활동 단계에서 정보조사와 분석단계에 소요되는 시간과 비용이 상당한 비중을 차지하는 것으로 조사되고 있다[1]. 이러한 이유로 한정된 자원을 효율적으로 투자하기 위해 최근 각종 국가연구개발사업의 과제기획시 연구동향분석에서 논문 및 특허의 선행조사를 의무화하도록 하고 있다. 한편 이러한 흐름에 발맞추어 2000년대 이후에 계량정보학이 급속히 연구되고 있다. 계량정보학은 과학문헌을 대상으로 어떤 특성을 도출하거나 혹은 특정한 연구문제를 해결하기 위한 수단으로써 과학문헌을 정량적으로 표현하는 연구분야이다. 계량정보학 분야를 지원하기 위해서 주로 선진국을 중심으로 분석시스템이 개발되고 있는데 한국의 연구자/연구기획자가 사용하기에는 여러 가지 문제점이 있어서 한국과학기술정보연구원에서 계량정보분석시스템 KnowledgeMatrix를 개발하였다.

II. 기존 연구

계량정보학이 발전하면서 이를 실제문제에 적용하기 위한 다양한 정보분석시스템이 개발되었다. 대표적으로 미국 조지아텍 대학의 VantagePoint, 오스트리아 연구회(ARC)의 BibTechMon, 미국 인디애나 대학의 CiteSpace 등의 문헌정보 분석시스템이 대표적인 툴이라 할 수 있다. 이 외에도 특허데이터베이스와 연동되어 분석기능을 제공하는 MicroPatent의 Aureka, Delphion Patlab 등도 있다. 그러나 이들 시스템은 몇 가지 문제점이 있는데 많은 분석시스템이 지나치게 특정 DB에 종속적이고, 정밀분석에 필요한 데이터 정제(cleansing)와 자유로운 편집이 불가능하고, 대다수 프로그램이 특정한 기능만을 수행하며, 가격이 고가이며 한글처리가 되지 않아 국내 연구자/연구기획자가 사용하기에는 불편한 점이 많다는 점이다[2].

III. 시스템 구성

1. 개발목표

본 시스템은 국제수준의 정보분석시스템 개발을 위해 VantagePoint와 BibTechMon 수준 이상의 기능을 구현하는 것을 목표로 하였다. 벤치마킹 대상인 VantagePoint는 행렬생성과 데이터 전처리 부분에서 강력한 기능을 보유하고 있고, BibTechMon은 시각화 기능에서 강력한 것으로 평가되고 있다. 본 시스템은 벤치마킹한 두 개의 소프트웨어의 장점을 모두 포괄하고 한글지원이 되며 수요자의 요구를 반영하여 활용로직을 탑재한 시스템 구현을 목표로 한다. 활용로직은 단축키 형태와 전략다이어그램 형태로 구현하였다.

2. 시스템 설계 및 구현

본 시스템을 기능적으로 분리해서 살펴보면 행렬생성, 클러스터링, 시각화, 데이터 전처리 부분으로 크게 구분할 수 있다. 본 시스템을 활용한 데이터 분석작업의 프로세스는 다음의 그림과 같다. 분석 목적으로 수집된 데이터 파일을 입력받고 나면 요약테이블이 생성된다. 전문적인 분석을 위해서는 요약테이블로부터 하나의 필드 정보를 선택한 후 데이터 전처리 과정이 필요하다. 요약테이블이 생성된 이후에는 다양한 행렬(발생행렬, 동시발생행렬, 유사도행렬)을 생성해서 확인할 수 있다. 또한 외부에서 바로 행렬값을 입력받을 수도 있다. 요약테이블과 행렬값으로부터 차트를 만들어서 볼 수 있고 행렬값에 대해서는 요약통계량을 만들 수 있다. 행렬 생성이후에는 클러스터링을 통해 군집의 계층구조를 확인할 수 있다. 마지막으로 시각화 결과로는 FDP[1], 전략맵(Strategic Diagram), 패스파인더네트워크(PFNet) 등을 보여준다. 클러스터링이 필요 없는 경우에는 이 과정을 생략할 수도 있다.

1 Force Directed Placement : 분석 대상 성분의 발생패턴을 비교하여 발생패턴이 유사한 것은 가까이 위치시키고 유사하지 않는 성분들끼리는 멀리 위치하도록 보여주는 그래프 기법.

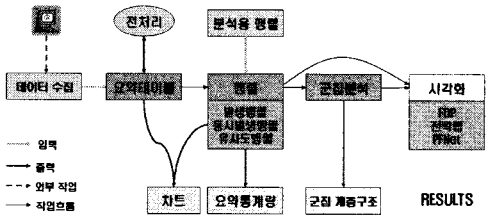


그림 1. 데이터 분석작업 프로세스

2.1 행렬생성

행렬에는 3가지 형태의 행렬이 구현되었는데, 발생행렬(occurrence), 동시발생행렬(co-occurrence), 유사도행렬(proximity)이다.

발생행렬을 생성하는 형태는 2가지로 나눌 수 있는데, 첫째 레코드를 매개로 발생행렬을 만드는 구조로 벤치마킹 대상인 VantagePoint에서 적용하는 방식이다. 두 번째 방식은 Record에 있는 인용문(reference) 정보를 별도로 추출하여 개체간의 관계를 살펴보는 관계로 Morris S. A.[3]가 제안한 방식을 간단한 구조로 수정하여 적용하였다. 다음의 두 가지 그림에서 좌우측과 중앙에 위치한 것은 정보분석시에 선택대상이 되는 각각의 필드이고 이 들을 연결하는 선은 두 필드간의 발생행렬을 표현하는데, 그 유형에 따라서 이진행렬(binary matrix), 계량행렬(valued matrix), 단위행렬(identity matrix) 형태로 구분된다.

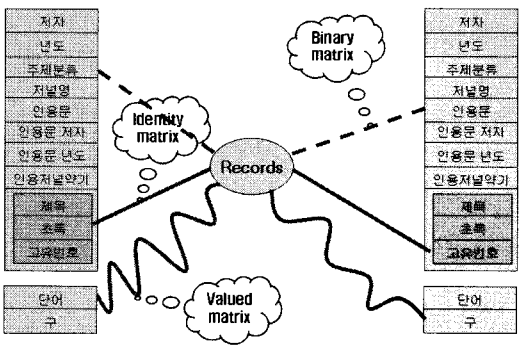


그림 2. 레코드 기반 발생행렬

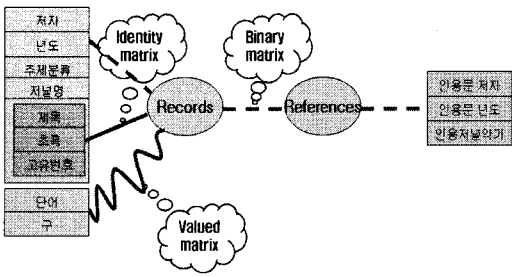


그림 3. 개체-관계 기반 발생행렬

한편 정보분석시에 행렬의 데이터 값을 살펴보고 의사결정을 내리는 경우가 많아서 행렬값에 대한 요약통계값과 행렬값을 이진화 또는 이분화(dichotomization)[2] 하는 기능을 구현하였다.

본 시스템에서 동시발생행렬은 발생행렬과 발생행렬의 전치행렬에 중복함수(overlap function)를 적용하여 계산한다.

유사도(proximity)는 두 값이 얼마나 유사한 가를 나타내는 유사성(similarity)과 두 값이 얼마나 다른가를 나타내는 비유사성(dissimilarity)이 있다. 그런데 데이터의 형태에 따라서 적용 가능한 유사도 계수는 매우 다양하며 본 시스템에서는 다음 표에 제시된 유형만을 적용한다. 통계학자들의 의견을 빌면 피어슨 계수와 코사인 계수는 상관계수의 성격을 가지므로 이진데이터에 적용하면 부적절할 수 있다고 하였으나 실제로 외국의 사례에서도 많이 사용되고 있고 행렬값의 성격에 따라서 사용자가 결정할 사안으로 판단하여 이진데이터에 대해서도 구현하였다.

표 1. 구현한 유사도 계수

구분	association 계수	상관계수	거리계수
이진데이터 (binary)	자카드 (jaccard) 다이스 (dice) 코사인 (cosine) 등가계수 (equivalence)	피어슨 (pearson correlation)	유클리디안 (euclidean) 제곱유클리디안 (squared euclidean) 민코프스키 (minkowski)
숫자데이터 (numerical)	코사인 (cosine)		

2 특정값을 기준으로 0과 1로 행렬값을 변경

한편 앞서 소개한 3가지 행렬 형태에 대해서 외부에서 직접 입력받아서 그 이후의 과정을 분석할 수 있도록 구현하였다. 이는 논문 및 특허의 데이터베이스 파일을 분석하기 보다는 행렬값을 통계분석에 활용하기 위하여 설계하였다.

2.2 클러스터링

본 시스템에서 클러스터링은 계층적 방법과 비계층적 방법을 구현하였는데, 계층적 방법으로 단일연결법, 완전연결법, 평균연결법, 와드연결법 등을 구현하였고[4], 비계층적 방법으로 K-평균군집화 방법을 구현하였다. 와드연결법[5]은 개념적으로 살펴보면 군집간 오차제곱합(Sum of Square Error; SSE)의 증분값 중에서 최소값을 갖는 군집끼리 연결하는 기법인데 실제 구현시에는 일반화된 와드연결법[4]으로 구현하였다. 비계층적 클러스터링 기법인 K-평균 클러스터링은 K개의 군집수와 초기군집중심점을 사용자가 지정하고 계속적으로 거리 계산을 통해서 군집중심점이 바뀌는 방식이다. 본 시스템에서는 이러한 방식을 구현하였고 또한 자동으로 초기군집중심점을 할당할 수 있는 기능도 구현하였는데 이는 대규모 데이터에 적합하다고 하는 MAX-MIN 방식을 이용하였다[6].

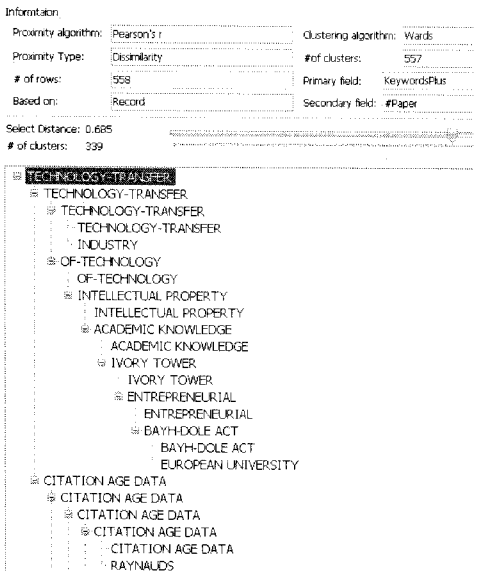


그림 4. 군집계층구조 사례

2.3 시각화

본 시스템에서 제시하고 있는 시각화는 챠트, FDP(force-directed displacement), 전략다이어그램(stragegic diagram), 패스파인더 네트워크(PFNet)를 말한다.

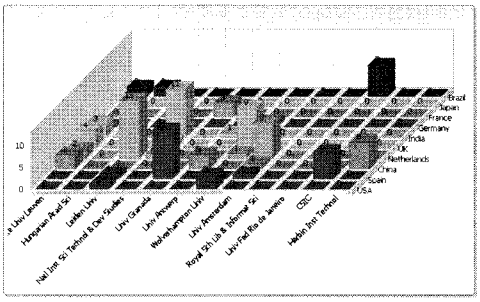


그림 5. 발생행렬 챠트 사례

본 시스템에서 챠트 기능은 논문 및 특허의 데이터베이스에서 추출한 여러 개의 필드정보 중에서 하나 또는 두 개의 필드에 대한 챠트를 보여줄 수 있도록 구현하였다.

본 시스템에서 FDP는 Eades[7] 방식과 Kamada와 Kawai[8] 방식, Fruchterman과 Reingold[9]의 방식을 비교한 후 Fruchterman과 Reingol이 제안한 방식으로 구현하였는데, Eades 방식에 비해서는 정확도가 높고, Kamada & Kawai 방식에 비해서는 구현이 단순하다. 이 방식은 모든 정점간에 인력과 척력을 계산하고 각 정점들의 이동 변위에 한계값을 주고 이동시킨다. 이러한 과정을 계속해서 반복하다보면 마지막에는 정점들의 좌표배치가 최적화되는 모습을 확인할 수 있다.

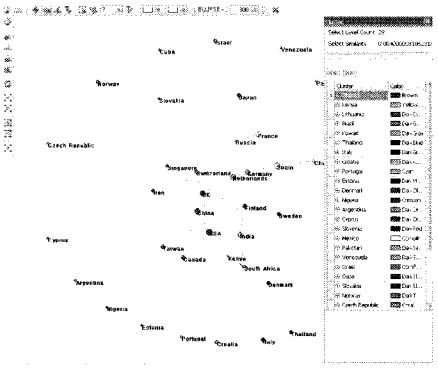


그림 6. FDP 구현 사례

패스파인더 네트워크[10]는 복잡하게 연결된 네트워크를 최대한 단순화시키는 목적으로 사용된다. 패스파인더 네트워크는 파라미터 q와 r을 고려하는데 q는 노드 사이의 경로거리를 계산할 때 고려하는 최대 링크의 수를 말하며 r은 민코프스키 거리공식에서 제곱수를 나타낸다. r 값이 무한대이면 경로를 구성하는 링크의 가중치 중 최대값이 경로의 거리가 된다[11]. 본 시스템에서는 r=∞를 디폴트로 설정하고 q 값을 사용자가 입력할 수 있도록 구현하였다.

전략다이어그램은 Callon[12]이 제시한 동시단어분석(co-word) 방법으로 특정 기술군을 선정하고 그 내부에서 몇 개의 세부 기술군간의 상대적인 발전도(활성화 정도)를 살펴보고자 할 때 적합하다. 본 시스템에서는 한국과학기술정보연구원에서 자체 연구개발한 방식으로 구현하였다[13].

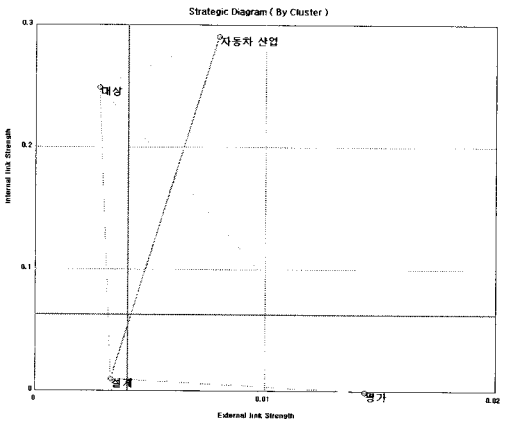


그림 7. Strategic Diagram 구현 사례

2.4 데이터 전처리

데이터 전처리(pre-processing)는 서지DB의 비일관성 문제를 해결하고 다양한 분석을 하기위한 필수적인 과정이다. 본 시스템에서 데이터 전처리는 필드정제, 그룹핑, 부분데이터집합 생성, 편집기, 자연어 처리 등으로 구성된다. 필드정제와 그룹핑 및 부분데이터집합 생성은 원하는 항목만을 선택해서 분석용 데이터만을 추출하기 위해서 구현하였다.

편집기에는 입력 편집기, 문자열 편집기, 시소러스 편집기를 구현하였다. 본 시스템에서 입력편집기(Import Editor)는 임의의 텍스트 데이터베이스 파일을 입력받을 수 있도록 설정하는 편집기이다.

문자열 편집기(String Editor)는 정확히 일치하지 않는 단어나 어절(phrase)을 그룹핑하는 규칙을 설정하는 편집기이다. 본 시스템에서는 여러 가지 기법을 제공하고 있는데 최장 공통 부분열(Longest Common Subsequence Ratio: LCSR)기법과 bi-gram 기법[15], 스테밍(stemming) 기법[14], 불용어(stopwords) 제거 기능 등을 구현하였다.

시소러스 편집기는 기존에 잘 구축된 방대한 시소러스 사전을 이용하는 것은 아니다. 즉 그룹핑 기능 등의 다른 과정을 통해서 도출된 시소러스나 미리 간단히 정의된 시소러스(국가 시소러스 등)를 이용하여 나중에 필드 항목을 정제할 때 사용하는 용도이다.

자연어 처리는 논문이나 특허의 제목이나 초록에서 단어나 어절을 추출할 필요가 있을 때 많이 사용되는 기능이다. 본 시스템에서 영문은 명사구만을 대상으로 하여 추출하였고, 한글은 형태소 분석기술을 이용하여 구현하였다.

Ⅳ. 시스템 구동 환경

본 시스템은 .NET 프레임워크를 이용하여 개발하였다. 본 시스템을 처음 설치할 때는 실행에 필요한 .NET 프레임워크를 인터넷에서 자동으로 다운로드 받는 과정이 필요하기 때문에 인터넷에 연결되어 있어야 한다. 시스템 구동 최소 환경은 다음과 같다.

- 메모리 1기가 바이트 이상
- 하드디스크 200M 이상
- OS 환경 Windows 2000/NT/XP/Vista(.Net2.0) 이상
- CPU Pentium 4 이상

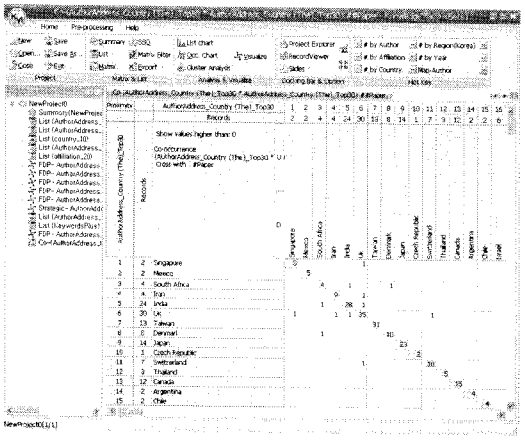


그림 8. Home 전체화면

V. 결론

본 시스템은 해외의 유명한 소프트웨어를 벤치마킹 하였고 기능면에서 상세하게 비교분석한 결과 전반적으로 좀 더 다양한 기능을 제공하는 것으로 파악되었다[16]. 장점에 대해서만 전반적으로 살펴보면 VantagePoint가 제공하는 행렬값의 형태보다 더욱 다양한 행렬 형태를 제공하고 있고 BibTechMon이 제공하는 시각화 기능과 유사한 정도로 시각화 기능을 제공하고 있다. 수요자의 활용상 편리를 위해 단축키도 제공하고 있는데 연도별/국가별/기관별/저자별/한국 지역별 건수 챠트와 국가별/기관별/저자별/한국 지역별 협력맵을 원버튼 클릭으로 결과를 확인할 수 있도록 개발하였다. 본 시스템에서 구현한 분석방법론이 군집분석 방법이 주류를 이루고 있어서 다방면에 활용하기에는 부족한 점이 있지만 해외의 유명한 소프트웨어보다 더욱 다양한 기능을 수행하면서 한글지원이 된다는 점에서 개발의 의미가 크다고 판단된다. 끝으로 KnowledgeMatrix에 관한 정보는 miso.yeskisti.net 사이트에서 확인할 수 있으며 일반사용자를 위한 다운로드도 제공하고 있다.

참 고 문 헌

[1] 문영호 외, *지식정보의 조사분석체계 구축 연구*, KISTI, 2004.

[2] 문영호 외, *차세대 R&D 정보분석·평가·예측 시스템 개발*, KISTI, 2007.

[3] S. A. Morris, *Unified mathmatical treatment of complex cascaded bipartite networks: the case of collections of journal papers*, Oklahoma State University, 2005.

[4] N. H, Timm, *Applied Multivariate Analysis*, Springer, 2002

[5] 임종원 외, *마케팅조사방법*, 법문사, 2001.

[6] W. Bae and S. W. Ron, "A study on k-means clustering", The Korean Communications in Statistics Vol.12, No.2, pp.497-508, 2005.

[7] P. Eades, "A heuristic for graph drawing", CONGRESSUS NUMERANTIUM, Vol.42, pp.149-160, 1984.

[8] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs", Information Processing Letters 31, pp.7-15, 1989.

[9] T.M.J. Fruchterman and E.M. Reingold, "Graph drawing by force-directed placement", Softwrar-practice and experience, Vol.21 No.11, pp.1129-1164, 1991.

[10] R.W. Schvaneveldt, *Pathfinder associative networks: studies in knowledge organization*, Ablex publishing corporation, 1990.

[11] 이재윤, "지적 구조의 규명을 위한 네트워크 형성 방식에 관한 연구", 한국문헌정보학회지, 제40권, 제2호, pp.333-355, 2006.

[12] M. Callon, J.P. Courtial, and F. Laville, "Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer

chemistry", Scientometrics, Vol.22, No.1, pp.155-205, 1991.

[13] B. Lee and Y. Jeong, "Mapping the Korea's national R&D domain of robot technology by using the co-word analysis", Scientometrics, Vol.77, No.1 (2008년 10월 출판 예정)

[14] M.F. Porter, "An algorithm for suffix stripping", Program, Vol.14 No.3, pp.130-137, 1980.

[15] B.Y. Ricardo, and R.N. Berthier, *Modern information retrieval*, ACM press, 1999.

[16] 이방래, 이준영, 여운동, 이창환, 문영호, 권오진, "서지데이터 분석 툴에 대한 특성 및 편의성 비교 분석", KOSTI 2007/한국콘텐츠학회 추계종합학술대회, pp.501-505, 2007.

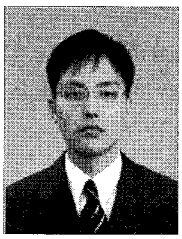
저 자 소 개

이 방 래(Bangrae Lee) 정회원

- 2002년 2월 : KAIST 전자전산학과 (공학석사)
- 2002년 4월 ~ 현재 : KISTI 선임연구원

<관심분야> : 데이터마이닝, 지식계량학, 사회네트워크분석

여 운 동(Woon-Dong Yeo) 정회원



- 2002년 2월 : 경북대학교 전자공학과 (공학석사)
- 2002년 4월 ~ 현재 : KISTI 선임연구원

<관심분야> : Datamining, Scientometrics

이 준 영(June-Young Lee) 정회원

- 2000년 2월 : 고려대학교 과학기술학 (이학사)
- 2001년 5월~현재 : KISTI 선임연구원

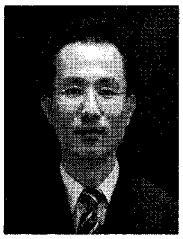
<관심분야> : 과학계량학, 지식네트워크 시뮬레이션

이 창 환(Chang-Hoan Lee) 정회원

- 1991년 8월 : 연세대학교 (공학석사)
- 2008년 2월 : 서울시립대학교 (공학박사)
- 1991년 6월 ~ 현재 : KISTI 책임연구원

<관심분야> : 계량정보분석, 데이터마이닝, 정보분석툴

권 오 진(Oh-Jin Kwon) 정회원



- 1994년 8월 : 광운대학교 전자계산학과 (이학석사)
- 2006년 2월 : 서울시립대학교 전산통계학과 박사수료
- 1991년 6월 ~ 현재 : KISTI 선임연구원

<관심분야> : 데이터마이닝, 고성능컴퓨팅, 정보분석시스템, 사회네트워크분석

문 영 호(Yeong-Ho Moon) 정회원

- 1996년 2월 : KAIST 건설환경공학과 (공학박사)
- 1986년 1월 ~ 2000년 12월 : 산업연구원, 산업기술정보원
- 2001년 1월 ~ 현재 : KISTI 정책연구실 실장/책임
- 2004년 1월 ~ 현재 : KISTI 정보분석센터 센터장

<관심분야> : 계량 정보분석, 미래 유망기술, 정보분석시스템, 네트워크 분석