거대 희소 행렬을 이용한 특허정보 유통 모형에 대한 연구

권오진° 서진이 김정호 노경란 김완종 *김진석
한국과학기술정보연구원, *서울시립대
{dbajin°, jinny, aresto, inforan, whkim}@kisti.re.kr, jskim*@venus.uos.ac.kr

A Study on Patent Information Dissemination Model using Large Scale Sparse Martix

OhJin Kwon°, Jinny Seo, J.H. Kim, K.R. Noh, W.J. Kim, J.S. Kim*
KISTI, *University of Seoul

요 약

최근 특정 주제의 지적 구조를 파악하기 위한 저자 동시인용분석, 동시단어분석, 서지결합법 등 계량정보분석에 대한 연구가 활발히 진행되고 있다. 그러나 국내의 경우 계량정보분석 기법을 활용한 정보 유통 프레임웍을 갖추고 있는 연구기관이나 대학이 아직 없는 실정이다. 그 이유는 특허나 과학문헌에 대한 인용정보를 보유한 곳이 없고, 거대 인용정보 행렬을 계산하기 위한 컴퓨팅 자원을 확보하지 못하고 있기 때문이다. 본 연구는 미국 특허 데이터베이스를 대상으로 인용 피인용 행렬을 생성한 후, 클러스터 컴퓨터를 사용하여 동시인용과 서지결합빈도를 계산하고 그 결과를 이용자에게 제공하는 정보 유통 서비스 모델을 제시하고자 한다.

키워드 : 서지결합법, 동시인용분석, 계량정보분석, 정보유통모형, 특허정보, 거대희소행렬

1. 서론

최근 특정 주제의 지적 구조를 파악하기 위해 저자 동시인용분석, 동시단어분석, 서지결합법과 같은 계량정보분석기법에 대한 연구가 활발히 진행되고 있다. 학술논문에 인용된 참고문헌을 대상으로 이루어지던 계량정보분석기법은 웹상에 존재하는 웹페이지로, 그리고 특허라는 다른 정보매체로 그 분석대상을 확대하고 있다. 그러나 국내에서는 아직 이러한 계량정보분석 기법을 활용한 정보 유통 프레임웍을 갖추고 있는 연구기관이나 대학이 없는 실정이다. 그 이유는 특허나 과학문헌에 대한 인용정보를 보유한 곳이 없고, 거대 인용정보 행렬을 계산하기 위한 컴퓨팅 자원을 확보하지 못하고 있기 때문이다. 따라서 이 연구는 미국 특허 데이터베이스를 대상으로 인용·피인용 행렬을 생성한 후 인용정보 행렬을 이용하여 클러스터 컴퓨터를 사용하여 동시인용빈도와 서지결합도를 계산한다. 그리고 계산된 결과를 이용자에게 서비스하는 정보유통 서비스 모델을 제시하고자 한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 동향에 대해 살펴보고, 3장에서는 서지결합과 동시인용 네트웍 분석에 대한 개념을 정의하고 대용량 정보 처리를 위한 희소 행렬 처리방법에 대해서 언급한다. 4장에서는 이를 활용한 유통 모델을 제시하고 5장은 결론으로 구성한다.

2. 관련 동향

Thomson ISI의 웹 데이터베이스인 Web of Science는 어떤 문헌을 검색했을 때 'Cited References'기능은 그 문헌이 인용하고 있는 문헌들을 보여주고, 'Times Cited'기능은 이 문헌을 인용하고 있는 문헌들을 보여준다. 그리고 'Related Records'기능은 서지결합기법을 이용하여 이 문헌과 공통되는 인용문헌을 하나 이상 가지고 있는 문헌들을 보여줌으로써 기존의 주제명 또는 저자명 검색으로 찾을 수 없었던 관련 연구분야를 효과적으로 검색할 수 있도록 한다.

한편 Citeseer의 'Related documents'기능은 2개의 인용문헌이 이후에 출판된 제3의 문헌에 동시에 인용된 상태를 보여주며, 두 문헌을 동시에 인용한 문헌의 수인 동시인용빈도를 보여준다. 즉, Citeseer의 대표적인 기능인 ACI (Autonomous Citation Indexing) 는 한 논문에서 인용된 논문들의 리스트를 원문에서 자동으로 인덱싱하여서 인용된 논문과 연결해 준다. 또한 인용된 논문들을 계속해서 새롭게 업데이트하여 링크하기 때문에, 이 논문과 관련된 최근의 연구 결과들을 손쉽게 찾아 볼 수

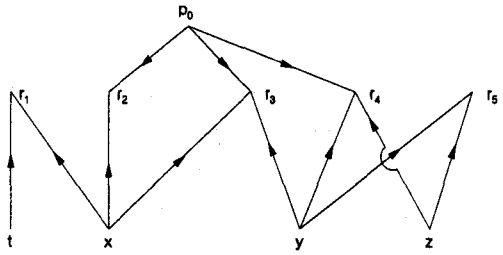
있도록 한다.

구글은 웹상의 방대한 정보를 효과적으로 끌어내는데 서지결합기법에 기초하여 역방향 인용링크를 이용한 페이지랭크라는 독특한 방식을 적용하였다.

3. 서지결합과 동시인용 네트웍 분석

3.1 서지결합과 동시인용 계산

서지결합법(BC: Bibliographic Coupling)은 <그림 1>과 같이 여러 개의 문헌이 공통으로 인용하는 문헌의 개수에 따라 주제의 공통성(affinity)을 측정하는 방식으로 두 가지 기준으로 정의된다.



<그림 1> 인용·피인용 관계 그래프

문헌 P_0와 적어도 하나 이상의 결합단위가 있다면 $Ga(P_0)$ 관련 집단을 갖고, 한 문헌 집단이 다른 문헌과 적어도 하나 이상의 결합단위를 가질때, 이들 문헌은 G_B라는 또 다른 관련 집단을 이룬다. <그림 2>에서 $Ga(P_0:2)=\{p0,x,y\}$는 두 개의 결합단위로 결합된 BC 집합으로서 $Ga(P_0)=\{P_0,x,y,z\}$의 부분 집합이다[1].

동시인용(CC : Co-Citation)은 먼저 발표된 두 편의 논문이 나중에 발표된 논문에 동시에 인용되는 것을 말한다. 동시인용빈도는 두 개의 논문이 동시에 인용되는 회수로 정의된다. 동시인용빈도가 높은 논문들은 서로 밀접한 관련이 있으며, 또한 인용빈도가 높은 문헌들이 어떤 핵심적인 개념이나 방법 등을 담고 있기 때문에 동시 인용패턴을 이용하여 핵심개념이나 방법론을 유추해 내는 것이 가능하다[2].

<그림 2>는 인용정보를 행렬로 표현한 것으로서 집합 $T=\{P_0,t,x,y,z\}$가 되고 집합 T가 가지고 있는 인용정보는 $B=\{r_1,r_2,r_3,r_4,r_5\}$가 된다. 이를 행렬로 표현하면 $[A]_{m\times n}$으로 정의 할 수 있다.

$A_{5\times 5}$ =

	r_1	r_2	r_3	r_4	r_5
P_0	0	1	1	1	0
t	1	0	0	0	0
x	1	1	1	0	0
y	0	0	1	1	1
z	0	0	0	1	1

<그림 2> 인용 정보의 행렬표현

BC는 다음과 같은 식으로 표현될 수 있다.

$BC_{m\times m} = A_{m\times n}\times \tilde{A}_{n\times m}$
where $\tilde{A}_{n\times m}$ is $A_{m\times n}$'s transpose matrix (1)

$BC_{5\times 5}: A_{5\times 5}\times \tilde{A}_{5\times 5}=$

	P_0	t	x	y	z
P_0	3	0	2	2	1
t	0	1	1	0	0
x	2	1	3	1	0
y	2	0	1	3	2
z	1	0	0	2	2

$BC_{m\times m}$ 행렬은 특허 T에 대한 대칭적인 상관관계 행렬이다. 따라서 행렬 곱셈으로 $A_{5\times 5}\times \tilde{A}_{5\times 5}$ 특허 T에 대한 서지결합도를 나타내는 BC 행렬을 구할 수 있다. 대각 행렬의 원소는 T집합에 대한 out-degree 수를 의미한다. 즉 $\{P_0,P_0\}$는 3개의 out-degree를 갖고 있으며 $\{P_0,t\}$는 서로 공통되는 out-degree가 없고 P_0과 x는 서로 두개의 reference를 공유하는 것을 알 수 있다. 즉 P_0는 t보다 x와 주제적으로 밀접하다고 할 수 있다. BC를 이용한 활용 분야는 검색 효율(정확률)을 향상시키기 위한 수단으로 사용가능하며 또한 문헌의 군집화에도 사용될 수 있다.

$CC_{n\times n}$는 다음과 같은 식으로 표현될 수 있다.

$CC_{n\times n}= \tilde{A}_{n\times m}\times A_{m\times n}$ (2)

$CC_{5\times 5}: \tilde{A}_{5\times 5}\times A_{5\times 5} =$

	r_1	r_2	r_3	r_4	r_5
r_1	2	1	1	0	0
r_2	1	2	2	1	0
r_3	1	2	3	2	1
r_4	0	1	2	3	2
r_5	0	0	1	2	2

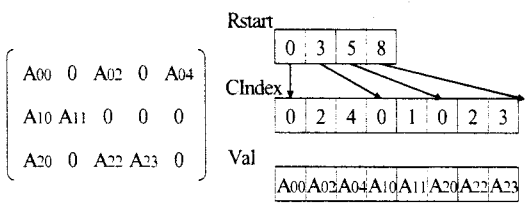
$CC_{n\times n}$ 행렬은 인용정보 B에 대한 대칭적인 상관관계 행렬이다. 따라서 행렬 곱셈으로 $A_{5\times5}\times A_{5\times5}$ 특허 B에 대한 동시인용도를 살펴볼 수 있는 $CC_{5\times5}$ 행렬을 구할 수 있다. 대각행렬은 B집합의 원소에 대한 in-degree 수를 의미하고 즉 $\{r_1,r_1\}$는 2개의 in-degree를 갖고 있으며 $\{r_1,r_2\}$ 서로 공통되는 in-degree값이 1이고 $\{r_1,r_4\}$는 공통되는 in-degree가 0이다. 즉 r_1은 r_4보다 r_2 와 주제적으로 밀접한 것을 알 수 있다.

3.2 희소 행렬의 표현

특허 정보의 동시인용빈도와 서지결합빈도를 계산하기 위해 행렬을 계산하였다. 행렬 연산은 계산 과학을 사용하는 분야에서 다양하게 사용되고 있으며 이때 사용되는 행렬의 크기는 풀고자 하는 문제의 변수의 개수에 따라 결정되므로 실제적인 문제에서는 크기가 너무 커서 모든 원소를 2차원 배열의 형태로 표현하였을 때 대부분 컴퓨터의 주 메모리 크기를 초과한다. 그러나 이 행렬을 잘 살펴보면 행렬의 원소는 연관이 있는 변수의 행과 열이 만나는 곳에서만 0이 아닌 값을 가지고 대부분의 경우에 0의 값을 가진다. 이러한 행렬은 0이 아닌 값만을 그 위치 정보와 함께 저장하는 희소 행렬의 형태로 저장된다[3]. 희소 행렬 형태로 저장하면 메모리의 사용량을 줄일 수 있을 뿐아니라 행렬 연산을 하는데 0값에 대한 연산을 생략하기 때문에 계산량을 줄일 수 있다.

희소행렬을 저장하는 자료 구조는 행렬의 특성, 응용분야의 특성에 따라 다양하다. 일반적으로 CSR (Compressed Sparse Row) 혹은 CSC(Compresed Sparse Column) 형태 중 선택하여 사용한다[4].

특허정보에 대한 네트웍을 규명하기 위해 3.1절에서 정의된 행렬을 기반으로 한국과학기술정보연구원에서 보유하고 있는 미국특허(1976-2005) 인용정보를 대상으로 $A_{2,536,707\times4,373,613}$ 행렬을 생성하였다. 이 행렬의 모든 원소를 2차원 형태의 일집 행렬로 처리할 경우 메모리 크기가 커지며 시간 복잡도 또한 n^3 이 되어 이를 계산하는데 많은 자원을 필요로 한다. $A_{2,536,707\times4,373,613}$ 행렬은 연관이 있는 변수의 행과 열이 만나는 곳에서만 1의 값을 갖고 대부분의 경우 0의 값을 갖는 희소 행렬이다. 시간과 메모리 절약을 위해 BLAS 포럼[4]에서 제안된 형식 중 CSR형식으로 변환하였다. CSR 형식은 3개의 배열로 표현된다. 희소 행렬의 위치 정보 중 각 행의 시작 위치를 가지고 있는 Rstart배열, 열의 위치 정보를 가지고 있는 CIndex배열, 0이 아닌 값을 가지고 있는 Val 배열이다. 희소행렬의 CSR 형식은 <그림 3>과 같다.



<그림 3> 희소 행렬의 CSR 표현

3.3 희소 행렬 계산을 통한 BC & CC 네트웍

$A_{2,536,707\times4,373,613}$ 행렬 중 1의 값을 갖는 개수는 24,870,516건이며 밀도(density)는 2.2×10^{-4}% 이다. BC와 CC 행렬의 계산을 위해 KISTI 슈퍼컴퓨터센터의 클러스터 컴퓨터를 활용하여 BC와 CC를 계산 하였다.

BC는 원행렬과 원행렬의 전치 행렬을 곱하여 계산되나, 이는 원행렬에 대한 각각의 행에 대해 동일한 열에 대한 1의 값을 갖는 수의 합과 동일한 결과를 갖는다. 즉 BC(1,2)의 값은 원행렬 1행과 2행에 대해 1과 2 행 중 동일한 열에 있는 1의 값의 수를 계산한다. 따라서 BC와 CC를 계산하기 위해 먼저 원행렬에 대한 전치 행렬을 구하고, 각 원행렬과 전치 행렬에 대한 correlation을 수행하였다. 1개의 CPU 활용시 약 215시간이 필요했지만 128개의 CPU를 활용할때 1.2시간이 소요되었다.

<표 1>은 클러스터 컴퓨터를 활용한 희소 행렬 $A_{2,536,707\times4,373,613}$의 BC & CC 연산을 수행한 시간이다.

<표 1> 클러스터 컴퓨터를 활용한 희소 행렬 $A_{2,536,707\times4,373,613}$의 BC & CC 수행시간

<성능 시험표> (단위:hour)

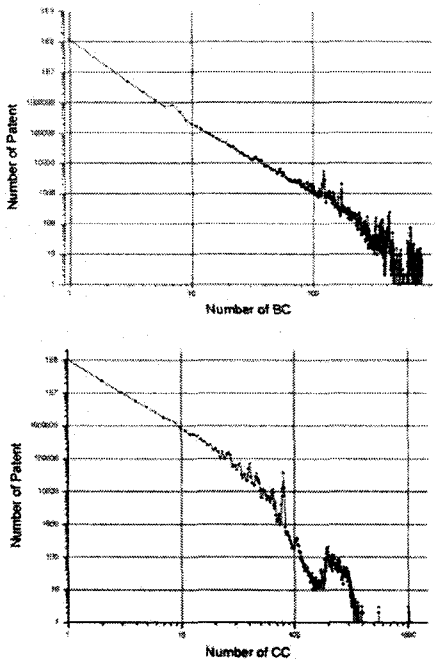
no of CPU	A×A^T	A^T ×A	Total	Speedup
1	80.01	134.92	214.93	1.0084
2	32.27	56.26	88.52	2.43
4	15.17	25.18	40.35	5.33
8	7.30	12.41	19.71	10.90
16	3.64	6.11	9.74	22.06
32	1.81	3.04	4.86	44.27
64	0.91	1.51	2.42	88.84
128	0.46	0.76	1.22	176.22

※ Speedup(성능향상률):n개의 프로세서를 이용했을 때의 순차성능에 대한 성능향상비

$$Sn = \frac{\{\text{Elapse time using 1 CPU}\}}{\{\text{Elapse time using n CPU}\}}$$

※ 사용 시스템: KISTI Hamel Cluster
※ 시스템 사양:
CPU - Intel Pentium Xeon 2.8GHZ (2CPUs/node)
Meomory - 3GB/node, Interconnection - Myrinet 2000

계산결과 미국특허의 BC와 CC에 대한 그래프는 <그림 4>와 같다. 서지결합빈도가 1인 특허의 수는 115,756,809건이며 최대 서지 결합도는 780회 였다. 동시인용 빈도가 1인 특허는 103,820,840건이며 특허의 최대 동시인용빈도의 값은 1,029건이 였다.



<그림 4> BC& CC 네트웍

특허인용정보를 대상으로 수행한 결과 또한 학술문헌을 이용하여 분석한 결과[5]와 유사하게 멱함수 체제로 감소하는 분포를 갖는다. 따라서 특허정보의 서지결합과 동시인용 네트웍은 멱함수 분포를 따르는 척도 없는 네트워크(scale-free network)이라고 말할 수 있다.

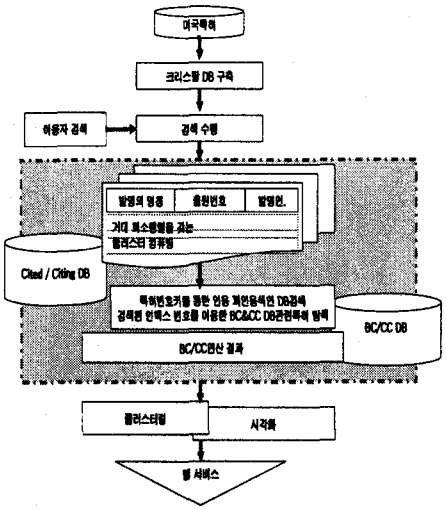
다음 장에서는 이를 기반으로 현재 한국과학기술정보연구원에서 서비스하고 있는 특허정보 서비스를 계량정보 분석이 포함된 시스템으로 대체하기 위한 서비스 모델을 제안하고자 한다.

4. 유통 서비스 모델

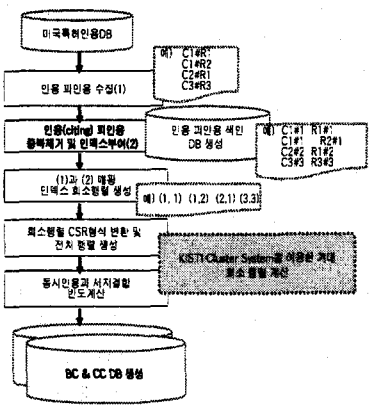
4.1 전체 프로세스

유통 서비스 모델은 <그림 5>과 같이 크게 3개의 모듈로 구성된다. 1)특허 데이터의 크리스탈 DB 적재 및 색인을 수행하기 위한 검색 모듈, 2)본 논문에서 제시한 인용정보 분석 모듈, 3)검색의 효율성을 제고시키기 위한 시각화 모듈로 구성된다.

이용자가 검색을 수행한 후 간략검색 화면에서 찾고자 하는 특허를 선택하여 상세검색 화면으로 이동화면 해당 특허에 대한 동시인용과 서지 결합된 특허 및 그 빈도에 대한 정보를 이용자에게 제공한다.



<그림 5> 유통 서비스 시스템 모델



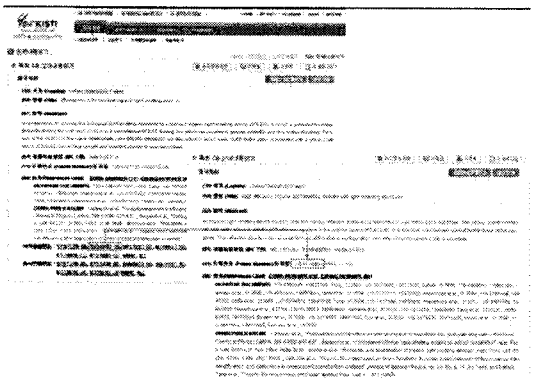
<그림 6> 인용정보 분석모듈

4.2 인용 정보 분석 모듈

인용정보 분석모듈의 구성은 <그림 6>와 같이 5단계로 구성된다. (1) 미국특허 DB로부터 인용/피인용 정보를 수집, (2) 수집된 인용정보로부터 인용 특허와 피인용 특허에 대해 유일한 인덱스 값을 부여하여 인용/피인용 색인 DB를 생성, (3) (1)과 (2)를 매핑하여 희소행렬 생성, (4) 생성된 희소행렬을 CSR 품으로 변환하고 CSR형태의 전치 행렬을 생성한다. 마지막으로 KISTI 클러스터 시스템을 활용하여 동시인용빈도와 서지결합빈도를 계산한 후 서지 결합과 동시인용 빈도 DB를 생성한다.

4.3 정보유통 서비스 예

4.2절에서 계산된 결과를 사용하여 한국과학기술정보연구원의 정보 유통 서비스에 적용하는 예는 <그림 7>과 같다.



<그림 7> 정보유통 서비스의 예

5. 결 론

정보의 홍수속에서 적합한 정보를 찾기란 쉬운일이 아니다. 사용자의 정보 탐색에 대한 부하를 줄이기 위해 본 논문에서는 여러 개의 문헌이 공통으로 인용하는 문헌의 개수에 따라 주제의 공통성을 측정하는 서지 결합 정보와 핵심개념이나 방법론을 유추해 내는 동시인용정보를 이용하여 사용자의 정보 탐색에 대한 부하를 줄이고자 하였다. 그리고 거대 희소 행렬을 갖는 인용정보의 BC & CC를 계산하기 위해 클러스터 컴퓨터를 사용하여 정보서비스 시스템과 연계하는 모델을 제안하였다. 향후 빈번히 발생하는 데이터베이스 갱신시 실시간으로 서비스하기 위한 신속하고 경제적인 희소행렬 계산 알고리즘에 대한 연구가 필요하다.

< 참 고 문 헌 >

[1] Kessler M. M. "Bibliographic Coupling Between Scientific Papers", AD,pp10-25, 1963.

[2] Kim.H.H, Kim,Y.H. "Bibliometric", KoomiTrade, p132-133, 1993.

[3] Eun-Jin Kim, Kyung-Hoon Kim, "Efficient Sparse Matrix-Matrix Multiplication for pursuit optimization", Korean Multimedia Society Autumn Conference, 2003.

[4]S.Blackford, G.Corliss, J.Dongarra, I.Duff, S.Hammarling, G.Henry, M.Heroux, C.Hu, W.Kahan, L.Kaufman, B.Kearfott, F.Krogh, X.Li, Z.Marrny, A.Petitet, R.Pozo, K.Remington, W.Walster, C.Whaley, and J.W. von Gudenberg, "Document for the Basic Linear Algebra Subprograms (BLAS) standard", BALS Technical Forum, 2001.

[5] Anthony F.J. Van Raan, "Reference-based publication networks with episodic memories", Scientometrics, Vol. 63, No. 3 pp.549-566, 2005.