Journal of the Korean Data & Information Science Society
2013, **24**(5), 999–1011

http://dx.doi.org/10.7465/jkdi.2013.24.5.999
한국데이터정보과학회지

거대 인용 자료를 이용한 문서 추천 방법[†]

채민우[1] · 강민수[2] · 김용대[3]

[13]서울대학교 통계학과 · [2]광개토연구소

접수 2013년 6월 30일, 수정 2013년 7월 23일, 게재확정 2013년 8월 5일

요 약

본 연구에서는 논문이나 특허 등의 문서들의 인용 정보를 활용하여 연관성이 높고 중요한 특허를 추천하는 방법을 제안한다. 문서 간의 연관성 지표인 공통피인용횟수와 중요도 지표인 HITS를 적절한 형태로 결합한 뉴먼 커널로부터 두 정보의 반영 정도를 조율하는 것이 핵심이다. 제안하는 방법은 미래의 인용에 대한 예측 오차를 최소화하는 것으로 이를 통해 뉴먼 커널의 조율모수 γ를 적절하게 선택할 수 있다. 또한, 거대 인용 자료를 분석하기 위해 필요한 계산 기술에 대해서 자세히 논의한다. 마지막으로, 미국 등록 특허 400만 건에 대한 실증적 자료 분석을 시행한다.

주요용어: 거대 자료, 뉴먼 커널, 성긴 행렬 연산, 인용 자료 분석, 추천.

1. 서론

특정 문서가 다른 문서와 연관성이 있는지, 그리고 얼마나 중요한 문서인지는 그 문서에 사용된 단어, 문장, 글의 흐름, 사용된 그림이나 표 등의 내용을 분석함으로써 알 수 있다. 언어, 특히 단어와 단어가 유기적으로 연결되어 있는 하나의 글을 모델링하는 것은 매우 어렵기 때문에, 통계적으로 쉽게 다룰 수 있는 간단한 형태로 문서를 변환한 후 이를 모델링 하는 것이 훨씬 효과적이다. 예를 들어, Blei 등 (2003)이 제안한 LDA (latent Dirichlet allocation) 모형은 문서를 단순히 단어를 모아놓은 집합 (bag of words), 즉 단어의 빈도를 나타내는 벡터로 변환한 후에 다항혼합분포 (mixture of multinomial distributions)를 사용하여 문서를 모델링한다.

LDA 모델의 발전과 더불어 지난 십수년 간 문서를 주제별로 분류하는 수많은 통계모형 및 알고리즘이 개발되었다. Teh 등 (2006)은 계층적 디리끌레 과정을 사용하여 문서를 대주제, 소주제 등의 계층구조를 갖는 주제 모형을 제안하였고 Li와 McCallum (2006)와 Blei와 Lafferty (2007) 등의 문헌에서는 주제들 간의 연관성을 고려하는 모형을 제안하였다. 이러한 방법들은 자동 분류 (Blei 등, 2003), 검색 (Wei와 Croft, 2006), 협업 필터링 (Hofmann, 2004), 추천 시스템 (Jannach, 2010) 등의 다양한 분야에서 활용되고 있다. 하지만 짧은 문서일지라도 수십 또는 수백 개의 단어를 포함하고 있기 때문에 수백만 건의 문서를 모델링하기 위해서는 수천 또는 수만 개의 단어를 포함하는 사전이 필요하고 이러한 거대자료를 처리하기 위해서는 엄청난 계산 비용이 소모된다.

한편, 엄청난 양의 문서 내용을 직접 분석하지 않고도 문서 간의 연관성이나 중요도를 평가할 수 있는데 인용 정보를 사용하는 것이 대표적인 방법이다 (Garfield와 Merton, 1979). 과학 분야의 논문을 예

[†] 본 논문은 농촌진흥청 공동연구사업 (과제번호: PJ907160)의 지원에 의해 이루어진 것임.
[1] 교신저자: (151-747) 서울특별시 관악구 관악 1로, 서울대학교 통계학과, 박사과정 연구원.
E-mail: demian87@snu.ac.kr
[2] (135-921) 서울특별시 강남구 역삼동 725-32, 우일빌딩 1층, 광개토연구소, 대표이사.
[3] (151-747) 서울특별시 관악구 관악 1로, 서울대학교 통계학과, 교수.

로 들자면 피인용 횟수가 많은 논문일수록 중요한 논문일 가능성이 크다는 것은 너무나 당연한 사실이다. 또한, 두 논문이 서로를 인용하고 있거나 공통으로 인용한 논문의 수가 많을 경우 두 논문은 연관성이 크다고 할 수 있다 (Kessler, 1963). 마찬가지로 두 논문을 공통으로 인용하고 있는 다른 논문의 수가 많을 경우에도 두 논문은 많은 관련이 있다고 볼 수 있다 (Small, 1973). 하나의 논문에 들어가는 단어의 수가 보통 수천 이상이지만 참고문헌의 수는 많아야 수십 개인 것을 감안하면 인용 정보를 분석하는 데 필요한 비용이 문서의 내용을 분석하는 것에 비해 훨씬 저렴할 거라 추측할 수 있다. 또한, 문서의 내용을 모델링하는 대표 주자인 토픽모델 (topic model)에 비해 인용 정보를 사용한 분석 결과는 그 의미가 훨씬 간결하기 때문에 보다 직관적인 해석을 내릴 수 있다. 다시 말해, 토픽모델의 결과를 이해하기 위해서는 각 주제가 포함하는 단어의 분포를 보고 인위적인 해석을 내려야 하지만 인용 정보를 활용할 경우 그럴 필요가 전혀 없다. 따라서 문서 간의 연관성이나 상대적 중요도를 추정하고자 할 경우에 인용 정보를 활용하면 내용을 온전히 분석하는 것에 비해 비용을 크게 줄일 수 있을 뿐만 아니라 결과물의 질을 향상시키는 데도 큰 도움이 된다.

본 논문에서는 인용 정보를 활용하여 해당 문서와의 관련이 깊으면서 동시에 중요성이 큰 문서를 추천하는 방법을 제안하고자 한다. 다시 말해, 인용 정보를 통해 문서 간의 상대적 중요도를 수치화하고 이를 통해 특정 문서를 접한 사용자에게 그 문서와의 상대적 중요도가 높은 다른 문서를 추천하는 방법론을 개발하는 것이 목적이다. 이러한 문서 추천에 관한 선행 연구로 Strohman 등 (2007)은 텍스트 내용과 인용 정보로부터 나온 변수를 선형 결합하여 문서 간의 연관성 측도로 사용하였으며 Katz 거리 (Liben-Nowell와 Kleinberg, 2007)가 매우 중요한 측도라는 결론을 내렸다. McNee 등 (2002)에서는 저자-인용행렬, 문서-인용행렬 및 공통피인용행렬 등의 여러 측도를 사용하여 문서의 상대적 순위를 매겼다. Tang와 Zhang (2009)은 토픽 모델과 인용 정보를 결합하여 해당 문서의 참고문헌에서 주제별로 문서를 추천하는 방법을 제안하였다. 이들 방법의 단점은 제안된 방법을 적용하기 위해서 추천할 만한 문서의 리스트 (참고문헌)가 사전에 주어져야만 한다는 것이다. 반면, He 등 (2010)은 비슷한 방법으로 이러한 리스트가 없어도 전체 문서에서 추천 리스트를 뽑을 수 있지만 이들은 인용정보보다는 문서의 내용을 중점적으로 활용하였다. 본 논문에서는 문서 추천을 주로 다루지만 이 방법은 문서뿐만이 아니라 인터넷 웹페이지의 하이퍼링크와 같이 인용 구조가 있는 어떠한 아이템에도 적용이 가능하다. 구글과 같은 인터넷 검색 엔진에서 하이퍼링크 정보를 활용하여 검색의 질을 크게 향상시킨 이야기는 매우 유명하다 (Brin와 Page, 1998). 본 연구의 결과는 연관성이 큰 논문이나 특허 추천, 관련 웹페이지 추천 등 수많은 네트워크 자료에서 활용이 가능할 것으로 보인다.

본 논문의 후반부에는 자료의 저장 및 처리에 관한 것을 주의 깊게 다루고자 한다. 사실 수십만 또는 수백만 건에 이르는 거대 문서로부터 인용 정보를 추출하면 그 자체만으로 엄청난 양이기 때문에 이를 일반적인 네트워크 자료를 처리하는 방법으로 저장하거나 계산하기는 힘들다. 왜냐하면 인용 정보는 모든 문서 쌍을 고려해야 하는데 이 숫자는 문서 수의 제곱에 비례해서 증가하기 때문이다. 보통 네트워크 자료는 그래프를 행렬로 변환하여 저장하는 것이 일반적인데 거대 네트워크 행렬을 일반적인 저장 방식으로 처리하면 현재 통용되는 컴퓨터의 메모리 용량으로 감당할 수가 없다. 따라서 행렬을 저장하고 계산하기 위해 Saad (1990)와 같이 성긴 행렬을 다루는 특별한 데이터 구조 및 알고리즘이 요구된다.

본 논문의 흐름은 다음과 같다. 2절에서는 인용 그래프에 대한 소개를 하고 그래프 노드의 연관성 및 중요성을 나타내는 지표들 중 본 논문에 필요한 것들을 설명한다. 3절에서는 2절에서 소개되는 지표들을 사용하여 문서를 추천하는 방법에 대해서 자세히 다룬다. 4절에서는 거대 인용 자료 분석 시 필요한 계산 기법에 대한 설명을 하고 5절에서는 미국 등록 특허 약 400만 건에 대한 실증적 자료 분석을 시행한다. 마지막으로 6절에서 결론 및 제언을 통해 글을 마무리한다.

2. 인용 그래프 및 노드의 특성

이 절에서는 인용 자료를 그래프로 표현하는 방법을 설명하고 인용 정보에 의해 도출되는 그래프 노드 상의 지표들을 소개한다. 먼저 인용 그래프를 정의하고 그래프의 인접행렬 (adjacency matrix)로부터 노드 간의 연관성 정도를 나타내는 공통인용행렬, 공통피인용행렬을 정의한다. 다음으로 노드의 절대적 중요도를 나타내는 HITS (hypertext-induced topic search)를 소개한다. 마지막으로 노드 간의 연관성 수치와 중요도 수치를 결합하여 특정 노드에 대한 상대적 중요도를 나타내는 뉴먼 커널에 대한 설명을 한다. 이번 절에서 설명하는 것보다 더 다양한 내용은 Shimbo와 Ito (2006)와 그 곳에 나오는 참고문헌들을 보면 찾을 수 있다.

2.1. 인용 그래프

그래프 이론을 사용하여 네트워크 자료를 표현하고 분석하는 방법은 많은 교재를 통해 잘 알려져 있다. Cook과 Holder (2006)나 Kolaczyk (2009)와 같은 책에 그래프 및 네트워크 데이터에 대한 기본적인 개념부터 시작하여 관련 이론과 다양한 응용 분야가 잘 정리되어 있다. 여기서는 본 연구에 필요한 간단한 개념만을 설명하기로 한다.

인용은 항상 방향성을 가지기 때문에 모든 인용 자료는 방향 그래프 (directed graph)를 사용하여 표현할 수 있다. 문서 전체 집합을 노드의 집합 V라고 정의하고 $V = \{1, \ldots, n\}$로 표기하자. 서로 다른 두 문서 $i, j \in V$에 대하여 문서 i가 문서 j를 인용하였으면 (i, j)를 모서리 집합 E에 포함시키자. $G = (V, E)$라고 하면 G는 방향 그래프가 된다. 이를 인용 그래프라고 정의하자. 그래프의 인접행렬을 $A = (a_{ij})_{1 \le i,j \le n}$라고 하자. 여기서 a_{ij}는 $(i, j) \in E$일 경우 1, 그 외의 경우 0으로 정의된 값이다. 반대로 $n \times n$ 이진행렬 A가 있을 때 A를 인접행렬로 갖는 그래프 $G = (V, E)$를 정의할 수 있다. 무향 그래프 (undirected graph)의 경우 인접행렬 A는 대칭행렬이 된다. 방향 그래프, 무향 그래프 모두 각 모서리에 음이 아닌 실수 (또는 정수) 값을 갖는 가중치를 줄 수 있다. 이 때는 인접행렬의 성분이 0과 1로만 이루어진 것이 아니라 음이 아닌 실수 값을 가질 수 있다.

2.2. 노드 간의 연관성 지표

인용 정보를 통해 두 문서의 연관성을 파악하는 가장 쉬운 방법은 서로 간의 인용 여부를 확인하는 것이다. 즉, 두 문서 i와 j가 있을 때 서로가 서로를 인용하고 있는 경우, 또는 둘 중 하나의 문서가 다른 하나를 인용하고 있는 경우 두 문서가 연관성이 높다고 볼 수 있다. 하지만 이 방식을 통해 나타낼 수 있는 연관성 수준은 오직 3개뿐이고 더욱 큰 문제는 특정 문서와 연관성이 있는 문서의 수가 그 문서의 참고문헌 또는 피인용 수에 비례한다는 것이다. 또한, 인용 자체가 문서를 작성한 사람의 행위에 의한 것이기 때문에 그 사람이 모르고 있던 문서 중 연관성이 높은 문서를 찾아내기가 어렵다. 이러한 이유로 직접적인 인용 여부만으로 문서 간의 연관성을 평가하는 것은 좋은 방법이라고 볼 수 없다.

문서 간의 연관성을 나타내는 가장 대표적인 지표로는 두 개의 문서를 공통으로 인용하고 있는 문서의 수 (co-citation), 두 문서가 공통으로 인용하고 있는 문서의 수 (bibliographic coupling)를 생각할 수 있다. 간단한 행렬 곱셈을 통해서 모든 문서 쌍에 대하여 이 두 값을 표현할 수 있다. 행렬 $A^T A$를 공통피인용행렬 (Small, 1973), 행렬 AA^T를 공통인용행렬 (Kessler, 1963)이라고 정의하자. 그러면 $A^T A$의 (i, j) 성분은 i와 j를 공통으로 인용하고 있는 문서의 수를 나타내고 AA^T의 (i, j) 성분은 i와 j가 공통으로 인용하고 있는 문서의 수를 나타낸다. 공통인용행렬과 공통피인용행렬에 대응하는 그래프를 생각할 수 있는데 둘 모두 대칭행렬이기 때문에 무향 그래프이다.

2.3. 절대적 중요도

비슷한 단어가 많이 나오는 문서들은 서로 연관성이 높다고 생각할 수 있다. 반면에 문서의 내용을 통계적으로 분석해서 추정하기 가장 힘든 것 중 하나로 해당 문서가 얼마나 중요한 지를 나타내는 지표이다. 그 문서의 중요도는 단순히 단어의 수를 세는 통계만으로 파악할 수 있는 것이 아니기 때문이다. 그러나 저자의 인위적 행위로 간주할 수 있는 인용 정보를 사용하면 이를 매우 쉽게 추정할 수 있는데 단순히 피인용 횟수를 세는 것만으로도 그 문서의 중요도를 대략적으로 파악할 수 있기 때문이다.

단순히 피인용 횟수를 세는 것보다는 조금 복잡한 방법 중 최근 많이 사용되는 것으로는 Kleinberg (1999)가 제안한 HITS와 Page와 Brin (1999)이 제안한 PageRank 등이 있다. 둘 모두 문서 하나의 인용을 같은 값으로 취급하지 않고 그 문서가 받은 인용 횟수에 따라 가중치를 다르게 준다는 것이 핵심 아이디어이다. 또한 둘 모두 반복적으로 행렬과 벡터를 곱하는 알고리즘을 통해 매우 큰 네트워크 자료에서도 비교적 쉽게 계산을 할 수 있다는 장점이 있다. 본 논문에서 다루고자 하는 뉴먼 커널은 HITS와 관련된 것이기 때문에 여기서는 이에 대한 내용만을 다루기로 한다.

HITS 알고리즘에는 권위 (authority) 스코어와 허브 (hurb) 스코어라는 두 종류의 값이 나오는데 반복적인 방법으로 이들을 정의할 수 있다. 우선 a_0와 h_0를 모든 성분이 1인 n차원 벡터라고 하자. 앞서 표기한 대로 A를 인용 그래프의 인접행렬이라고 하고

$$a_{m+1} = \frac{A^T h_m}{|A^T h_m|} \qquad h_{m+1} = \frac{A a_m}{|A a_m|}$$

이라고 정의하자. 여기서, $x = (x_1, \ldots, x_n)$에 대하여 $|x| = (\sum_{i=1}^n x_i^2)^{1/2}$이다. 권위벡터는 $a = \lim_{m\to\infty} a_m$, 허브벡터는 $h = \lim_{m\to\infty} h_m$로 정의하고 각각의 i번째 성분을 노드 i의 권위 스코어, 허브 스코어라고 하자. 이렇게 정의한 $\{a_m\}$과 $\{h_m\}$이 잘 수렴한다고 가정하면

$$a_{m-1} \approx a_{m+1} = \frac{A^T h_m}{|A^T h_m|} = \frac{A^T A a_{m-1}}{|A^T A a_{m-1}|}$$

을 만족한다. 허브벡터에 대해서도 비슷한 이야기를 할 수 있기 때문에 직관적으로 권위벡터와 허브벡터가 인용행렬 및 피인용행렬의 고유벡터 (eigenvector)와 밀접한 관련이 있을 것이라는 것을 알 수 있다. 만약 A^TA와 AA^T의 가장 큰 고유값 (eigenvalue)의 다중도 (multiplicity)가 1이면 스펙트럴 정리를 사용하여 권위벡터 및 허브벡터가 잘 정의되고 각각 A^TA와 AA^T의 가장 큰 고유값에 대응되는 고유벡터와 같아진다는 것을 보일 수 있다. 더욱이 이 스코어 값들은 모두 음이 아닌 값을 갖는다는 것 또한 증명이 가능하다.

한편, 권위 스코어와 허브 스코어는 $a = A^T h$와 $h = Aa$ 관계를 만족하기 때문에 이들에 대하여 매우 직관적인 해석을 내릴 수 있다. $a_i = \sum_{k=1}^n A_{ki} h_k$이므로 권위 스코어가 높은 노드는 중요한 노드들이 많이 인용을 하고 있는 노드이고 $h_i = \sum_{k=1}^n A_{ik} a_k$를 고려하면 허브 스코어가 높은 노드는 중요한 노드를 많이 인용하고 있는 노드이다. 따라서 두 스코어는 약간 다른 의미에서 각 노드에 대한 전반적인 중요도를 나타내는 값으로 이해할 수 있다.

2.4. 뉴먼 커널에 의한 상대적 중요도

이번 절의 앞선 두 장에서 설명한 연관성 및 중요도 지표는 연관성 또는 중요성을 바탕으로 한 문서 추천 시스템에서 중요한 점수로 사용될 수 있다. 하지만 문서 i에 관심이 있는 사용자에게 단순히 i와 연관성만 높은 문서를 추천하거나 문서 i와의 연관성을 전혀 고려하지 않고 중요도가 높기만 한 문서를 추천하는 것은 바람직하지 않다. 가장 좋은 것은 둘을 유기적으로 결합하여 i와의 연관성도 높으면서 동시에 중요성도 높은 문서를 추천하는 것이다. 이번 장에서는 대표적 방법인 뉴먼 커널 (Kandola 등,

2003)에 대해서 설명하기로 한다. 이는 2.2와 2.3에서 다룬 두 종류의 측도를 자연스럽게 결합한 측도이다.

먼저, 대칭행렬 B가 주어졌을 때 $\rho(B)$를 행렬 B의 스펙트럴 반경이라고 정의하자. 임의의 $\gamma < \rho(B)$에 대하여 $N_\gamma(B)$를 다음과 같이 정의할 수 있다.

$$N_\gamma(B) = B(I - \gamma B)^{-1} = B\sum_{k=0}^{\infty}(\gamma B)^k \tag{2.1}$$

γ가 $\rho(B)$보다 작을 때에는 $I - \gamma B$의 역행렬이 잘 정의되고 우변의 무한급수 또한 잘 정의된다. 이제 인용그래프의 뉴먼 커널 $\widehat{K}$와 $\widehat{M}$을 다음과 같이 정의하자.

$$\widehat{K}_\gamma = N_\gamma(A^TA) = A^TA(I - \gamma A^TA)^{-1} = A^TA\sum_{k=0}^{\infty}\gamma^k(A^TA)^k \tag{2.2}$$

$$\widehat{M}_\gamma = N_\gamma(AA^T) = AA^T(I - \gamma AA^T)^{-1} = AA^T\sum_{k=0}^{\infty}\gamma^k(AA^T)^k \tag{2.3}$$

조금 복잡해 보이지만, 수식 (2.2)와 (2.3)에 대하여 명료한 해석을 내릴 수 있다. 먼저, $\widehat{M}$은 $\widehat{K}$와 비슷한 방식으로 생각할 수 있기 때문에 $\widehat{K}$에 국한에서 설명하기로 하자. $(A^TA)^k$의 (i,j) 성분은 공통피인용행렬 A^TA에 대응되는 그래프에서 노드 i와 노드 j 사이의 길 (path) 중에서 길이가 k인 것의 개수에 해당하는 수이다. 이 때, A^TA의 (i,j) 성분이 k이면 대응되는 그래프의 노드 i와 노드 j를 잇는 서로 다른 k개의 모서리가 있다고 간주한다. 이를 인용 그래프 G상에서 설명하자면, k개의 중간 단계를 거쳐서 i와 j를 공통으로 인용하고 있는 문서의 수를 말한다. Figure 2.1을 보면 그 의미를 쉽게 이해할 수 있다. 따라서 k값이 작을 때의 $(A^TA)^k$ 행렬의 (i,j) 성분은 두 문서 간의 연관성을 재는 측도로 간주할 수 있다. 반면 k가 큰 경우를 생각해보자. 만약 $\lambda = \rho(A^TA)$의 다중도가 1이라면 행렬의 스펙트럴 분해에 의해 $(A^TA/\lambda)^k$이 행렬 aa^T로 수렴하는 것을 알 수 있다. λ를 고정된 상수라고 생각한다면 이는 k값이 클 경우 A^TA의 (i,j) 성분이 노드 i와 노드 j의 권위 스코어 값을 곱한 값에 거의 비례하는 것을 의미한다. 즉, k가 클 경우에는 이를 두 노드의 중요도 정도로 생각할 수 있다. 결론적으로 뉴먼 커널의 (i,j) 성분은 중요도와 연관성이 적절한 형태로 결합되어 있는 것으로 생각할 수 있다. 하나의 노드 i를 고정시켜 놓고 모든 가능한 j를 생각하면 이는 노드 i에 대한 상대적 중요도로 간주할 수 있다.

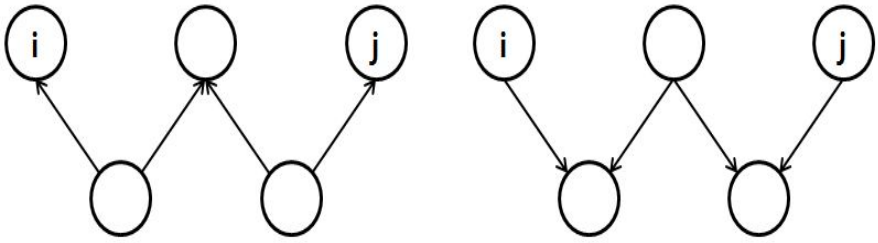

Figure 2.1 Graphical interpretation for the (i,j) element of $(A^TA)^2$ (left) and $(AA^T)^2$ (right)

3. 상대적 중요도에 의한 문서 추천 방법

3.1. 조율모수 γ의 의미

뉴먼 커널에서 조율모수 γ는 매우 중요한 의미를 갖고 있다. 이번 절에서도 $\widehat{M}$의 경우는 같은 방식으로 이해할 수 있기 때문에 $\widehat{K}$에 치중해서 설명을 하기로 한다. 수식 (2.2)를 보면 조율모수 γ값이 0으로 수렴함에 따라 $\widehat{K}_\gamma$는 A^TA로 수렴한다는 것을 알 수 있다. A^TA는 공통피인용행렬을 의미하기 때문에 γ의 값이 0에 가깝다는 것은 뉴먼 커널이 거의 연관성 측도로 편향되어 있다는 것을 의미한다. 반면 γ의 값이 커지는 경우를 생각해보자. 뉴먼 커널이 잘 정의되기 위해서는 γ의 값이 $1/\lambda$보다 작아야 한다. 여기서 λ는 A^TA의 스펙트럴 반경, 즉 가장 큰 고유값이다. 따라서 γ가 $1/\lambda$로 증가할 때의 뉴먼 커널의 극한을 생각하면 큰 γ 값에 대한 뉴먼 커널의 의미를 이해할 수 있다. 먼저 아래의 정리를 살펴보자.

정리 3.1 공통피인용행렬 A^TA의 가장 큰 고유값 λ의 다중도가 1이면 $\gamma \nearrow \lambda^{-1}$일 때

$$(\lambda^{-1}-\gamma)N_\gamma(A^TA) \to aa^T$$

이 성립한다.

증명은 Shimbo와 Ito (2006)의 Theorem 12.2를 보면 된다. 이 정리를 살펴보면 γ가 증가함에 따라 뉴먼 커널이 권위 스코어에만 의존한다는 사실을 알 수 있다. 즉, γ값이 커질수록 뉴먼 커널은 중요도 쪽으로 편향되는 경향을 보인다. 따라서 이 둘의 조합을 통해 결론을 내리자면 뉴먼 커널에서의 조율모수 γ는 커널의 값이 연관성과 중요성 중 어느 쪽에 더 비중을 둘 것인가를 결정하는 역할을 한다. 값이 0에 가까울수록 연관성 쪽으로 치우치는 반면, 값이 $1/\lambda$로 커질수록 중요도 쪽으로 큰 비중을 두게 된다.

3.2. γ 조율을 통한 추천 방법

White와 Smyth (2003)에서 언급한 바와 같이 조율모수 γ를 결정하는 일반적인 규칙이나 방법은 없다. 3.1에서 설명했듯이 γ는 연관성과 중요성의 정도를 결정하는 모수이기 때문에 고려하는 문제의 목적에 맞는 적절한 방법을 선택해야 한다. 또한, 문제의 목적이 명확하다고 하더라도 특정한 γ 값에 대해서 뉴먼 커널의 값이 중요도 50%, 연관성 50%와 같이 명확하게 구분되는 것이 아니기 때문에 이를 적절한 값으로 조율하는 것은 쉬운 작업이 아니다. 본 연구의 목적은 인용 정보를 활용하여 연관성이 높은 중요한 문서를 추천하는 것이기 때문에 이번 장에서는 이에 기반한 조율모수 결정법을 제안하도록 한다.

우선, 분석 대상인 문서 전체에 대하여 인용이 발생한 시기를 알 수 있다고 가정하자. 논문이나 특허 자료의 경우 게재년도 또는 출원년도 등을 알 수 있기 때문에 참고문헌의 모든 인용이 이 시기에 이루어졌다고 생각해도 큰 무리가 없다. 반면 인터넷 웹페이지의 경우 모든 웹페이지에 대하여 링크가 생긴 날짜를 아는 것은 어렵기 때문에 이와 같은 방법을 적용하기는 힘들다. 일단, 인용 시기를 안다고 가정한 후에는 교차검증방법과 마찬가지로 검증 자료에 대한 예측 오차를 줄이는 방법을 생각할 수 있다. 제안하는 방법의 적용 대상은 주로 논문 또는 특허이기 때문에 문서의 작성 시기와 그 문서에 나오는 인용의 시기가 일치한다고 가정하기로 한다.

먼저, 기준 시점 T를 정하고 모든 인용 정보를 T 시점 전과 후의 자료 둘로 나누자. 그리고 T 시점 전의 자료만을 사용하여 모든 γ값에 대하여 뉴먼 커널을 계산한다. 이제 T 시점 이후의 인용 자료를 바탕으로 T 시점 이후 발생한 각 문서 쌍에 대한 공통피인용횟수를 구할 수 있다. 이제 T 시점 이전의 자료로 구한 뉴먼 커널과 T 시점 이후의 자료로 구한 공통피인용행렬을 비교함으로써 두 행렬의 추천 성능이 가장 비슷해지는 γ값을 선택하면 된다. 여기서 두 행렬의 추천 성능이 비슷해진다는 것은 각 행렬에 의해 추천되는 문서가 비슷해진다는 것을 의미한다. 특정 문서 i에 대하여 각 행렬의 (i,j) 성분이 높

은 순서대로 다른 문서를 추천해주기 때문에 문서 i에 대한 추천 성능이 비슷하다는 것은 두 행렬, T 시점 이전의 자료로 구한 뉴먼커널과 T 시점 이후의 자료로 구한 공통피인용행렬의 i번째 행을 추출했을 때 순위상관계수가 1에 가깝다는 것을 의미한다. 전체적인 추천 성능을 비교하기 위해서 모든 i에 대해 얻어진 순위상관계수를 평균내어 그 값이 가장 높은 γ를 선택한다.

전체 자료를 사용하여 구한 공통피인용행렬의 스펙트럴 반경은 T 시점까지의 자료만으로 구한 공통피인용행렬의 스펙트럴 반경보다 값이 조금 더 크기 때문에 위의 방법으로 추정한 γ는 약간의 보완이 필요하다. γ의 범위가 0부터 $1/\lambda$이기 때문에 $1/\lambda$에 대한 γ의 비율을 추정값으로 사용하면 이를 전체 자료에서 구한 뉴먼 커널에 적용하는데 사용할 수 있다. 즉, γ를 추정하는 것이 아니라 $\gamma\lambda$의 값을 추정해 사용하는 것이다. $\gamma\lambda$의 범위는 자료에 의존하지 않고 항상 0에서 1 사이의 값을 갖기 때문에 이 방법을 사용하면 T 시점 이전의 자료만으로 추정한 값을 전체 자료로 확장하더라도 아무 문제가 없다. 또한, 추정된 γ는 문서 추천에 있어 중요도와 연관성의 정도를 결정짓는 모수이기 때문에 기준 시점 T를 어떻게 정하든지 크게 상관이 없다. 따라서, 자료의 수를 보고 T 시점 이전의 인용 자료가 70% 정도가 되는 시점으로 결정하는 것을 추천한다. 이번 장에서 설명한 방법은 Algorithm 3.1에 요약되어 있다.

Algorithm 3.1 Documents recommendation by tuning Neumann kernels

1. Divide whole citation data into two groups D_1 and D_2, where D_1 are data before time T and D_2 are remainders.
2. Using D_1, calculate $\hat{K}_\gamma$ for all γ.
3. Using D_2, calculate a co-citation matrix B^TB.
4. For each corresponding columns between $\hat{K}_\gamma$ and B^TB, find rank correlation coefficients.
5. Find the estimator $\widehat{\gamma\lambda}$ as the minizer of mean of rank correlation coefficients obtained in 4.
6. Calculate the Neumann kernel at $\widehat{\gamma\lambda}$ using whole citation data.
7. Recommend documents based on the Neumann kernel obtained in 6.

4. 계산 문제

이번 절에서는 뉴먼 커널 계산 시 고려해야 하는 사항을 다룬다. 우선 왜 이런 문제를 고려해야 하는지 생각해보자. 문서의 수가 기껏해야 수백에서 수천 정도일 때에는 이번 절에서 다루는 내용을 전혀 고려할 필요가 없다. 하지만 문서의 수 n이 커짐에 따라 고려해야 하는 문서 쌍의 수는 n^2의 차수로 증가하기 때문에 n이 30,000 정도만 되어도 뉴먼 커널을 저장하기 위해서는 9억 개의 실수 자료를 저장할 수 있는 공간이 필요하다. 이를 메모리로 환산하면 대략 7기가바이트에 달하는 양으로 현재 사용되는 일반적인 계산용 컴퓨터 수준을 고려했을 때 여러 γ 값에 대한 뉴먼 커널을 계산하는 것이 사실상 불가능하다. 또한, 뉴먼 커널을 계산하기 위해서는 행렬의 역행렬 계산이 필수적인데 n의 값이 큰 경우 이는 매우 어려운 문제이다.

다행히도 논문이나 특허와 같은 문서 하나에 나오는 참고문헌의 수는 기껏해야 수십 개 정도이기 때문에 인용행렬 A는 대부분의 성분이 0인 매우 성긴 행렬이다. 따라서, 행렬의 모든 성분을 저장하지 않고 0이 아닌 성분의 위치와 값만을 저장하면 행렬이 매우 크고 성긴 경우 공간을 크게 절약할 수 있다. 하지만 이렇게 행렬을 저장할 경우 메모리에 접근하는 방식에 따라 행렬 간의 덧셈이나 곱셈의 속도 차이가 클 수 있기 때문에 이에 맞는 데이터 구조가 필요하다.

한편, 거대 행렬의 역행렬을 구하는 것은 대상 행렬이 성긴 행렬일지라도 매우 어려운데, 수식 (2.2)나 (2.3)을 보면 뉴먼 커널 계산 시 필요한 역행렬 연산이 행렬의 합과 곱으로 근사될 수 있다는 것을 알 수 있다. 행렬의 역행렬 계산에 비해 행렬의 덧셈과 곱셈은 훨씬 간단한 문제이기 때문에 이를 통해 뉴먼 커널의 근사값을 계산하면 된다. 행렬의 역행렬을 구하는 계산 복잡도는 대략 $O(|V|^3)$이지만 첫 k개의 항만을 계산하여 근사함으로써 계산 시간을 크게 줄일 수 있다. 또한, 이 방법의 매우 큰 장점 중 하나는 분산 처리가 가능하다는 것이다. 행렬의 덧셈이나 곱셈은 역행렬 계산과 달리 각 성분을

독립적으로 계산할 수 있는데 계산 분야에서 최근 각광받는 GPU (Sanders와 Kandrot, 2010)나 하둡 시스템 (Lam, 2010)을 통해 분산처리를 하면 계산 시간을 크게 단축할 수 있다.

물론 인용 그래프의 인접행렬이 매우 성긴 행렬이라고 하더라도 뉴먼 커널은 그렇지 않은 경우가 대부분이다. 수식 (2.2)나 (2.3)을 보면 뉴먼 커널을 계산하기 위해서 A^TA 또는 AA^T를 계속해서 곱해나가야 하는데 이 과정을 진행하면 0이 아닌 성분의 수가 급격하게 늘어난다. 따라서 그래프의 규모가 큰 경우 뉴먼 커널 전체를 구하는 것은 사실상 불가능한 일이다. 그러나 대부분의 경우, 뉴먼 커널 전체를 구하는 것에 관심이 있는 것이 아니라 특정 노드 i에 대한 다른 문서의 상대적 중요도를 구하는 것이 주요 쟁점이기 때문에 전체 행렬 연산을 할 필요가 없다. 뉴먼 커널을 사용하여 문서 i에 대한 상대적 중요도를 구하기 위해서는 $(A^TA)^k\mathbf{e}_i$ 또는 $(AA^T)^k\mathbf{e}_i$를 구하는 것으로 충분하다. 여기서 $\mathbf{e}_i$는 i번째 성분이 1이고 나머지 성분은 모두 0인 단위벡터를 말한다. A가 매우 성긴 행렬이면 A^TA 또한 매우 성긴 행렬이기 때문에 이 연산을 위해서 필요한 메모리 용량 거의 n에만 비례한다. 따라서 이는 웬만큼 거대한 자료가 아니고서는 충분히 감당할 수 있는 수준이다.

현재 성긴 행렬의 연산에 대한 연구는 많이 진행되어 있는 상태이고 사용 가능한 오픈소스 라이브러리 또한 다양하다. Golub과 Van Loan (2012)는 행렬 연산에 대한 내용을 다루는 책인데 여기에 거대 성긴 행렬에서의 여러 이론과 알고리즘이 자세히 나와 있다. 현재 구현되어 있는 라이브러리 중 가장 대표적인 것으로는 포트란 언어로 작성된 SPARSKIT (Saad, 1990)이 있다. 본 연구에서 사용한 프로그램은 통계 패키지 R의 Matrix라는 라이브러리이다. Matrix 라이브러리에서는 성긴 행렬을 저장하는 두 가지 방식을 제공한다. 이들 방식을 사용하기 위해서는 행렬 입력 형식에 약간의 변화를 주면 된다. 각각의 방법은 행렬의 성긴 성질을 사용하여 값이 0인 아닌 성분만을 저장하기 때문에 메모리 용량을 대폭 절약할 수 있지만 그 비용으로 행렬의 각 성분에 접근하는 시간이 일반적인 2차원 행렬 구조보다 약간 더 소모된다. 하지만, 본 연구에서 다루는 인용 그래프의 인접행렬은 대부분의 성분이 0이기 때문에 그 비용은 매우 미미하다. 메모리 구조에 익숙치 않은 보통 사용자에게 가장 친숙한 방법은 0이 아닌 성분의 행의 위치 i, 열의 위치 j, 행렬의 값 x, 이 세 개의 값을 (i, j, x) 형태로 저장하는 것이다. 이는 Matrix 라이브러리의 sparseMatrix 함수를 사용하면 된다.

행렬의 덧셈 및 곱셈뿐만 아니라 뉴먼 커널을 조율하기 위해서 조율모수 γ의 범위를 알아야 한다. 이 범위는 공통피인용행렬 A^TA 또는 공통인용행렬 AA^T의 가장 큰 고유값에 의존하기 때문에 A^TA의 고유값을 구해야 한다. 성긴 거대 행렬에서의 고유값을 구하는 빠른 알고리즘은 Golub과 Van Loan (2012)의 10장에 자세히 나와 있다. R에서의 구현은 igraph 라이브러리에 있는 arpack 함수를 사용해야 한다. R의 arpack 함수는 거대 성긴 행렬에서 고유값을 구하는 ARPACK (Lehoucq 등, 1998) 패키지의 일부를 R 상에서 구현한 것이다.

5. 특허 자료 분석

5.1. 자료 설명

이번 절에서 분석에 사용하는 자료는 미국 특허청에 등록된 특허 약 400만 건 중 일부이다. 각각의 특허는 평균적으로 대략 23개의 참고문헌을 갖고 있으며 이 중에는 미국 등록 특허인 것도 있으며 다른 나라의 특허나 논문 등 미국 등록 특허가 아닌 것도 있다. 참고문헌 중 미국 등록 특허는 평균 16개 정도이며 이들만으로 전체 특허에 대한 인용 그래프를 구성하였다. 이 자료를 분석하기 위해 사용한 컴퓨터는 Intel(R) Core(TM) i7-3770 3.40GHz 프로세서에 16Gbyte의 메모리 용량을 갖고 있으며 운영체제로는 Window 7 64bit가 설치되어 있다. 특허 데이터는 IPC (International Patent Classification) 기준에 따라 여러 계층의 기술 분야로 나눌 수 있는데 가장 큰 분류 기준은 IPC1으로 총 8가지 종류의 분야가 있다. 대략 50만 건의 특허는 IPC 체계에 의해 분류되지 않는 특허인데 이들은 분석 대상에서 제

외하였다. Table 5.1에 IPC1에 의한 특허 분류 체계가 정리되어 있다. 본 연구에서는 특정 특허와 연관성이 깊은 중요 특허를 추천하는 것이 주요 목적이기 때문에 IPC1이 동일한 특허만을 대상으로 분석을 시행하였다. 물론 400만 특허 전체를 사용하는 것이 가장 좋은 방법이지만, 이럴 경우 아무리 성긴 행렬의 성질을 사용하더라도 일반 PC의 메모리 용량으로는 버겁기 때문에 하둡과 같은 특별한 파일시스템 및 계산 방식을 사용해야 한다. 이는 추후 과제로 남겨두기로 하고 본 연구에서는 일반 PC에서 쉽게 다룰 수 있을 정도의 용량만을 처리하기로 한다. 그리고 전체 인용의 60% 정도가 동일한 IPC1 기술분야 내에서 발생하기 때문에 인용 정보를 통해 다른 기술분야에서 추천할 만한 문서는 그리 많지 않다. 따라서, 같은 기술분야 내에서의 인용정보만을 사용한다고 하더라도 본 연구의 목적에 크게 어긋나지 않는다고 할 수 있다.

Table 5.1 Patent classification based on IPC1: N represents the number of patents in each group, R1 is the mean of references for each patents and R2 is the mean of references which count citations only in the same group.

IPC1	Technique fields	N	R1	R2
A	human necessities	525,735	18.9	14.8
B	performing operations, transporting	678,211	16.5	7.0
C	chemistry, metallurgy	384,432	12.9	6.5
D	textiles, paper	36,736	15.1	5.1
E	fixed constructions	103,490	20.2	9.2
F	machanical engineering, lighting, heating, weapons, blasting	307,132	13.6	6.3
G	physics	961,522	15.3	11.4
H	electricity	795,817	13.5	10.5
	etc	488,319	12.0	
		4,281,394	15.5	9.9

Table 5.1을 보면 기술 분야 별로 그래프 노드의 수와 평균 모서리의 수를 알 수 있다. IPC1 기준으로 G, H 분야는 다른 분야에 비해 상대적으로 큰 분야이기 때문에 동일 분야 내에서의 인용 횟수가 다른 분야에 비해 상대적으로 많다. 반면 D와 같은 소규모 기술 분야에서는 같은 기술 분야 내에서의 인용에 비해 다른 분야의 특허를 많이 인용하는 것을 알 수 있다. A 기술군의 경우 인용 수가 크고 동시에 같은 기술 분야 내에서의 인용 비율 또한 크다.

5.2. 분석 결과

우선 특허의 숫자가 가장 작은 D 기술분야의 전체 인용 데이터를 살펴보자. D 기술분야의 공통피인용행렬 A^TA는 $36,736 \times 36,736$ 대칭행렬로 약 10억개의 성분을 가진다. 뉴먼 커널을 근사하기 위해서는 $(A^TA)^k$를 계산해야 한다. 4절에서 언급했듯이 k가 커질수록 0이 아닌 성분의 수가 매우 빠르게 늘어나기 때문에 금방 메모리의 한계치에 부딪친다. 본 연구에 사용한 컴퓨터에서는 $(A^TA)^5$까지 계산이 가능하며 이를 저장하기 위해 약 10Gbyte의 메모리 용량이 소요되었다. 이에 대한 간략한 정보는 Table 5.2에 요약되어 있다. R에서 데이터를 처리하는 구조로 인해 k가 4 이상만 되어도 각 성분에 접근하는 시간이 매우 길기 때문에 표의 3열과 같은 평균을 구하는 것도 쉽지 않다. k가 4보다 클 때 Table 5.2의 3열 값은 샘플링을 통해 추정한 것이다. igraph 라이브러리의 arpack 함수를 사용하여 전체 인용 자료로부터 얻은 공통피인용행렬 A^TA의 스펙트럴 반경은 대략 2000 정도이다. 기준시점 T 이전의 자료가 전체의 약 70%가 되도록 잡고 T 시점 이전의 자료만으로 공통피인용행렬을 만들어 스펙트럴 반경을 구하면 대략 1400 정도가 나온다.

Table 5.2 Summary of $(A^TA)^k$: The second column is the numbers of nonzero elements in $(A^TA)^k$ and the last column is mean values of nonzero elements in IPC1 group D. Both numbers are growing very fast as k increases.

$(A^TA)^k$	number of nonzero elements	mean of nonzero elements
A^TA	1,101,681	2
$(A^TA)^2$	16,246,377	113
$(A^TA)^3$	86,413,153	22252
$(A^TA)^4$	267,207,672	8×10^6
$(A^TA)^5$	469,637,726	2×10^{11}

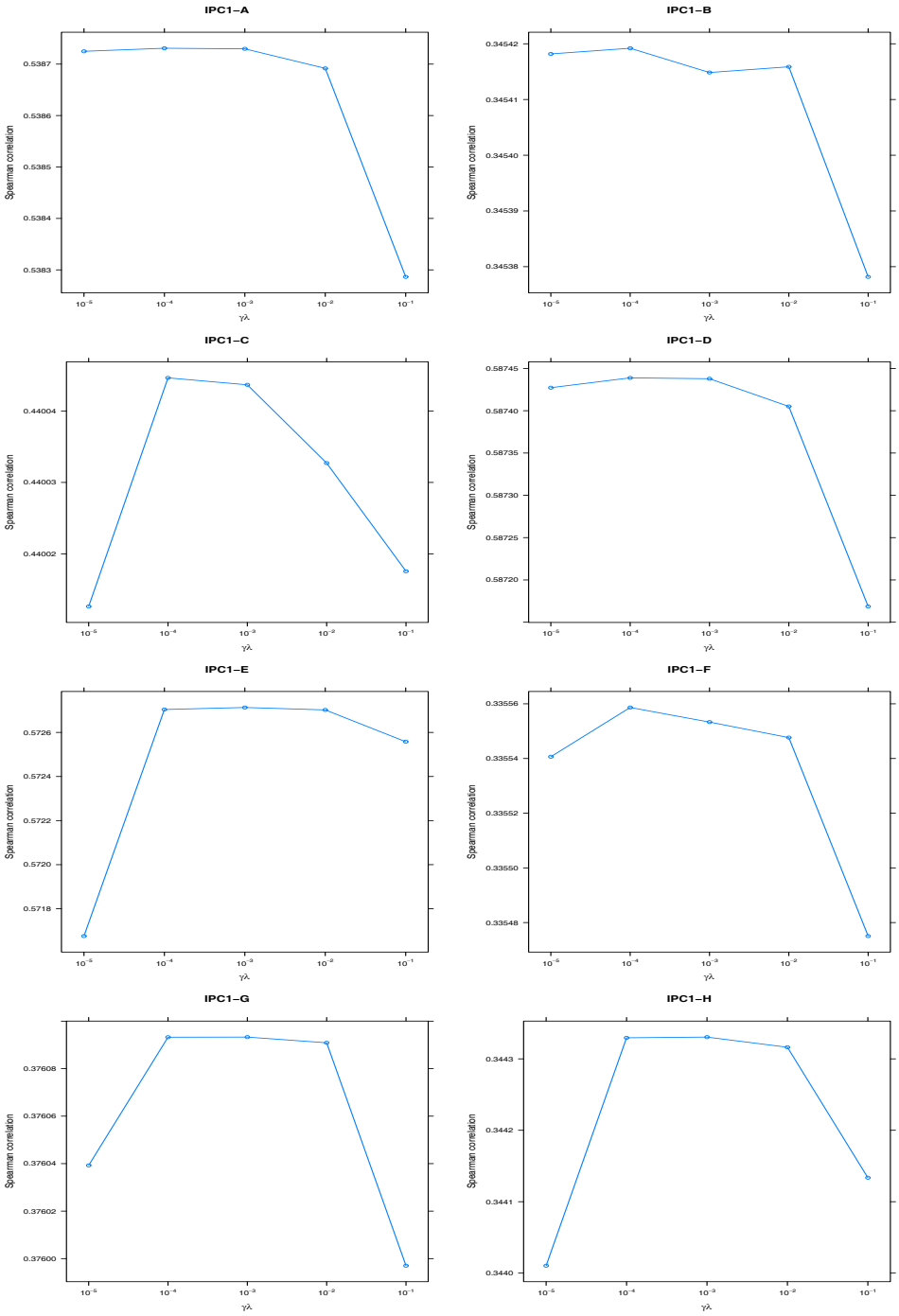

Figure 5.1 The results of tuned γ in each IPC1 group: The values of optimal $\gamma\lambda$ vary between 10^{-4} and 10^{-2}, so the association measure (co-citation) is more important than global importance measure (HITS) in our examples.

Figure 5.1은 3절에서 설명한 내용을 바탕으로 γ를 조율한 것이다. x축은 $\gamma\lambda$의 값을 나타낸다. 이 값이 0에 가까울수록 뉴먼 커널은 공통피인용행렬에 가까워지고 1에 가까울수록 HITS에 가까워진다. 결과를 살펴보면 대부분 $\gamma\lambda$ 값이 10^{-4}에서 10^{-2} 사이에서 결정되는 것을 알 수 있다. 절대적인 값은 아니지만 이를 통해 노드 간의 연관성 정도가 조금 더 중요하다고 결론지을 수 있다. 이는 같은 기술분야 내에서도 기술의 종류가 워낙 다양하기 때문에 굉장히 중요한 특허일지라도 다른 세부적인 기술과는 특별히 관계가 없는 경우가 많은 것으로 해석할 수 있다.

6. 결론 및 제언

본 논문에서는 문서의 인용 정보만으로 특정 문서와의 연관성이 높은 문서 중 중요한 문서를 추천하는 방법론을 제안하였다. 본 연구에서 사용 방법은 뉴먼 커널을 활용하는 것으로 뉴먼 커널의 조율모수 γ를 추정하여 연관성과 중요성의 영향 정도를 결정하는 것이 핵심이다. 본 논문에서 사용한 방법은 미래 자료에 대한 예측 오차를 최소화함으로써 중요도와 연관성의 반영 정도를 결정하는 것이다. 이를 실제 미국에 등록된 특허 문서에 적용함으로써 본 논문에서 제안한 방법이 중요도와 연관성의 적절한 중간 지점에서 예측 오차를 최소화하는 것을 보였다.

이러한 방법론을 적용함에 있어서 가장 큰 문제는 거대한 인용 자료로부터 나오는 그래프의 인접 행렬을 저장하고 계산하는 것이다. 본 연구에서는 실제 자료 분석에 있어서 일반 PC 수준에서 계산 가능한 정도로 자료를 분할하였다. 그 결과 5절에서는 특정 특허에 대하여 IPC1 기술분야 기준으로 같은 기술분야 내의 특허만을 추천하였다. 하지만 기술 분야가 다르더라도 최근 여러 분야의 기술들이 융합되어 새로운 기술이 개발되는 점을 감안한다면 다른 기술 분야의 특허를 전혀 추천하지 않는 것은 바람직하지 않다. 이를 극복하기 위해서는 전체 400만 건의 자료를 한꺼번에 계산할 수 있는 환경이 필요하다. 뉴먼 커널은 행렬과 벡터의 곱으로만 계산할 수 있기 때문에 최근 대두되고 있는 하둡과 같은 분산 시스템이 이런 문제에 대한 매우 좋은 해결책이다.

특허를 출원할 때에는 반드시 해당 특허가 속할 기술 분야를 정해야 한다. 특허 기술 분야는 5절에서 언급했듯이 IPC 기준에 의해 나뉘는데 IPC1 - IPC2 - IPC3 - $\cdots$ 등에 따라 여러 계층으로 구분된다. 보통 특허를 출원하는 사람은 이를 결정하는데 어려움을 겪기 때문에 많은 특허 전문가들이 이를 자동화 또는 완전 자동화가 아니더라도 해당 특허가 속할만한 기술 분야를 자동으로 추천해주길 원한다. 현재 키워드를 활용한 자동화 시스템 개발은 상당한 어려움을 겪고 있는데 여기에 인용 정보를 활용한다면 매우 좋은 추천 성능을 보일 것으로 예상하고 있다. 5절에서 확인했듯이 실제 문서의 인용 중 큰 비율이 같은 기술 분야 내에서 일어나기 때문이다. 따라서, 인용 정보를 활용하여 기술군 결정을 부분적으로 자동화할 수 있는 통계 모형을 개발하면 그 수요가 매우 클 것으로 예상한다.

References

Blei, D. M. and Lafferty, J. D. (2007) A correlated topic model of science. *The Annals of Applied Statistics*, **1**, 17-35.

Blei, D. M., NG, A. Y. and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual (web) search engine. *Computer Network and ISDN Systems*, **30**, 107-117.

Cook, D. J. and Holder, L. B. (2006). *Mining graph data*, John Wiley & Sons, New Jersey.

Garfield, E. and Merton, R. K. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*, Wiley, New York.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, Johns Hopkins University Press, Baltimore.

He, Q., Pei, J., Kifer, D., Mitra, P. and Giles, C. L. (2010). Context-aware citation recommendation. *Proceedings of the 19th International Conference on World Wide Web*, 421-430.

Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, **22**, **22**, 89-115.

Jannach, D., Zanker, M., Felfernig, A. and Friedrich, G. (2010). *Recommender systems: An introduction*, Cambridge University Press, New York.

Kandola, J., Shawe-Taylor, J. and Cristianini, N. (2003). Learning semantic similarity. In *Neural Information Processing Systems*, 673-680.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, **14**, 10-25.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, **46**, 604-632.

Kolaczyk E. D. (2009). *Statistical analysis of network data: Methods and models*, Springer, New York.

Lam, C. (2010). *Hadoop in action*, Manning Publications Company, Stamford.

Lehoucq, R. B., Sorensen, D. C. and Yang, C. (1998). *ARPACK users' guide: Solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*, **6**, Siam, Philadelphia.

Li, W. and McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd International Conference on Machine Learning*, 577-584.

Liben-Nowell, D. and Kleinberg, J. (2007). The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, **58**, 1019-1031.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., Konstan, J. A. and Riedl, J. (2002). On the recommending of citations for research papers. *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, 116-125.

Page, L. and Brin S. (1999). *The PageRank citation ranking: Bringing order to the web*, Stanford InfoLab, California.

Shimbo, M. and Ito, T. (2006). *Kernels as link analysis measures*, John Wiley & Sons, New Jersey, 283-310.

Saad, Y. (1990). *SPARSKIT: A basic toolkit for sparse matrix computations*, Research Institute for Advanced Computer Science, NASA Ames Research Center Moffet Field, CA.

Sanders, J. and Kandrot, E. (2010). *CUDA by example: An introduction to general-purpose GPU programming*, Addison-Wesley Professional, Boston.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, **24**, 265-269.

Strohman, T., Croft, W. and Jensen, D. (2007). Recommending citations for academic papers. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 705-706.

Tang, J. and Zhang, J. (2009). A discriminative approach to topic-based citation recommendation. *Advances in Knowledge Discovery and Data Mining*, 572-579.

Teh, Y. W., Jordan M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, **101**, 1566-1581.

Wei, X. and Croft W. B. (2006). LDA-based document models for ad-hoc retrieval. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 178-185.

White, S. and Smyth P. (2003). Algorithms for estimating relative importance in networks. *Proceedings of the KDD'03*, 266-275.

Journal of the Korean Data & Information Science Society
2013, **24**(5), 999–1011

http://dx.doi.org/10.7465/jkdi.2013.24.5.999
한국데이터정보과학회지

Documents recommendation using large citation data[†]

Minwoo Chae[1] · Minsoo Kang[2] · Yongdai Kim[3]

[13]Department of Statistics, Seoul National University
[2]Kwang Gae To Laboratory

Received 30 June 2013, revised 23 July 2013, accepted 5 August 2013

Abstract

In this research, we propose a document recommendation method which can find documents that are relatively important to a specific document based on citation information. The key idea is parameter tuning in the Neumann kernal which is an intermediate between a measure of importance (HITS) and of relatedness (co-citation). Our method properly selects the tuning parameter γ in the Neumann kernal minimizing the prediction error in future citation. We also discuss some comutational issues needed for analysing large citation data. Finally, results of analyzing patents data from the US Patent Office are given.

Keywords: Big data, citation data analysis, Neumann kernel, recommendation, sparse matrix computation.

[†] This work was carried out with the support of "Cooperative Research Program for Agriculture Science & Technology Development (Project No. PJ907160)" Rural Development Administration, Republic of Korea.

[1] Corresponding author: Ph. D. candidate, Department of Statistics, Seoul National University, Seoul 151-747, Korea. E-mail: demian87@snu.ac.kr
[2] Representative director, Kwang Gae To Laboratory, Seoul 135-921, Korea.
[3] Professor, Department of Statistics, Seoul National University, Seoul 151-747, Korea.