국내 논문 저자의 소속 연구기관 분석을 위한 시스템*†

홍현욱*, 권상은*, 임채호*, 김병규**
*한국과학기술원 정보보호대학원
**한국과학기술정보연구원 국내정보팀
e-mail:{karmatia, kse, chlim}@kaist.ac.kr, yourovin@kisti.re.kr

System to Analyze Affiliation of Domestic Paper Author

Hyun-Wook Hong*, Sang-Eun Kwon*, Chae-Ho Lim*, Byung-kyo Kim**
*Dept. of Graduate School of Information Security, Korea Advanced Institute of Science and Technology
**Domestic Information Team, Korea Institute of Science and Technology Information

요 약

연구기관의 연구 활동에 대한 평가는 연구기관에 소속된 연구자들의 논문으로 연구 활동을 분석함으로써 이루어질 수 있다. 논문으로 연구 활동을 분석하여 연구기관에 대한 평가를 하기 위해서는 논문을 작성한 저자의 소속기관이 가장 먼저 식별되어야 한다. 본 논문에서는 국내 과학학술지 논문에 대한 저자의 소속 연구기관을 식별하는 시스템을 구축하여 각 연구기관의 연구 활동에 대하여 분석해 보았다. 이러한 시스템을 기반으로 연구기관별 논문 수, 피인용 횟수, 1회 이상 피인용 된 논문 수 등의 기본 연구 활동을 분석하여 얻은 결과는 연구 기관 평가의 근거로 사용할 있으며, 나아가 특정 분야에서 강세를 보이는 연구기관과 연구기관끼리의 협업 관계를 분석하여 정책에 활용할 수도 있다.

1. 서론

연구기관의 연구 활동에 대한 평가를 하기 위해서는 연구기관의 연구 결과를 평가하여야 한다. 연구자들은 연구에 대한 결과를 논문으로 정리하여 학술지에 게재하며 이 실적을 분석함으로써 연구기관의 연구 활동에 대한 평가를 할 수가 있다. 논문으로 연구 활동을 분석하여 연구기관에 대한 평가를 하기 위해서는 논문을 작성한 저자의 소속기관이 가장 먼저 식별되어야 한다. 논문에 대한 저자의 소속 연구기관이 식별되면, 연구기관별 논문 수, 피인용 횟수, 1회 이상 피인용 된 논문 수 등의 기본 연구 활동을 쉽게 알 수 있으며 나아가 특정 분야에서 강세를 보이는 연구기관과 연구기관끼리의 협업 관계도 분석해 낼 수 있다.

본 연구에서는 KSCI[1](Korea Science Citation Index)의 데이터베이스로 활용되는 KSCD(Korea Science Citation Database)[2]를 기반으로 국내 과학학술지 논문에 대한 저자의 소속 연구기관을 식별하는 시스템을 구축하였다. 그리고 구축한 시스템을 기반으로 하여 각 연구기관들의 연구 활동에 대해 분석해 보았다. 분석 대상 연구기관은 정부출연연구소 46개 및 전국 4년제 대학교 208개이며 KSCD의 2002년부터 2010년 동안 전체 1,189,529 건의 논문 및 저자에 관한 데이터를 특정 데이터 처리 기준을 두어 처리하였다.

* 이 연구는 대한민국 지식경제부 정보통신진흥기금으로 수행되었으며, 정보통신산업진흥원(NIPA)의 관리로 진행된 사이버보안 연구센터 지원사업(과제코드 NIPA-H0701-12-1001)임.
† 이 연구는 한국과학기술정보연구원의 KSCD를 사용하여 연구되었음.

2. 시스템 설계

2.1. 데이터베이스 구축

저자 소속기관 분석을 위해 구축된 데이터베이스의 스키마는 <표 1>과 같다. 소속 기관에 대한 정의는 기관 테이블(ART_AFFILIATION TABLE)에 되어있다. 이 소속기관 테이블은 정부출연연구소 46개 및 전국 4년제 대학교 208개(2012년 3월 기준)를 대상으로 기관 변경이력을 분석 후 이 내용을 포함하여 구성하였다. 정부출연연구소와 대학교 모두 연도에 따라 통·폐합 및 기관명 변경이 있기에 이를 반영하기 위해 변경된 기관과 변경된 날짜 그리고 현재 기관에 대한 정보를 담을 수 있는 컬럼을 설계하였으며, 대학교끼리 통·폐합이 발생할 경우 캠퍼스별로 구분되는 경우가 있어 캠퍼스 별로 대학교를 조사하여 데이터베이스를 구성할 수 있도록 하였다. 그리고 대학교의 연구 분야 중 의학 분야가 많은 비중을 차지하기에 의대(의학, 약학, 치의학, 한의학, 수의학, 전문대학원)를 포함하고 있는 대학교의 경우 따로 표시를 할 수 있도록 하였으며, 추후 지역별 협력 관계를 분석할 수 있도록 기관의 위치 정보를 저장할 수 있는 컬럼을 설계하였다.

논문과 해당 저자에 관한 정보를 담는 논문 및 저자 테이블(ART_AUTHOR TABLE)은 KSCD에서 필요한 항목을 추출하여 데이터를 재가공하여 구축하였는데, 기존의 데이터베이스에서 주어진 정보는 논문 식별 번호, 논문에 대한 전체 저자의 한글 이름 또는 영문 이름, 한글 소속기관명 또는 영문 소속기관명, 그리고 저자들의 이메일이었다. 이 정보를 이용하여 각 저자들을 기준으로 저자 순번을 구분하고 구분된 저자들의 내용을 데이터베이스로 구성하였다.

<표 1> 저자 소속기관 분석을 위해 구축한 데이터베이스 스키마

TABLE ART_AFFILIATION (소속기관 테이블)		
데이터 의미	컬럼명	데이터 타입
기관 식별 번호	AFF_SEQ	NUMBER
기관 한글 이름	KOR_NAME	VARCHAR2
대학교 캠퍼스 구분	CAMPUS	VARCHAR2
변경된 기관 식별 번호	CHG_SEQ	NUMBER
기관명 변경 날짜	CHG_DATE	VARCHAR2
기관 변경 사유	REASON	CHAR
기관 홈페이지	HOMEPAGE	VARCHAR2
기관 위치(도 단위)	LOC	VARCHAR2
기관 위치(시, 구 단위)	LOC_CITY	VARCHAR2
기관 영문 이름	ENG_NAME	VARCHAR2
기관 한자 이름	CHI_NAME	VARCHAR2
기관 한글명 약어	KOR_ABB_NAME	VARCHAR2
기관 영문명 약어	ENG_ABB_NAME	VARCHAR2
기관 구분 (정부출연연구소 또는 대학교)	DIV	CHAR
대학교 의학분야 포함 여부	MEDICAL	CHAR
현재의 기관 식별 번호	CUR_SEQ	NUMBER
TABLE ART_AUTHOR (논문 및 저자 테이블)		
데이터 의미	컬럼명	데이터 타입
논문 식별 번호	ART_SEQ	VARCHAR2
저자 순번 (주저자, 공저자 구분)	AUT_SEQ	NUMBER
저자 한글 이름	AUK	VARCHAR2
저자 영문 이름	AUE	VARCHAR2
저자 소속기관 식별 번호	AFF_SEQ	NUMBER
저자 한글 소속기관명	CSK	VARCHAR2
저자 영문 소속기관명	CSE	VARCHAR2
저자 이메일	AUT_EM	VARCHAR2

2.2. 저자 소속 연구기관 식별을 위한 데이터 처리 기준

KSCD의 2002년부터 2010년까지의 1,189,529 건의 논문 저자에 관한 데이터를 대상으로 <표 2>와 같은 기준으로 데이터를 처리하였다. 데이터 중 한 저자가 여러 곳의 소속에 포함되어 있는 경우, 제일 먼저 서술한 것을 가장 우선순위가 높다고 가정하여 가장 먼저 서술한 것을 소속기관으로 식별하였다. 하지만 가장 먼저 서술된 것이 식별 대상이 아닌 경우 그 다음 서술된 것을 기준으로 식별하였고 이 또한 식별 대상이 아닌 경우 그 다음 것을 식별하는 방식으로 데이터를 처리하였다. 그리고 한글 소속기관명과 영문 소속기관명이 의미하는 기관이 다를 경우 한글 소속기관을 우선순위에 두고 식별하며 앞의 방법과 비슷하게 한글 소속기관이 식별 대상에 포함되지 않는 경우 영문 소속기관으로 식별하였다. 또한 대학교의 경우 4년제 대학교를 식별대상으로 선정하였으나 대학교의 학과마다 2년부터 4년까지의 과정으로 모두 달라 이 경우 구분하기가 쉽지 않아서 이 모두를 4년제 대학교로 식별하였다. 마지막으로 의학 분야(의학, 약학, 치의학, 한의학, 수의학, 전문대학원)를 연구하고 있는 대학교의 경우 이에 소속된 병원이 존재하는 경우가 있는데 대학교 이름이 포함된 병원의 경우 해당 대학교로 식별하여 데이터를 처리하였다.

<표 2> 저저 소속기관 식별을 위한 데이터 처리 기준

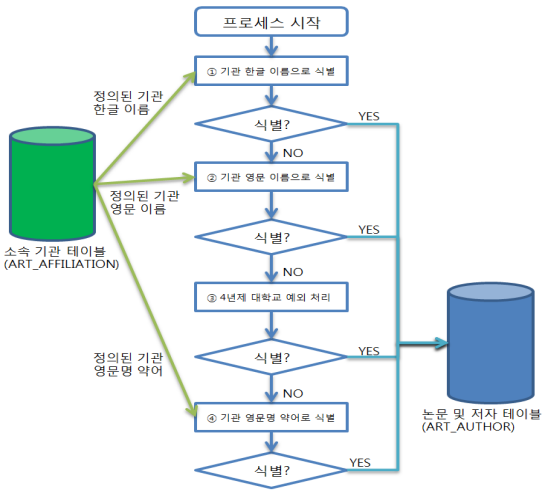
- 한 저자의 소속이 여러 개인 경우 제일 먼저 서술한 것을 기준으로 정리, 먼저 서술한 것이 식별 대상이 아닌 경우 그 다음 것을 기준으로 함
  - (연구원 A + 연구원 B) → 연구원 A 소속
  - (연구원 A + 대학교 A) → 연구원 A 소속
  - (민간기업 A + 대학교 A) → 대학교 A 소속
- 저자 한글 소속기관명과 영문 소속기관명이 의미하는 기관이 다를 경우 한글 소속을 우선순위에 두고 식별, 한글 소속이 식별 대상이 아닌 경우 영문 소속으로 정리
  - 한글 소속(대학교A), 영문소속(연구원A) → 대학교 A 소속
  - 한글 소속(민간기업A), 영문소속(대학교A) → 대학교 A소속
- 2년제 및 4년제가 혼재된 학교는 4년제 대학교로 식별함
- 의학, 약학, 치의학, 한의학, 수의학, 전문대학원에 소속된 병원은 해당 대학교로 식별하여 포함

2.3. 저자 소속 연구기관 식별 프로세스

각 저자들의 소속기관을 (그림 1)과 같은 프로세스를 거쳐 식별하였다. 각 프로세스에서 앞서 식별된 기관은 다음 프로세스에서 새롭게 식별되지 않도록 식별 대상에서 제외시키도록 하였다.

① 소속기관 테이블의 기관 한글 이름으로 식별

미리 구축하였던 소속기관 테이블의 기관 한글 이름 데이터를 이용하여 논문 및 저자 테이블에서 저자 한글 소속기관명에 소속기관 테이블의 기관 한글 이름이 포함되어 있으면 해당 기관으로 식별한다. 가령, 소속기관 테이블의 기관 한글 이름이 “한국과학기술원”이라는 데이터가 있고 논문 및 저자 테이블에서 저자 한글 소속기관명이 “한국과학기술원 정보보호대학원”이라면 이 경우 논문 및 저자 테이블의 저자 소속기관 식별 번호는 소속기관 테이블의 한국과학기술원 기관 식별 번호를 가지게 된다.



(그림 1) 저자 소속 연구기관 식별 프로세스

이러한 식별 과정은 PL/SQL[3], JDBC[4] 등으로 자동화할 수 있다.

② 소속기관 테이블의 기관 영문 이름으로 식별

소속기관 테이블에 기관 영문 이름 데이터를 이용하여 논문 및 저자 테이블에서 저자 영문 소속기관명에 소속기관 테이블의 기관 영문 이름이 포함되어 있으면 해당 기관으로 식별한다. 가령, 소속기관 테이블의 기관 영문 이름이 "Korea Advanced Institute of Science and Technology"이라는 데이터가 있고, 논문 및 저자 테이블에서 저자 영문 소속기관명이 "Dept. Graduate School of Information Science, Korea Advanced Institute of Science and Technology"라면 이 경우 논문 및 저자 테이블의 저자 소속기관 식별 번호는 소속기관 테이블의 한국과학기술원 기관 식별 번호를 가지게 된다. 이 프로세스는 대소문자 구분에 의해 같은 기관이 해당 기관으로 식별되지 않는 일이 발생하지 않도록 해당 영문명은 모두 소문자로 바꿔서 프로세스를 진행하였으며 이러한 식별 과정은 PL/SQL, JDBC 등으로 자동화할 수 있다.

③ 4년제 대학교 예외 처리 과정

4년제 대학교의 경우, 대학교 이름 전체를 제대로 서술하지 않아 생겨나는 예외가 많다. 의대를 포함하는 대학교의 경우 **의대, **의과대학, **의과대학교 등으로 서술되는 경우가 있어 이러한 경우 <표 3>과 같이 정규식을 활용하여 식별하였다. 이러한 정규식 또한 완벽하게 모든 경우를 식별할 수 없으며 예외가 존재하기에 정규식으로 생성된 결과가 올바른지 점검을 하여야 한다.

④ 소속 기관 테이블의 기관 영문명 약어로 식별

논문 및 저자 테이블의 저자 한글 소속기관명 또는 저자 영문 소속기관명에 단순히 기관 영문명을 약어로 서술해두거나 포함해둔 경우가 많다. 이런 경우는 소속기관 테이블의 기관 영문명 약어 데이터를 이용하여 저자 소속기관을 식별할 수 있다. 가령, 논문 및 저자 테이블의 저자 영문 소속기관명에 "KISTI"라는 데이터는 소속기관 테이블의 한국과학기술정보연구원의 기관 영문명인 "KISTI" 데이터와 일치하면 이 경우 논문 및 저자 테이블의 저자 소속기관 식별 번호는 소속기관 테이블의 한국과학기술정보연구원 기관 식별 번호를 가지게 된다. 하지만 4년제 대학교의 경우 중복되는 약어 이름을 가지는 경우가 많아 이 프로세스를 적용하기 어려우며, 정부출연연구소의 경우도 이 프로세스 적용 후 올바르게 기관이 식별되었는지 점검을 하여야 한다.

<표 3> 예외 처리를 위한 정규식

<대학교 이름>( )?(의)?(공)?(과)>대(학)?( )(교)?
예) 충남( )?(의)?(공)?(과)>대(학)?( )(교)?

3. 저자 소속 연구기관 분석 결과

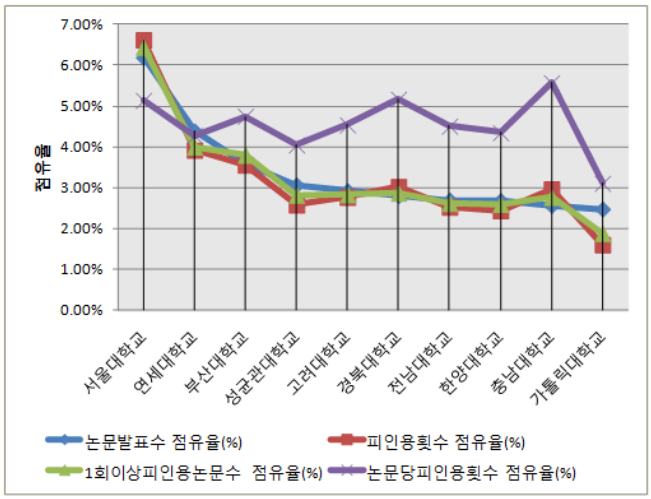
2.1.에서 구축된 데이터베이스를 이용하여 2.2.에서 제시한 데이터 처리기준으로 2.3.의 프로세스를 거쳐 저자 소속 연구기관을 식별하였다. 이렇게 식별한 결과를 바탕으로 분석한 연구기관의 연구 활동을 분석 비교하였는데 분석 결과는 상위 10개 기관에 대해서만 서술하였다.

3.1. 전국 4년제 대학교 연구 활동

<표 4>에서 전국 4년제 대학교의 연구 활동 내역에 관한 정보를 확인할 수 있다. 논문 발표 수를 기준으로 1위부터 10위까지의 대학교에 대해 피인용 횟수, 1회 이상 피인용 된 논문 수, 논문 당 피인용 횟수를 분석하였다. 또한 (그림 2)에서 분석한 상위 10개의 대학교에 대한 연구 활동 내역을 쉽게 비교할 수 있다. 논문 발표 수, 피인용 횟수, 1회 이상 피인용 된 논문 수는 서울대학교가 우수하며 논문 당 피인용 횟수는 선정된 상위 10개 대학 중 충남대학교가 우수한 것을 알 수 있다.

<표 4> 전국 4년제 대학교 연구 활동 내역(상위 10위)

대학교명	논문발표수		피인용횟수		1회이상피인용 논문수		논문당피인용 횟수	
	순위	논문수 점유율(%)	순위	횟수 점유율(%)	순위	논문수 점유율(%)	순위	횟수 점유율(%)
서울대학교	1	13,290 (6.19%)	1	8,879 (6.62%)	1	4,939 (6.44%)	7	0.67 (5.14%)
연세대학교	2	9,442 (4.39%)	2	5,258 (3.92%)	2	3,065 (4.00%)	17	0.56 (4.28%)
부산대학교	3	7,744 (3.60%)	3	4,779 (3.56%)	3	2,913 (3.80%)	10	0.62 (4.75%)
성균관대학교	4	6,590 (3.07%)	7	3,468 (2.58%)	6	2,162 (2.82%)	18	0.53 (4.05%)
고려대학교	5	6,290 (2.93%)	6	3,716 (2.77%)	5	2,170 (2.83%)	13	0.59 (4.54%)
경북대학교	6	6,039 (2.81%)	4	4,061 (3.03%)	4	2,203 (2.87%)	6	0.67 (5.17%)
전남대학교	7	5,792 (2.70%)	8	3,395 (2.53%)	8	2,011 (2.62%)	14	0.59 (4.51%)
한양대학교	8	5,781 (2.69%)	10	3,270 (2.44%)	9	1,993 (2.60%)	16	0.57 (4.35%)
충남대학교	9	5,516 (2.57%)	5	3,992 (2.97%)	7	2,120 (2.77%)	4	0.72 (5.57%)
가톨릭대학교	10	5,330 (2.48%)	20	2,156 (1.61%)	16	1,449 (1.89%)	20	0.4 (3.11%)



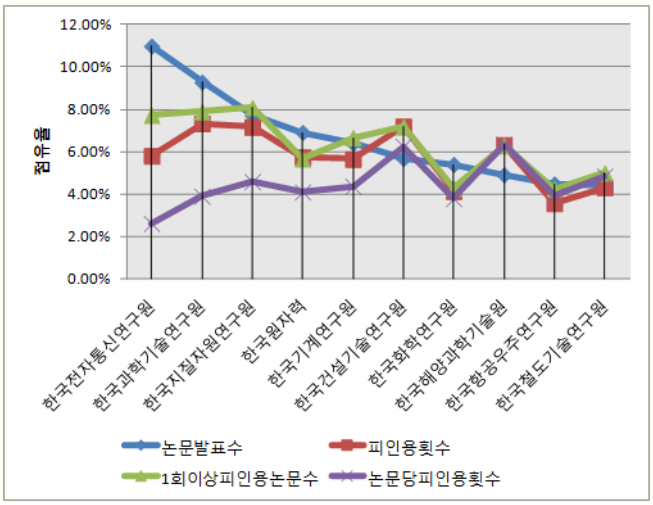
(그림 2) 전국 4년제 대학교 연구 활동 그래프(상위 10위)

3.2. 정부출연연구기관 연구 활동

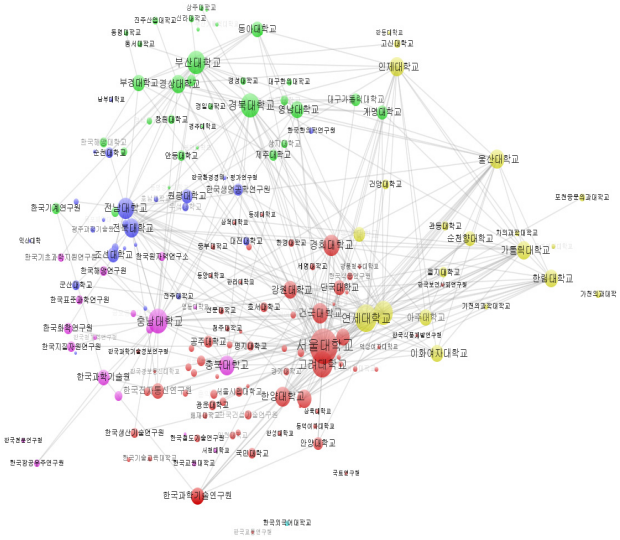
<표 5>에서 정부출연연구기관의 연구 활동 내역에 관한 정보를 확인할 수 있다. 논문 발표 수를 기준으로 1위부터 10위까지의 연구기관에 대해 피인용 횟수, 1회 이상 피인용 된 논문 수, 논문 당 피인용 횟수를 분석하였다. 또한 (그림 3)에서 분석한 상위 10개 정부출연연구기관에 대한 연구 활동 내역을 쉽게 비교할 수 있다. 논문 발표 수, 피인용 횟수, 1회 이상 피인용 된 논문 수는 한국과학기술연구원이 고르게 우수하며 한국해양과학기술원은 논문 발표 수를 제외한 나머지 지표에서 모두 우수한 것을 알 수 있다.

<표 5> 정부출연연구기관 연구 활동 내역(상위 10위)

연구기관명	논문 발표수		피인용 횟수		1회 이상 피인용논문수		논문 당 피인용 횟수	
	순위	논문수 (점유율%)	순위	횟수 (점유율%)	순위	논문수 (점유율%)	순위	횟수 (점유율%)
한국전자통신연구원	1	1,769 (10.98%)	6	655 (5.80%)	3	484 (7.73%)	19	0.37 (2.61%)
한국과학기술연구원	2	1,495 (9.28%)	2	829 (7.33%)	2	494 (7.89%)	12	0.55 (3.90%)
한국지질자원연구원	3	1,241 (7.71%)	4	809 (7.16%)	1	506 (8.08%)	8	0.65 (4.59%)
한국원자력연구원	4	1,113 (6.91%)	7	649 (5.74%)	8	354 (5.66%)	10	0.58 (4.10%)
한국기계연구원	5	1,030 (6.40%)	8	638 (5.65%)	6	417 (6.66%)	9	0.62 (4.36%)
한국건설기술연구원	6	915 (5.68%)	3	813 (7.19%)	5	450 (7.19%)	5	0.89 (6.25%)
한국화학연구원	7	869 (5.40%)	12	468 (4.14%)	11	267 (4.27%)	13	0.54 (3.79%)
한국해양과학기술원	8	790 (4.91%)	5	713 (6.31%)	7	393 (6.28%)	4	0.90 (6.35%)
한국항공우주연구원	9	721 (4.48%)	13	402 (3.56%)	12	266 (4.25%)	11	0.56 (3.92%)
한국철도기술연구원	10	706 (4.38%)	11	487 (4.31%)	10	313 (5.00%)	7	0.69 (4.85%)



(그림 3) 정부출연연구기관 연구 활동 그래프(상위 10위)



(그림 4) 연구 기관 협업 관계도

3.3. 연구기관 협업 관계도

연구기관의 협업 관계도는 한 논문에서 서로 다른 기관 소속의 공저 관계를 분석하여 데이터를 추출하였다. (그림 4)는 추출된 데이터를 기반으로 VOSviwer[5] 프로그램을 활용하여 협업 관계도를 그린 것이다. 같은 색으로 표시된 연구 기관들은 같은 그룹의 연구 기관으로 대개 같은 지역 위치하고 있어 협업 관계가 많았음을 알 수 있다.

4. 결론

본 논문에서는 국내 과학학술지 논문에 대한 저자의 소속 연구기관을 식별하는 시스템을 구축하여 각 연구기관의 연구 활동에 대하여 분석해 보았다. 이러한 시스템을 기반으로 연구기관별 논문 수, 피인용 횟수, 1회 이상 피인용 된 논문 수, 논문 당 피인용 횟수 등의 기본 연구 활동을 분석하여 연구 기관 평가의 근거로 사용할 수 있을 뿐 아니라, 향후 분야별로 강세를 보이는 연구기관 분석하며 연구기관끼리의 협업 관계, 지역별 협업 관계 등을 분석하여 결과를 얻는다면 연구를 위한 정책을 설정하는 기반 자료로 활용할 수 있을 것으로 기대된다.

참고문헌

[1] KSCI(Korea Science Citation Index), [cited 2011. 5] http://ksci.kisti.re.kr/

[2] KSCD(Korea Science Citation Database), [cited 2011.5] http://ksci.kisti.re.kr/

[3] PL/SQL(Procedural Language extension of SQL), http://plsql-tutorial.com/index.htm

[4] JDBC(The Java Database Connectivity), http://www.oracle.com/technetwork/java/javase/jdbc/index.html

[5] VOSviwer, http://www.vosviewer.com/