시간적 효과가 고려된 페이지랭크 함수를 이용한 핵심특허 탐색방법

이혁재* · 양혜영**

Ⅰ. 서론

구글의 페이지랭크(RageRank) 알고리즘(Brin and Page, 1998)은 17년 전 등장한 이후 지금까지 구글 검색엔진의 핵심이다. 페이지랭크는 웹 문서 간 연결(하이퍼링크)을 이용하여 웹 문서의 중요도를 평가하는 알고리즘으로, 웹 문서에 연결된 하이퍼링크의 수에 따라 정규화를 하고, 어떤 웹 문서로부터 하이퍼링크를 받았는지에 따라 중요도가 측정되는 방식이다. 웹 문서의 중요도를 평가하는 개념은 과학기술 문헌, 즉 논문이나 특허 문헌의 중요도를 평가하는 개념과 일맥상통한다. 페이지랭크 알고리즘은 과학기술 문헌의 중요도를 평가할 때 전통적으로 사용되어 온 피인용수의 단점을 보완할 수 있을 것이라는 기대감에 의해 논문 또는 특허 등 과학기술 문헌에 적용되어 연구되었다.

논문이나 특허 문헌의 피인용수는 문헌의 중요도를 측정하는 전통적인 방법으로 널리 활용되고 있다. 그런데 피인용수는 논문이나 특허 문헌의 중요도를 측정함에 있어 몇 가지 한계를 가진다. 첫째, 분야마다 피인용수의 분포가 다르므로 분야별 편차를 고려하는 방법이 필요하다. 둘째, 모든 개별 인용은 균등하게 1로 측정되는데, 모든 인용이 동일한 과학기술적 가치를 지니는 것이 아니므로 문헌의 중요도를 정확히 측정한다고 보기 어렵다. 셋째, 오래된 문헌일수록 피인용수가 클 수밖에 없어 최근 문헌에 대한 평가가 공정하지 못하게 된다.

그동안 피인용수의 한계를 보완하기 위해 여러 연구가 이루어졌다. 첫 번째 한계를 보완하기 위한 방법은 van Raan(2005) 등의 연구에서 찾아볼 수 있다. 페이지랭크 알고리즘은 피인용수의 대표적인 세 가지 한계 중 두 번째를 보완하기 위하여 적용되었다. 선행연구 조사 결과, 논문 문헌에 페이지랭크 알고리즘을 적용한 경우가 다수 존재(Chen 등, 2006, Sayyadi and Getoor, 2009)하며, 특허 문헌에 적용된 경우(Lukach and Lukach, 2007)도 있다. 페이지랭크 알고리즘도 피인용수의 세 번째 한계와 마찬가지로 시간적 효과를 고려하지 않으므로 오래된 웹 문서일수록 더 많은 하이퍼링크를 받게 된다는 문제가 존재한다. 이를 극복하기 위해 페이지랭크에 시간적 효과를 고려한 함수를 제안한 사례(Yu et al., 2005, Walker et al, 2007)가 존재한다. 그러나 특허에 대하여 시간적 효과가 고려된 페이지랭크 알고리즘을 적용한 사례는 아직 존재하지 않는다.

Yu 등(2005)과 Walker 등(2007)의 연구는 시간적 효과를 고려한 페이지랭크 함수를 제안하고 이를 논문 문헌에 대하여 적용한 결과를 제시하고 있다. 두 함수가 문헌의 중요성을 얼마나 정확하게 측정하는지에 대한 유효성을 증명하기는 어려우나, 구글의 페이지랭크 알고리즘에 대하여 시간적 효과를 고려하기 위한 방법을 제안한 것만으로도 의미가 있다고 볼 수 있다.

본 연구는 시간적 효과가 고려된 페이지랭크 함수를 특허 문헌에 적용한 연구이다. Yu 등(2005)과 Walker 등(2007)의 연구에서 제안한 시간적 효과를 고려한 페이지랭크 함수를 이용하여 KISTI가 개발한 COMPAS(Kim and Lee, 2015)에 탑재된 미국등록특허에 적용하였다. 이때 각 함수의 시간관련 파라미터에

* 이혁재, 한국과학기술정보연구원 책임연구원, 02-3299-6059, hlee@kisti.re.kr
** 양혜영, 한국과학기술정보연구원 선임연구원, 02-3299-6069, hyyang@kisti.re.kr, 교신저자

변화를 주어 계산하였고, 가장 적절한 파라미터와 함수를 선택하여 핵심특허를 탐색하는 방법으로 제안하고자 한다.

Ⅱ. 시간적 효과를 고려한 페이지랭크 함수, TimedPageRank와 CiteRank

1. TimedPageRank

Yu 등(2005)은 페이지랭크 함수[1]에 감쇄율(DecayRate)을 포함하는 가중치를 적용하고 이를 TimedPageRank 함수로 명명하여 제안하였다. 즉, 오래된 피인용의 가중치를 적게 두고, 최근 피인용의 가중치를 크게 두는 것이다. 가중치는

$$w_i = DecayRate^{(y-t_i)/12}$$

으로 정의되었다. 여기서 $(y-t_i)$는 현재 시점(y)과 논문 i의 출판 시점(t_i)과의 개월 차(the time gap in months)를 뜻한다. 감쇄율 $DecayRate$은 임의의 파라미터로 0과 1 사이의 상수이다. $DecayRate$이 1에 가까워질수록 가중치는 시간에 따라 천천히 감소하고, 1이 되면 페이지랭크 함수와 동일해지는 알고리즘이다. Yu 등은 $DecayRate$을 0.2, 0.3, 0.5, 0.7, 0.8, 1.0 등으로 설정하여 민감도를 측정하였고, 0.3에서 0.7 사이의 값이 적절하다는 의견을 제시하였다. TimedPageRank 함수가 잘 작동하는지를 확인하기 위하여 고피인용 논문과 잘 일치하는지 검토하였다. 연구에 활용한 논문은 2003년도 고에너지입자물리 논문 아카이브인 KDD CUP 2003이다.

2. CiteRank

Walker 등(2007)은 페이지랭크 함수에 감쇄시간(decay time)을 적용하고 이를 CiteRank 함수로 명명하여 제안하였다. 이들의 논리에 의하면 연구자들은 연구주제를 탐색할 때 임의의 논문을 선택하고 그 논문의 전후방 인용관계를 이용해 탐색을 진행하게 된다. 이 과정에서 특정 논문이 선택될 확률 ρ_i는 최신 논문일수록 높으며 다음과 같은 함수로 주어진다.

$$\rho_i = e^{-age_i/\tau}$$

여기서 age_i는 논문 i의 나이, τ는 임의의 파라미터로 해당 논문 인용 네트워크의 특성을 나타내는 감쇄시간 상수이다. 그리고 연구자가 선택한 논문을 읽고 만족할 확률, 즉 더 이상의 참고문헌을 탐색하지 않을 확률을 α로, 선택한 논문을 읽고 만족하지 못하여 다른 참고문헌을 탐색할 확률을 $1-\alpha$로 정의하였다. 이들은 연구자가 연구주제 탐색 시 인용 네트워크를 타고 움직이다가 논문 i로 흘러들어갈 트래픽(traffic)을 파라미터 τ와 α의 함수로 정의하였고, 두 파라미터의 상관관계 계수가 가장 커지는 값을 찾는 방식으로 τ와 α의

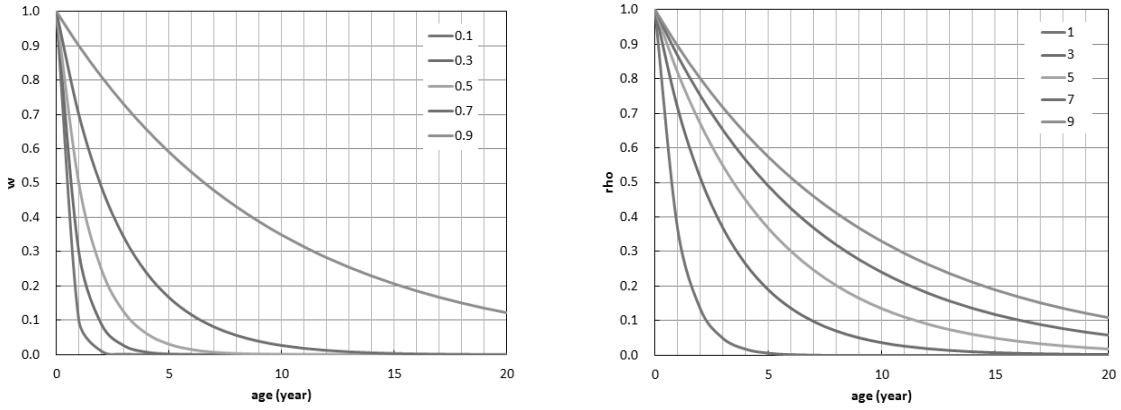
---

1) $PR(A) = \frac{(1-d)}{N} + d(\frac{PR(T1)}{C(T1)} + ... + \frac{PR(Tn)}{C(Tn)})$, $PR(A)$는 웹문서 A의 페이지랭크 값, N은 웹문서의 개수, d는 감쇄상수, $C(Tn)$은 웹문서 Tn에서 출발하는 하이퍼링크의 개수를 의미

값을 논문 인용 네트워크의 특성 파라미터로 결정하였다. 이들은 The Physical Review 저널의 인용 네트워크(Physrev)와 고에너지물리논문(The high-energy physics archive)의 인용 네트워크(Hep-th)에 대하여 실험했다. Pearson 상관계수를 이용할 경우 각각의 논문 인용 네트워크에 대하여 τ는 2.6년(Physrev), 1년(hep-th)으로 측정되었다. 그러나 Spearman 상관계수를 이용할 경우 τ는 8년(Physrev), 1.6년(hep-th)으로 측정되어 측정 방식에 따라 큰 차이가 있는 것이 이들 연구의 한계이다.

3. 특허 문헌에 대한 적용 가능성 검토

시간적 효과를 고려한 페이지랭크 함수, 즉 TimedPageRank와 CiteRank 함수를 특허에 적용할 수 있는지 가능성을 검토해보았다. 우선 TimedPageRank와 CiteRank 함수의 시간고려 함수 항인 $w_i = DecayRate^{(y-t_i)/12}$과 $\rho_i = e^{-age_i/\tau}$가 시간 파라미터 $DecayRate$과 τ에 따라 어떤 양상을 나타내는지 그려보았다.



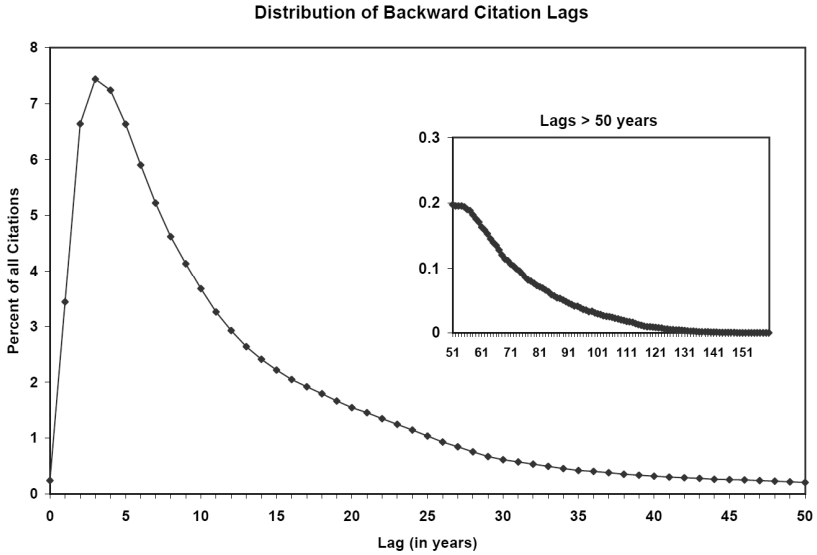
(a) $DecayRate$이 0.1, 0.3, 0.5, 0.7, 0.9일 경우 문서 나이에 따른 선택 확률

(b) τ가 1, 3, 5, 7, 9일 경우 문서 나이에 따른 선택 확률

(그림 1) TimedPageRank와 CiteRank 함수의 시간 파라미터 $DecayRate$과 τ에 따른 확률분포함수 w_i와 ρ_i

(그림 1)은 (a) TimedPageRank 함수의 $DecayRate$ 값을 0.1, 0.3, 0.5, 0.7, 0.9로, (b) CiteRank 함수의 τ를 1, 3, 5, 7, 9로 변화시켜가면서 그린 w_i와 ρ_i이다. x축은 문서의 나이를 의미한다. 이 그림은 오래된 문서일수록, 즉 x 값이 클수록 그 문서가 연구자에 의해 선택될 확률이 낮아짐을 의미한다. TimedPageRank 함수의 경우, 시간 파라미터 값이 작을수록 w_i가 매우 급격하게 감쇄한다. 예를 들어 $DecayRate$ 값이 0.1인 경우, 1년만 지나도 확률이 0.1 이하로 떨어진다. 출판된 지 1년 이상 지난 문서가 선택될 확률은 불과 10% 이내라는 의미이다. 반면 최근 1년 이내 출판된 논문이 선택될 확률은 90% 이상이다. 문서 나이가 2년만 되어도 선택될 확률은 거의 0에 가까워진다. $DecayRate$ 값이 0.3인 경우, 선택 확률이 10% 이하로 낮아지는 문서 나이는 2년 정도이고, 문서 나이가 4년만 되어도 선택될 확률은 거의 0에 근접한다. $DecayRate$ 값이 0.9인 경우, 확률분포함수의 감쇄는 매우 더뎌져 문서 나이가 10년이 되더라도 선택될 확률은 여전히 30% 이상으로 나타난다. CiteRank 함수의 경우, 시간 파라미터 값이 작을수록 오래된 문헌이 인용될 확률이 급격히 낮아지는 경향은 TimedPageRank와 동일하지만 상대적으로 완만한 감소를 보인다. 시간 파라미터 τ가 1일 경우, 문서가 선택될 확률이 10% 이하로 낮아지는 문서의 나이는 대략 2년이다. TimedPageRank 함수의 경우 문서가 출판된 지 1년만 지나도 선택될 확률이 10% 이하로 낮아지는 것에 비하면 덜 급격한 변화이고,

TimedPageRank 함수에서 $DecayRate$ 값이 0.3인 경우와 유사한 경향이다. 문서 나이가 5년이 되면 선택될 확률은 0에 가까워진다. τ가 3일 경우, 문서 나이가 5년이면 선택될 확률이 약 20% 정도로 낮아지고, 선택될 확률이 0에 근접하는 문서 나이는 14년 정도이다. 두 함수 모두, 특히 TimedPageRank 함수의 경우 문서의 나이가 일정 수준 이상인 경우 선택될 확률이 0에 근접한다. 이는 특허 문서에 과거 특허의 인용이 거의 발생하지 않음을 의미하나 이는 현실과 다르다.



(그림 2) 1975년 이후 등록된 미국특허에 인용된 특허의 나이(인용 발생 시점에 대한)
(그림 출처) Hall, B. H. et al.(2001)

Hall 등(2001)의 연구 결과에 의하면, 1975년 이후 등록된 미국특허에 의해 인용된 특허의 나이, 즉 인용한 특허와 인용된 특허의 시간 차이(lag)의 빈도는 (그림 2)와 같다. 3년 전 특허에 대한 인용이 가장 많고, 그보다 더 오래된 특허에 대한 인용 비율은 점진적으로 감소한다. 그러나 감쇄하는 정도는 매우 완만하다. 심지어 등록된 지 50년이나 지난 특허에 대한 인용비율은 동일 연도에 등록된 특허를 인용하는 비율만큼 높다. (그림 2)의 시간 차이가 50년을 초과하는 경우에 대한 세부 그림을 보면 그 비중은 매우 낮지만 100년이 지난 특허에 대한 인용도 존재함을 알 수 있다. 따라서 시간적 효과를 고려한 페이지랭크 함수를 특허 문서에 적용할 때 시간 파라미터를 어떻게 설정할 것인지에 대하여 매우 신중하게 접근해야 한다고 판단된다.

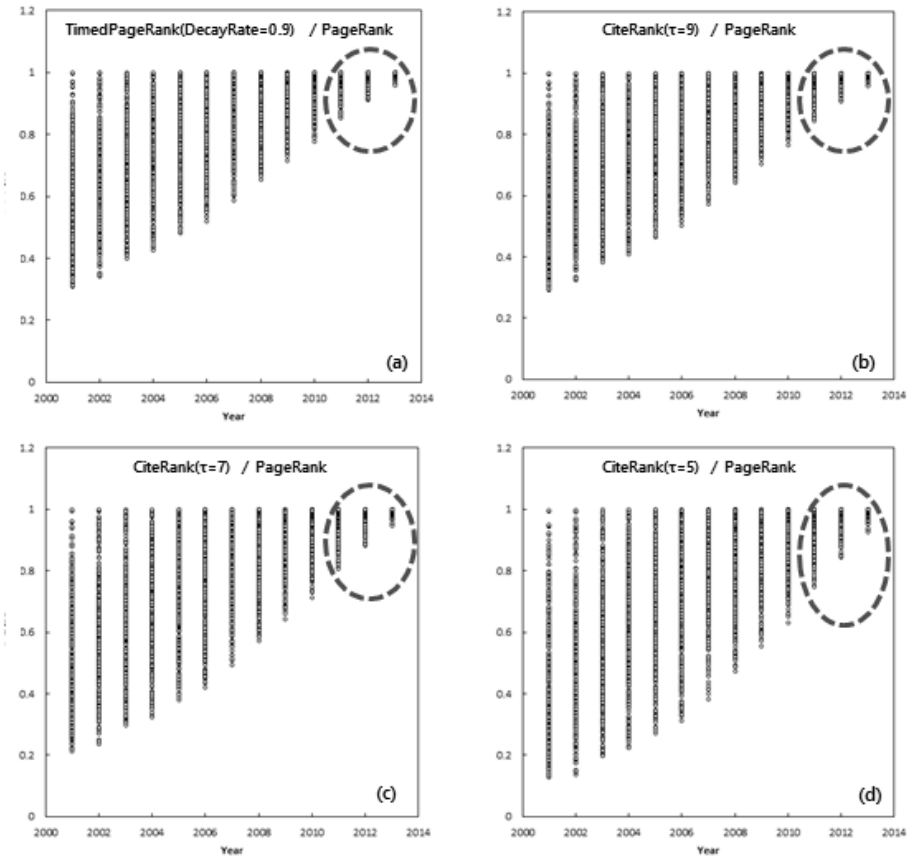
III. 분석 방법 및 결과

1. 분석 방법

본 연구에서는 특허의 중요도를 측정하기 위한 방법으로 시간적 효과를 고려한 페이지랭크 함수를 적용하였다. 2장에서 설명한 TimedPageRank 함수와 CiteRank 함수의 시간 파라미터를 달리 설정하여 미국등록특허의 인용 네트워크에 적용하였다. 분석에 사용한 데이터는 2001년부터 2013년까지의 모든 미국등록특허이다. 결과를 확인하기 위하여 동일 기간의 특허 중 출원인이 Toyota인 미국등록특허 5,470개에 대해 TimedPageRank 값, CiteRank 값, 페이지랭크 값, 피인용수 등을 분석하였다.

2. 분석 결과

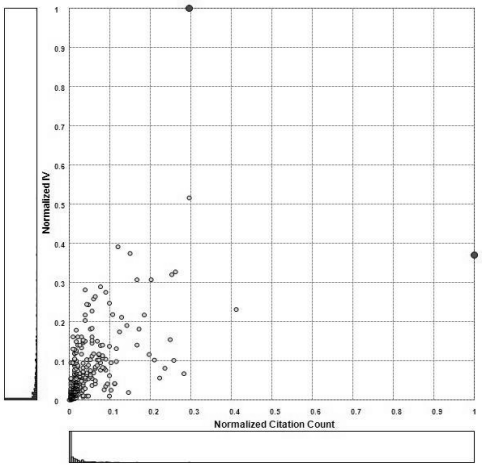
Toyota의 미국등록특허 5,470개에 대한 TimedPageRank 값, CiteRank 값, 페이지랭크 값, 피인용수를 비교하였다. TimedPageRank 값과 CiteRank 값은 시간 파라미터 $DecayRate$와 τ가 각각 0.1, 0.3, 0.5, 0.7, 0.9인 경우와 1, 3, 5, 7, 9인 경우에 대하여 계산하였다. 적절한 시간 파라미터의 종류와 값을 판단하기 위한 특별한 기준이 존재하지 않으므로 본 연구에서는 다양한 관점에서 결과를 논의하였다. 우선 2장 3절에서 서술한 바와 같이, 나이가 5년 이상인 문서가 인용될 확률이 20% 이하로 낮아지는 경우, 즉 $DecayRate$이 0.7 이하인 경우와 τ가 3이하인 경우는 모두 기각되는 것이 타당하다. 따라서 실험한 시간 파라미터 값 중에서 $DecayRate$가 0.9인 경우와 τ가 5, 7, 9인 네 가지 경우가 남는다. 한편 $DecayRate$이 0.9인 경우와 τ가 9인 경우인 (그림 3)의 (a)와 (b)에 의하면, 최근 특허문서에 대한 TimedPageRank 값과 CiteRank 값은 페이지랭크 값과 크게 차이 나지 않는다. 따라서 페이지랭크의 한계를 보완하려는 목적에 크게 부합하다고 볼 수 없다. (그림 3)의 (c)와 (d)는 τ가 7과 5인 경우에 대한 CiteRank 값과 페이지랭크 값의 비율이다. 최근 특허문서의 시간적 효과 보정 목적을 고려할 때 (d)의 경우가 가장 합당한 결과로 보인다. 다만 오래된 문서에 대한 보정효과도 커지는 단점이 존재한다. 이를 극복하기 위하여 핵심특허 탐색 시 (그림 4)와 같이 피인용수와 CiteRank(τ=5) 값을 2차원 평면에서 동시에 나타내는 방안을 고안하였다.



(그림 3) 연도별 특허의 TimedPageRank 값(a) 또는 CiteRank(b, c, d) 값과 페이지랭크 값의 비율. 빨간 점선 원은 최근 3년간 특허문서에 대한 결과를 의미.

Ⅳ. 결론 및 시사점

본 연구는 시간적 효과를 고려한 페이지랭크 함수가 특허의 중요도를 판단하는 지표로 활용 가능한지를 탐색한 것이다. 문헌조사 결과 TimedPageRank 함수와 CiteRank 함수를 후보 함수로 검토하였고, KISTI가 개발한 COMPAS 시스템에 구축된 미국등록특허에 대하여 시간 파라미터별로 측정하였다. Toyota 특허에 대해 TimedPageRank 값, CiteRank 값, 페이지랭크 값을 비교하였고 가장 적합한 시간 파라미터를 결정하였다. 본 연구결과는 COMPAS 시스템 내 '핵심특허 분석모델'로 구현되어 있으며, 앞서 설명한 바와 같이 특허의 피인용수 지표와 함께 2차원 평면상에 구현되어 사용자가 적절히 핵심특허를 탐색할 수 있도록 한다. 시간적 효과가 고려된 페이지랭크 함수가 특허데이터에 적용된 최초의 사례연구이므로 의미 있는 연구라 할 수 있으며, 정교화를 위한 후속 연구를 거쳐 핵심특허 탐색 시 도움이 될 것으로 기대된다.



(그림 4) COMPAS 시스템 내 핵심특허 분석모델 구현 모습(점 하나는 특허문서 하나를 의미하고, x축은 정규화된 피인용 수, y축은 CiteRank 함수를 이용해 구한 ImpactValue의 정규화된 값을 의미.)

참고문헌

Brin, S and Page, L. (1998), "*The anatomy of a large-scale hypertexual Web search engine*", Computer Networs and ISDN Systems, 30, 107-117.

van Raan, A. F. J., (2005), Measuring Science, Handbook of Quantitative Science and Technology Research, Chapter 1.

Yu, P., Li, X., and Liu, B., (2005), "*Adding the Temporal Dimension to Search : A Case Study in Publication Search*", Web Intelligence, 2005. Proceedings of The 2005 IEEE/WIC/ACM International Conference, 543-549.

Chen, P. et al., (2006), "*Finding scientific gems with Google's PageRank algorithm*", Journal of Informetrics, 1, 8-15.

Walker, D. et al., (2007), "*Ranking scientific publications using a model of network traffic*", Journal of Statistical Mechanics: Theory and Experiment, Volume 2007.

Sayyadi, H. and Getoor, L. (2009), "*FutureRank: Ranking Scientific Articles by Predicting their Future PageRank*", Proceedings of the 2009 SIAM International Conference on Data Mining.

Lukach, R. and Lukach, M., (2007), "*Ranking USPTO Patent Documents by Importance Using Random Surfur Method (PageRank)*", Social Science Research Network, Available at SSRN: http://ssrn.com/abstract=996595.

Hall, B. H. et al., (2001), "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools", Working Paper 8498, National Bureau of Economic Research, http://www.nber.org/papers/w8498.

Kim, S. Y. and Lee, H. J, (2015), "*COMPAS – Competitive Analysis Service for Informed Decision-Making*", 2015 춘계 기술혁신학회.