# 특허인용 예측모형 구축에 관한 연구*

## A Study on Developing a Prediction Model of Patent Citation Counts

유재복(Jae-Bok Yoo)**
정영미(Young-Mee Chung)***

### 초 록

이 연구에서는 특허의 인용에 영향을 미치는 주요 변수들을 토대로 특허의 피인용횟수를 예측하기 위한 모형을 제시하였다. 이를 위해 미국특허를 대상으로 5개 주제분야에 걸쳐 특허의 피인용횟수와 일정 수준 이상의 상관관계, 즉 5% 이상의 설명력을 갖는 것으로 밝혀진 페이지 수, 청구항 수, 참고문헌 평균 피인용횟수, 서지결합도, 문헌간유사도 등 5개 변수들을 토대로 다중회귀분석을 실시하였다. 연구결과에 따르면, 제시된 5개 주제분야의 특허인용 예측모형의 설명력은 주제분야에 따라 58.3%~89.6%로 나타났으며, 예측변수로 사용된 5개의 독립변수 중 특허 피인용횟수에 가장 영향력이 높은 변수는 '문헌간 유사도'로 나타났다. 또한 이 연구에서 추정된 주제분야별 예측모형을 토대로 산출한 특허 피인용횟수에 대한 예측값과 실제값을 비교한 결과 이들 예측모형은 5개 주제분야에서 모두 적합한 것으로 나타났다.

### ABSTRACT

The purpose of this study is to develop a prediction model of patent citation counts based on major factors which affect patent citation. To this end, we performed multiple regression analysis between the patent citation counts and five explanatory variables such as the number of pages, the number of claims, the reference-average-citation rate, the strength of bibliographic coupling, and the document similarity proved as having 5% or more standardized variances($r^2$) with patent citation counts, with a test dataset of U.S. patents in five subject fields. As a result, our prediction models showed 58.3% to 89.6% predictability depending on subject fields and revealed the document similarity has the highest impact on citation counts among the five predictive variables in all the subject fields. The result of comparison between the predicted citation counts and the actual ones confirmed the usefulness of the citation prediction models built for each subject field.

키워드: 인용분석, 인용예측, 특허인용, 인용예측모형
citation analysis, citation prediction, patent citation, citation prediction models

---

* 이 논문은 박사학위 논문의 일부를 요약·정리한 것임.

** 한국원자력연구원 책임연구원(jbyoo@kaeri.re.kr) (제1저자)
*** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr) (공동저자)

■ 논문접수일자 : 2010년 11월 15일 ■ 최초심사일자 : 2010년 11월 20일 ■ 게재확정일자 : 2010년 11월 25일
■ 정보관리학회지, 27(4): 239-258, 2010. 〔DOI:10.3743/KOSIM.2010.27.4.239〕

# 1. 서 론

우리 사회가 산업사회에서 지식정보화 사회로 급속히 변화함에 따라 지식과 정보가 부의 수단에서 그 자체가 부의 원천이 되고 있으며, 토지나 자본 등 유형자산보다는 기술력, 브랜드, 디자인 등과 같은 무형의 지식재산이 국가나 기업의 경쟁력의 핵심요소로 등장하였다. 그 가운데 특허는 무형자산의 대표적인 지적 산물로 개인과 기업 및 국가의 기술수준과 혁신역량을 가늠하는 데 있어 중요한 객관적인 척도로 활용되고 있다. 그에 따라 특허기술의 가치평가가 크게 강조되고 있지만 특허기술의 가치를 객관적이고 정량적으로 측정하는 것은 결코 쉬운 일이 아니다.

특허기술의 가치를 평가하는 수단으로 여러 가지 정량적인 척도가 제안되어 있는데 지금까지는 주로 특허건수에 기반한 양적 지표가 주류를 이루어왔다. 그러나 특허건수에 기반한 지표는 개별 특허마다의 이질성, 즉 질적 측면을 전혀 고려하지 않았다는 점에서 문제가 되어 왔다(박종용 2008). 특허는 그 기술적 또는 경제적 가치가 상이하여 가치가 거의 없는 특허에서부터 엄청난 가치를 가진 특허에 이르기까지 다양하다. 그들 중 아주 적은 일부 특허만이 실제 활용되거나 후행특허에 영향을 준다(Ashton and Sen 1989). 따라서 개별 특허의 가치를 동일하다고 가정하는 단순 특허건수의 경우 이와 같은 특허가치의 이질성에 대한 고려가 원천적으로 배제되었다는 문제점이 제기될 수 있다.

이로 인해 특허의 양적인 측면은 물론 질적인 측면도 고려되어야 한다는 주장이 꾸준히 제기되었으며, 그 해결방안을 모색하는 일환으로 특허 인용에 기반한 다양한 연구가 이루어졌다. 특허문헌에 인용된 비특허문헌을 이용하여 과학과 기술간의 연계구조를 분석하거나(Carpenter, Narin, and Woolf 1981; Carpenter and Narin 1983; Narin, Noma, and Perry 1987), 특허의 인용정보를 이용하여 기업간, 지역간 및 국가간 지식의 흐름을 파악하거나(Jaffe, Trajtenberg, and Henderson 1993; Jaffe and Trajtenberg 1996; Verspagen 1997; Jaffe and Trajtenberg 1999; Jaffe, Trajtenberg, and Fogarty 2000; Fung and Chow 2001), 특허의 피인용횟수를 이용하여 각 기업이나 국가의 기술수준을 측정하는 연구가 이루어졌다(Narin, Carpenter, and Woolf 1984; Albert, Avery, Narin, and McAllister 1991; Lanjouw, Pakes, and Putnam 1998). 게다가 특허의 피인용횟수를 이용하여 개별 특허의 기술적·경제적 가치와 연계시킨 연구 또한 활발하게 이루어졌는데, 즉 후행특허들에 의해 인용을 많이 받을수록 자주 인용된 특허는 그만큼 기술적 가치가 높을 뿐만 아니라(Narin, Carpenter, and Woolf 1984; Chakrabarti 1991; Karki 1997; Harhoff, Narin, Scherer, and Vopel 1999; Yang, Lie, and Liu 2008), 경제적 가치도 높다는 것이 증명되었다(Trajtenberg 1990; Hall, Jaffe, and Trajtenberg 2000; Huang, Chen, and Zhang 2003; Hall, Jaffe, and Trajtenberg 2005).

이로써 보면, 단순한 특허건수보다는 피인용횟수가 가중된 특허건수가 지식의 흐름은 물론 질적 우수성을 판단하는데 매우 유용하게 사용될 수 있음을 알 수 있다. 또한 많은 실험 연구결과, 특허의 피인용횟수가 특허기술의 가치평

가에 매우 유용한 척도 중의 하나로 광범위하게 받아들여지고 있음을 알 수 있다.

그러나 특허의 인용은 논문과 마찬가지로 발표, 즉 공개 또는 등록된 후 일정기간이 경과해야만 활발하게 이루어지는 특성으로 인해 최근에 출원되어 발표된 특허의 피인용횟수를 제대로 파악할 수 없다는 문제점이 있다. 즉 특허가 후행 특허로부터 더 많은 인용을 받기 위해서는 일정 기간이 경과되어야 하는데, Hall 등(Hall, Jaffe, and Trajtenberg 2001)에 따르면 특허의 누적 피인용 비율이 50%에 이르기까지는 특허가 공개된 후 최소한 10년 이상 경과되어야 한다고 발표하였다. 결국 특허의 가치를 측정하는 중요한 척도인 피인용횟수는 이처럼 시간상의 제약으로 인해 즉각적인 활용이 어렵고, 특히 최근에 발표된 특허의 경우 사실상 피인용횟수를 정확하게 파악하는 것은 거의 불가능하다고 할 수 있을 것이다.

이로 인해 특허의 인용에 영향을 미칠만한 각종 요인에 대한 분석이 2000년대 들어 비교적 활발하게 이루어지고 있는데(Lanjouw and Schankerman 2004; Lee, Lee, and Song 2006; Lin, Chen, and Wu 2007), 이들 연구는 주로 프론트 페이지(front page)에 나타난 일부 데이터만을 대상으로 한 형태적인 측면의 요인을 토대로 분석한 사례로, 기술적인 측면이나 개념적인 측면에 대해서는 거의 연구가 이루어지지 않았다. 그에 따라 최근에 필자들에 의해서 특허의 인용에 영향을 미칠만한 형태적・기술적・개념적 측면에서의 제반 요인들에 대한 종합적인 분석이 이루어졌다(유재복, 정영미 2010). 그 결과 페이지 수, 청구항 수, 참고문헌 평균 피인용횟수, 기술분야 특허증감율, 서지결합도, 동시인용도 및 문헌간유사도 등 7개의 변수가 특허의 피인용횟수와 통계적으로 일정 수준 이상의 유의미한 상관관계, 즉 5% 이상의 설명력을 갖고 있음을 밝혀냈으며, 분산분석 결과 이들 7개 변수는 모두 전반적으로 대부분의 주제분야 간에 있어서 평균값의 차이가 있음을 밝혀냈다.

그러나 지금까지의 연구는 특허의 인용에 어떠한 요인들이 얼마만큼의 상관관계가 있는지를 파악하는 수준에 머물러 있을 뿐, 특허의 피인용횟수를 실제로 예측할 수 있는 예측모형이 제시된 사례는 아직까지 발표된 바가 없다.

이 연구는 필자들의 최근 선행연구에 대한 후속연구로서, 필자들의 선행연구 결과 도출된 특허의 인용에 일정 수준 이상의 영향을 미치는 제반 변수들을 토대로 특허의 피인용횟수를 예측하기 위한 예측모형을 설계하였다.

## 2. 연구방법

이 연구에서는 필자들의 최근 선행연구에서 사용한 데이터를 그대로 사용하였다. 또한 그 선행연구에서 밝혀진 특허의 인용에 영향을 미치는 주요 변수들을 대상으로 특허의 피인용횟수에 대한 예측모형을 설계하였다. 단, 여기에서는 선행연구로부터 도출된 7개의 변수 가운데 2개의 변수, 즉 '기술분야 특허증감율'과 '동시인용도'를 제외하였다. 전자는 특허단위가 아닌 해당 주제분야의 기술발전도 구간에 따라 산출된 변수값이므로 개별 특허에 대한 식별력이 미흡하고, 후자는 변수값에 따른 변량의 증감율이 일정하지 않을 뿐만 아니라 변수의 특

성상 절대적으로 후행 특허의 영향을 받으므로 인용변수로 사용하기에는 적절하지 않기 때문이다.

### 2.1 데이터 수집 및 전처리

여기에서는 미국 특허청에 등록된 특허만을 대상으로 분석하였다. 미국특허는 다른 나라의 특허와는 달리 다량의 폭넓은 인용정보를 기계가독형태로 제공하고 있고, 세계 기술시장을 대표하는 특성을 가지고 있어서(노경란 2006) 세계 각국의 중요한 신기술은 거의 대부분 미국특허로 권리화하고 있기 때문이다(Narin, Hamilton, and Olivastro 1997; Iversen 2000).

각종 분석에 필요한 기본적인 데이터는 국내 대표적인 특허검색 사이트인 (주)윕스의 특허검색시스템에서 제공하는 IPC(International Patent Classification, 이하 IPC) 탐색기능을 통해 검색한 후 '다운로드' 기능을 통해 수집하였다. 단, NT분야의 경우 별도의 IPC가 부여되어 있지 않아 키워드 검색을 통해 데이터를 확보하였다. 또한 일부 변수를 산출할 때 필요하지만 (주)윕스에서 제공하지 않은 1976년 이전의 미국 등록특허에 대한 각종 데이터는 미국특허청(U.S. Patent and Trademark Office)의 특허검색 사이트를 통해 수집하였다.

분석대상 주제분야는 국가과학기술위원회(National Science & Technology Council)에서 2001년 12월 15일에 미래유망 신기술로 선정한 5T, 즉 BT(Biology Technology), ET(Environment Technology), IT(Information Technology), NT(Nano Technology), ST(Space Technology) 분야를 그 대상으로 하였다. 주제분야별 분석대상 범위는 특허검색시스템의 IPC 탐색기능을 통해 탐색된 결과를 토대로 각 주제분야별로 등록된 특허건수의 형평성을 고려하여 그 범위를 적절하게 축소 또는 확대하였는데, 각 분야별 세부 기술분야는 〈표 1〉과 같다.

분석대상 특허의 등록년도 기간은 주제분야마다 상이하다. 주제분야별로 특허의 피인용반감기(cited half-life)를 적용하여 분석대상 기간을 설정하였기 때문이다. 일반적으로 특허의 경우 해당 주제분야의 기술발전 속도가 빠를수록 피인용반감기가 짧고 느릴수록 피인용반감기가 길어지는 경향이 있다. 이에 이 논문에서는 피인용반감기에 따른 주제분야별 차이로 인한 형평성 문제를 없애고자, (주)윕스의 특허검색

〈표 1〉 주제분야별 세부 주제분야

| 주제분야 | 세부 주제분야(IPC) |
|---|---|
| BT | Micro-organisms or Enzymes; Compositions thereof; Propagating, Preserving, or Maintaining Micro-organisms; Mutation or Genetic Engineering; Culture Media(C12N) |
| ET | Processes or Means, e.g. Batteries, for the Direct Conversion of Chemical Energy into Electrical Energy(H01M) |
| IT | Multiplex Communication(H04J) |
| NT | Application Fields of Nano-technology |
| ST | Aircraft; Aviation; Cosmonautics(B64) |

시스템에서 제공하는 미국등록특허 중 1976.1.1~2008.12.31까지의 등록특허를 대상으로 각 주제분야별로 검색한 결과 수집된 인용정보를 토대로 해당 주제분야의 특허 피인용반감기를 산출·적용하여 분석대상 기간을 설정하였다. 여기에서의 피인용반감기는 Thomson Reuters사의 JCR(Journal Citation Reports)에서 사용하는 피인용반감기 산출방법을 토대로 산출하였다.

〈표 2〉는 주제분야별 특허 피인용반감기를 적용한 분석대상 특허의 등록년도 및 특허건수이다. 분석대상 특허건수는 BT분야 5,337건, ET분야 5,929건, IT분야 4,354건, NT분야 2,313건, 그리고 ST분야 3,742건 등 총 21,675건이다.

## 2.2 변수 산출방법

이 연구에서는 특허의 피인용횟수를 예측하기 위한 예측모형을 설계하기 위해 필자들의 선행연구 결과로부터 도출된 페이지 수, 청구항 수, 참고문헌 평균 피인용횟수, 서지결합도 및 문헌간유사도 등 5개의 변수를 독립변수로 사용하였다. 이 연구에서의 유일한 종속변수인 특허의 피인용횟수와 이들 5개의 독립변수들에 대한 변수값 산출과정을 살펴보면 다음과 같다.

### 2.2.1 피인용횟수

이 연구에서의 유일한 종속변수인 '피인용횟수'는 특허의 등록년도에 따른 형평성 문제를 없애고자 해당 주제분야의 특허 피인용반감기를 분석대상이 되는 각각의 특허에 적용시켜 산출하였다. 논문과 마찬가지로 특허 역시 발표 즉, 공개 또는 등록된 시점에 따라 그 피인용횟수가 달라지기 때문이다. 따라서 특허의 등록년도 차이로 인한 피인용횟수의 형평성 문제를 없애고 분석대상이 되는 모든 특허에 대해서 동일한 조건을 부여하고자 각각의 특허에도 해당 주제분야별 피인용반감기를 적용하여 피인용횟수를 산출하였다.

분석대상 특허의 피인용횟수에 대한 산출과정을 정리하면 다음과 같다.

① 주제분야별로 특허 피인용반감기를 적용시킨 분석대상 특허집단의 각 특허의 등록번호와 등록일자를 추출하여 A집단 데이터세트를 구성하였다.

② A집단 내 각 특허에 대한 인용특허의 등록번호와 등록일자를 추출하여 서브 데이터세트를 구성한 후 이를 토대로 B집

〈표 2〉 주제분야별 분석대상 특허 등록년도 및 특허건수

| 주제 분야 | 전체 특허건수 (1976~2008년) | 반감기 | 분석대상 특허 | |
|---|---|---|---|---|
| | | | 등록년도 | 특허건수 |
| BT | 21,130건 | 12년(12.35년) | 1976년~1996년 | 5,337건 |
| ET | 18,134건 | 14년(14.10년) | 1976년~1994년 | 5,929건 |
| IT | 14,868건 | 13년(13.07년) | 1976년~1995년 | 4,354건 |
| NT | 11,514건 | 11년(10.53년) | 1976년~1997년 | 2,313건 |
| ST | 10,541건 | 18년(17.89년) | 1976년~1990년 | 3,742건 |
| 계 | 76,187건 | | | 21,675건 |

단 데이터세트를 구성하였다.

③ A집단 내 각 특허의 등록일자에 주제분야별 특허 피인용반감기를 적용시킨 후, B집단의 대응되는 서브 데이터세트 내 특허의 등록일자와 비교하여 피인용반감기 범위 내에 포함된 B집단의 특허건수를 A집단의 각 특허에 대한 피인용횟수로 산출하였다.

한 예로, BT분야의 특허인 US 4,226,938(등록일자: 1980.10.7)의 피인용횟수의 경우, 실제 총 피인용횟수는 25회이지만 BT분야의 피인용반감기를 적용하여 산출한 결과 9회로 줄어들었다. 즉 상기 특허가 등록된 1980년 10월 7일부터 1992년 10월 6일까지 12년간 다른 미국특허로부터 인용받은 횟수는 총 25회 중 9회이며, 나머지 16회는 피인용반감기 이후에 인용되었으므로 산출대상에서 제외된 것이다.

2.2.2 페이지 수

페이지 수는 수작업을 통해 해당 특허의 원문 페이지 수를 일일이 확인하여 산출하였다.

2.2.3 청구항 수

청구항 수는 특허검색시스템으로부터 다운로드한 데이터를 그대로 사용하였다.

2.2.4 참고문헌 평균 피인용횟수

'참고문헌 평균 피인용횟수'는 분석대상 특허에 대한 각 참고문헌 내 특허의 피인용횟수를 산출한 후, 이들 피인용횟수의 합을 참고문헌 내 특허 건수로 나누어 산출하였다. 단, 분석대상 특허의 참고문헌 중 피인용 정보가 수록되어 있지 않은 특허, 즉 1970년 이전의 모든 특허와 1975년 이전의 특허 중 피인용 정보가 수록되지 않은 일부 특허의 경우에는 참고문헌 피인용횟수 산출대상에서 아예 제외하였다.

2.2.5 서지결합도

'서지결합도'는 분석대상 특허집단을 A집단으로 구성하고, A집단 내 각각의 특허를 인용한 특허집단을 B집단으로 구성한 후, B집단의 2개의 특허가 A집단 내 각각의 특허를 공통적으로 몇 건 인용하고 있는지를 계산함으로써 2개 특허 간의 서지결합도를 산출하였다.

이 논문에서의 서지결합도 산출과정을 살펴보면 다음과 같다.

① 분석대상 특허집단의 각 등록번호로 A집단 데이터세트를 구성하였다.

② A집단 내 각 특허에 대한 인용특허의 등록번호로 각각의 서브 데이터세트를 구성한 후 이들 서브 데이터세트를 토대로 B집단 데이터세트를 구성하였다.

③ 이 논문에서의 분석대상은 A집단 내 2개 특허 간의 서지결합도이므로, B집단의 특허 중에서 A집단에 포함되지 않은 특허를 제거시켜 최종적으로 B집단 데이터세트를 구성하였다.

④ A집단의 각 특허에 대응되는 B집단의 서브 데이터세트를 읽은 후 해당 서브 데이터세트의 특허번호를 2개씩 쌍 결합시켰다.

⑤ 생성된 각 쌍마다 서지결합도 '1'을 부여한 후 서지결합된 특허 쌍을 대상으로 C집단 데이터세트를 구성하였다.

⑥ 각 특허 쌍이 C집단 데이터세트에 몇 번

출현하는지를 계산함으로써 2개 특허 간의 서지결합도를 산출하였다.

#### 2.2.6 문헌간유사도

'문헌간유사도'는 각 주제분야별로 분석대상 특허의 제목과 초록에 출현한 용어를 토대로 산출하였다. 먼저 제목과 초록에 출현하는 용어들을 포터 알고리즘(Porter's stemming algorithm)을 이용하여 형태소 단위로 분리한 후 어간과 어근을 추출하여 색인어를 선정하였다(Porter 1980). 선정된 색인어에는 아래 공식과 같이 TF*IDF 가중치 $w_{ik}$를 부여하였다.

$$w_{ik} = tf \cdot log(N/df)$$

위 공식에서 tf(term frequency)는 용어가 문헌에 출현한 횟수를, df(document frequency)는 용어가 출현한 문헌의 수를 의미하며, N은 총 문헌의 수를 의미한다. 각 특허 문헌을 TF*IDF 용어가중치 벡터로 표현한 후 문헌간유사도를 산출하기 위해 코사인 유사계수를 이용하였다. 코사인 유사계수는 문헌간유사도를 산출하는데 가장 일반적으로 많이 사용되는 유사계수로 문헌 $D_x$와 $D_y$의 유사도를 산출하는 공식은 다음과 같다.

$$SIM(D_x, D_y) = \frac{\sum_{i=1}^{k} t_{xi} \cdot t_{yi}}{\sqrt{\sum_{i=1}^{k} (t_{xi})^2 \cdot \sum_{i=1}^{k} (t_{yi})^2}}$$

### 2.3 분석방법

이 연구에서는 특허의 피인용횟수를 예측하기 위한 예측변수로 필자들의 최근의 선행연구에서 특허의 인용과 통계적으로 일정 수준 이상의 유의미한 상관관계가 있는 페이지 수, 청구항 수, 참고문헌 평균 피인용횟수, 서지결합도 및 문헌간유사도 등 5개의 변수를 사용하였다. 아울러 그 연구에서의 분산분석 결과, 예측변수로 사용될 이들 5개 변수들이 모두 주제분야 간에 변수값의 평균의 차이가 있는 것으로 나타남에 따라 특허의 피인용횟수에 대한 예측모형을 주제분야별로 추정하였다.

참고로, 이 연구에서 예측변수로 사용될 5개의 변수들 중 '서지결합도'와 '문헌간유사도'의 경우 그 변수값에 따라 예측모형이 달라질 수 있다. 따라서 원칙적으로는 모든 경우의 수를 고려하여 예측모형을 설계하여야 하나, 여기에서는 다음 2가지 경우만을 대상으로 예측모형을 설계하였다. 즉 필자들의 선행연구에서의 최소 조건인 분석대상 특허와의 '서지결합도'가 1 이상이거나 '문헌간유사도'가 0.3 이상인 경우(이하 A-유형 모형)와, 5개 주제분야는 물론 5개 주제분야를 통합한 전체분야에서 해당 변수값이 모두 들어 있으면서 전체분야를 대상으로 상관관계분석을 실시한 결과 변량이 가장 높은 구간으로 나타난 '서지결합도'가 3 이상이거나 '문헌간유사도'가 0.8 이상인 경우(이하 B-유형 모형)로 나누어 분석하였다.

이 연구에서는 특허의 피인용횟수에 대한 예측모형을 추정하기 위하여 SPSS for windows 12.0을 사용하였다. 먼저, 5개의 예측변수들이 그들 상호간에 독립적인지의 여부를 확인하고자 각 주제분야별로 다중공선성(multicollinearity)을 진단하였다. 다음, 다중공선성 진단 결과를 토대로 각 주제분야별로 예측모형을 추정하기 위하

여 다중회귀분석(multiple regression analysis)을 실시하였다. 여기에서 종속변수는 특허의 피인용횟수를, 독립변수로는 5개의 예측변수를 사용하였으며, 변수의 선택은 실험집단의 크기가 다른 경우에 가장 보편적으로 사용되는 방법인 '단계선택(stepwise)'을 사용하여 주제분야별 모형에서 유의하지 않은 변수를 제외시켰다. 끝으로, 다중회귀분석을 통해 주제분야별로 추정된 예측모형으로부터 산출된 특허 피인용횟수의 예측값과 그 실제값 간의 차이를 비교하였다.

## 3. 연구결과

### 3.1 예측변수의 유의성 검증

예측모형에서 예측변수로 사용될 독립변수들은 그들 간에 서로 상관관계가 있는지를 반드시 확인할 필요가 있다. 독립변수들 간에 상관관계가 높을 경우 추정된 예측모형의 회귀계수가 유의하지 않게 나타나 종속변수에 미치는 영향력을 올바르게 설명하지 못할 수가 있기 때문이다. 따라서 이들 독립변수들 간의 상관관계를 확인하고자 다중회귀분석을 통해 다중공선성 분석을 실시하였다.

여기에서의 다중공선성 진단은 보편적으로 가장 널리 사용되고 있는 분산팽창계수(variance inflation factor; 이하 VIF)를 사용하였다. 통상 VIF값이 10 이상일 경우 해당 변수는 다중공선성이 존재한다는 것을 의미한다. 따라서 다중공선성을 유발하는 변수가 2개 이상 있을 경우에는 보통 다중공선성을 가장 크게 유발시키는 독립변수, 즉 VIF값이 가장 높은 변수를 제거하거나 또는 예측모형에서 반드시 포함해야 하는 변수를 포함시키고, 상대적으로 중요도가 낮거나 변수값 산출이 어려운 변수는 제거하여 다중공선성 문제를 해결하는 것이 일반적이다.

예측변수로 사용될 5개 독립변수들의 유의성을 각 주제분야별로 진단하기 위한 다중공선성 분석은 위에서 언급한 서지결합도와 문헌간 유사도에 따른 2가지 유형으로 나누어 실시하였는데, 그 결과는 〈표 3〉, 〈표 4〉와 같다.

〈표 3〉은 종속변수인 특허의 피인용횟수와 5개 독립변수들 간의 관계에 대한 A-유형 모형의 다중공선성 통계량을 비교한 결과로, 각 주제분야별은 물론이고 5개 주제분야를 통합한 전체분야의 경우에서도 5개의 독립변수의 VIF값이 10보다 작게 나타남으로써 공선성 문제는 없는 것으로 나타났다.

〈표 4〉는 특허의 피인용횟수와 5개 독립변

〈표 3〉 A-유형 모형에 대한 주제분야별 독립변수의 다중공선성 통계량(VIF)

| | BT | ET | IT | NT | ST | 전체 |
|---|---|---|---|---|---|---|
| 페이지수 | 1.145 | 1.141 | 1.099 | 1.193 | 1.326 | 0.862 |
| 청구항수 | 1.097 | 1.120 | 1.109 | 1.179 | 1.325 | 0.915 |
| 참고문헌평균피인용횟수 | 1.105 | 1.002 | 1.326 | 1.105 | 1.246 | 0.858 |
| 서지결합도 | 1.105 | 1.148 | 1.240 | 1.175 | 1.241 | 0.764 |
| 문헌간유사도 | 1.073 | 1.130 | 1.175 | 1.230 | 1.189 | 0.792 |

〈표 4〉 B-유형 모형에 대한 주제분야별 독립변수의 다중공선성 통계량(VIF)

| | BT | ET | IT | NT | ST | 전체 |
|---|---|---|---|---|---|---|
| 페이지수 | 3.293 | 1.096 | 1.385 | 2.504 | 2.936 | 1.903 |
| 청구항수 | 1.515 | 1.210 | 1.557 | 1.949 | 2.615 | 1.319 |
| 참고문헌평균피인용횟수 | 1.365 | 1.074 | 1.821 | 1.163 | 1.485 | 1.366 |
| 서지결합도 | 32.273 | 10.203 | 6.236 | 2.029 | 7.227 | 12.394 |
| 문헌간유사도 | 36.515 | 10.061 | 6.456 | 2.344 | 7.326 | 12.908 |

수들 간의 관계에 대한 B-유형 모형의 다중공선성 통계량을 비교한 결과로, BT, ET분야와 5개 주제분야를 통합한 전체분야의 경우에서 '서지결합도'와 '문헌간유사도'의 VIF값이 10 이상으로 나타남에 따라 공선성 문제가 있는 것으로 밝혀졌다. 즉 이들 주제분야에서는 '서지결합도'와 '문헌간유사도' 간에 상관관계가 높기 때문에 이들 변수를 동시에 투입할 경우 다중공선성 문제를 유발시키므로 2개 중 하나는 제거할 필요가 있는 것으로 나타났다. 이에 따라 이 연구에서는 해당 주제분야에 대한 예측모형을 추정할 때 특허의 인용에 가장 영향을 미칠 것으로 판단되는 '문헌간유사도'를 채택하는 대신, '문헌간유사도'에 비해 상대적으로 변수값의 산출이 복잡하고 특허의 피인용횟수에 영향이 덜할 것으로 추정되는 '서지결합도'를 제거하였다.

### 3.2 주제분야별 특허인용 예측모형 추정

특허의 피인용횟수에 대한 예측모형을 추정하기 위한 다중회귀분석은 각 주제분야별로 수행하였다. 독립변수 중 '서지결합도'와 '문헌간유사도'의 경우 각 변수값에 따라 예측모형이 달라질 수 있기 때문에 앞의 다중공선성 분석에서와 마찬가지로 A-유형 모형과 B-유형 모형으로 나누어 분석하였다. 또한 앞 절에서의 다중공선성 진단 결과, 특히 B-유형 모형의 경우 BT, ET분야와 5개 주제분야를 통합한 전체분야의 경우에서 '서지결합도'와 '문헌간유사도'가 다중공선성 문제가 있으므로 이들 주제분야를 분석할 때에는 '서지결합도'를 제외한 4개의 독립변수만을 대상으로 다중회귀분석을 실시하였다.

여기에서의 분석대상 특허는 각 주제분야별로 특허의 피인용반감기를 적용시킨 5개 주제분야의 모든 특허, 즉 21,675건을 대상으로 하였다. 이 분석에서 '서지결합도'와 '문헌간유사도'의 변수값은 불가피하게 개별 특허 단위로 산출해야 하기 때문에, 앞에서 언급한 '변수값 산출방법'에 따라 산출된 '서지결합도'와 '문헌간유사도'를 가중치로 사용하여, 분석대상 특허와 쌍결합된 대응특허의 피인용횟수에 그 가중치를 곱한 값의 평균값을 해당 변수의 변수값으로 사용하였다.

〈표 5〉는 특허의 피인용횟수와 5개 독립변수들 간의 관계에 대해 A-유형 모형과 B-유형 모형으로 나누어 각 주제분야별로 실시한 다중회귀분석 결과를 정리한 것이다. 그 결과를 비교·정리하면 다음과 같다.

첫째, 2가지 유형에서의 유의변수를 보면, A-유형 모형에서는 BT, ET, IT분야와 5개 주

제분야를 통합한 전체분야의 경우 5개의 독립변수가 모두 유의수준 .05에서 유의한 것으로 나타났고, NT분야와 ST분야는 각각 '페이지 수'와 '청구항 수'를 제외한 각각 4개의 변수가 유의한 것으로 나타났다. 반면 B-유형 모형에서는 독립변수 중 '문헌간유사도'만 5개 주제분야와 5개 주제분야를 통합한 전체분야에서 유의수준 .05에서 유의한 것으로 나타났고, '페이지 수'는 BT분야와 전체분야, '서지결합도'는 IT분야 그리고 '참고문헌 평균 피인용횟수'는 NT, ST분야와 전체분야에서 유의한 것으로 나타났으며, '청구항 수'는 모든 주제분야에서 유의하지 않은 변수로 나타났다.

둘째, 각 유형에서의 유의한 변수들 중 종속변수인 특허의 피인용횟수에 대한 설명력, 즉 영향력을 보면 A-유형 모형에서는 영향력이 높은 변수가 대체로 주제분야마다 다르게 나타났는데, 즉 BT분야와 전체분야에서는 '문헌간유사도', ET분야에서는 '서지결합도', IT분야와 ST분야에서는 '참고문헌 평균 피인용횟수' 그리고 NT분야에서는 '청구항 수'가 영향력이 가장 높은 것으로 나타났다. 반면 B-유형 모형에서는 5개 주제분야 모두와 5개 주제분야를 통합한 전체분야에서도 '문헌간유사도'의 영향력이 가장 높은 것으로 나타났다.

셋째, 각 유형에서의 설명력을 나타내는 결정계수 $R^2$값을 보면 A-유형 모형에서는 $R^2$값이 .263~.325로 나타나 종속변수인 특허 피인용횟수의 26.3%~32.5%를 예측할 수 있는 것으로 나타난 반면, B-유형 모형에서는 $R^2$값이 .583~.896으로 나타나 특허 피인용횟수의 58.3%~89.6%를 예측할 수 있는 것으로 나타났다. 또한 모든 주제분야에 있어서 A-유형 모형보다 B-유형 모형의 예측력이 거의 2배가량 우수한 것으로 나타났다.

〈표 5〉와 같이 2가지 유형의 다중회귀분석 결과를 비교한 결과, A-유형 모형은 각 주제분야별로 유의한 예측변수의 수가 4개 이상이지만 영향력이 높은 변수가 대체로 상이하고 특히 예측모형의 설명력이 낮은 것으로 나타났다. 반면 B-유형 모형은 각 주제분야별로 유의한 예측변수의 수는 3개 이하이지만 영향력이 높은 변수가 모든 주제분야에서 동일하고 특히 예측모형의 설명력은 A-유형 모형보다 훨씬 우수한 것으로 나타났다.

예측모형에서 가장 중요한 것은 설명력, 즉 예측력을 추정할 수 있는 결정계수 $R^2$값이다. 일반적으로 예측모형의 설명력이 50% 수준이면 신뢰성이 있는 예측모형을 추정했다고 할 수 있고, 80% 이상이면 매우 신뢰성이 높은 예측모형을 추정했다고 할 수 있다. 이러한 관점에서 볼 때, 위의 2가지 유형에 대한 다중회귀분석 비교결과, B-유형 모형이 A-유형 모형보다 훨씬 더 우수한 것으로 나타남에 따라 여기에서는 B-유형 모형만을 대상으로 각 주제분야별로 예측모형을 추정하였는데 그 결과는 다음과 같다.

첫째, BT분야의 특허를 대상으로 다중회귀분석을 실시한 결과, $R^2$값은 .590으로 BT분야 특허의 총 피인용횟수의 59%를 설명, 즉 예측할 수 있는 것으로 나타났다. 이 모형에서는 5개의 독립변수 중 '페이지 수'와 '문헌간유사도'만이 유의한 것으로 나타났으며, 이들 변수의 회귀계수는 '페이지 수'가 .086, '문헌간유사도'가 .368으로 나타났다. 이 모형에서 유의한 독립변수의 영향력은 '문헌간유사도'(.700), '페이지 수'(.132) 순으로 높은 것으로 나타났다.

〈표 5〉 주제분야별 다중회귀분석 결과

| | | A-유형 모형 | | | | | B-유형 모형 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *b* | *S.E.(b)* | *β* | *t(p)* | $R^2$ | *b* | *S.E.(b)* | *β* | *t(p)* | $R^2$ |
| BT | (상수) | .382 | | | | .325 | 1.855 | | | | .590 |
| | 페이지 수 | .098 | .021 | .124 | 4.583(.000) | | .086 | .025 | .132 | 3.448(.001) | |
| | 청구항 수 | .057 | .025 | .061 | 2.290(.022) | | 유의하지 않음 | | | | |
| | 참고문헌 평균 피인용횟수 | .010 | .004 | .066 | 2.455(.014) | | | | | | |
| | 서지결합도 | .119 | .034 | .102 | 3.454(.001) | | 다중공선성 문제로 제거함 | | | | |
| | 문헌간유사도 | .403 | .029 | .426 | 13.957(.000) | | .368 | .020 | .700 | 18.254(.000) | |
| ET | (상수) | -.978 | | | | .293 | 1.577 | | | | .591 |
| | 페이지 수 | .154 | .038 | .082 | 4.082(.000) | | 유의하지 않음 | | | | |
| | 청구항 수 | .076 | .018 | .083 | 4.161(.000) | | | | | | |
| | 참고문헌 평균 피인용횟수 | .029 | .012 | .046 | 2.411(.016) | | | | | | |
| | 서지결합도 | .585 | .032 | .384 | 18.214(.000) | | 다중공선성 문제로 제거함 | | | | |
| | 문헌간유사도 | .269 | .029 | .194 | 9.179(.000) | | .666 | .031 | .769 | 21.220(.000) | |
| IT | (상수) | -6.911 | | | | .279 | 2.233 | | | | .760 |
| | 페이지 수 | .084 | .029 | .053 | 2.886(.004) | | 유의하지 않음 | | | | |
| | 청구항 수 | .249 | .043 | .107 | 5.770(.000) | | | | | | |
| | 참고문헌 평균 피인용횟수 | .324 | .023 | .293 | 14.296(.000) | | | | | | |
| | 서지결합도 | .081 | .033 | .049 | 2.442(.015) | | .364 | .174 | .354 | 2.087(.042) | |
| | 문헌간유사도 | .474 | .035 | .269 | 13.664(.000) | | .512 | .162 | .536 | 3.158(.003) | |
| NT | (상수) | 1.884 | | | | .399 | -5.817 | | | | .896 |
| | 페이지 수 | 유의하지 않음 | | | | | 유의하지 않음 | | | | |
| | 청구항 수 | .337 | .065 | .301 | 5.219(.000) | | | | | | |
| | 참고문헌 평균 피인용횟수 | .068 | .027 | .151 | 2.558(.011) | | .264 | .076 | .431 | 3.494(.000) | |
| | 서지결합도 | .248 | .076 | .230 | 3.258(.001) | | 유의하지 않음 | | | | |
| | 문헌간유사도 | .193 | .051 | .266 | 3.788(.000) | | .578 | .091 | .786 | 6.372(.000) | |
| ST | (상수) | -1.837 | | | | .263 | -6.434 | | | | .718 |
| | 페이지 수 | .242 | .034 | .191 | 7.097(.000) | | 유의하지 않음 | | | | |
| | 청구항 수 | 유의하지 않음 | | | | | | | | | |
| | 참고문헌 평균 피인용횟수 | .278 | .028 | .291 | 9.846(.000) | | .504 | .214 | .361 | 2.349(.037) | |
| | 서지결합도 | .166 | .038 | .135 | 4.402(.000) | | 유의하지 않음 | | | | |
| | 문헌간유사도 | .249 | .041 | .184 | 6.035(.000) | | 1.144 | .235 | .746 | 4.861(.000) | |
| 전체 | (상수) | -.937 | | | | .291 | 2.542 | | | | .583 |
| | 페이지 수 | .081 | .016 | .057 | 5.135(.000) | | .071 | .020 | .069 | 3.477(.001) | |
| | 청구항 수 | .167 | .018 | .098 | 9.069(.000) | | 유의하지 않음 | | | | |
| | 참고문헌 평균 피인용횟수 | .038 | .005 | .081 | 7.167(.000) | | -.014 | .005 | -.059 | -3.019(.003) | |
| | 서지결합도 | .270 | .018 | .184 | 15.026(.000) | | 다중공선성 문제로 제거함 | | | | |
| | 문헌간유사도 | .489 | .018 | .339 | 27.239(.000) | | .562 | .015 | .753 | 36.344(.000) | |

이를 토대로 추정한 BT분야의 예측모형은 다음과 같다.

# 예측모형 1: BT분야

피인용횟수 = 1.855 + (0.086 × 페이지 수) + (0.368 × 문헌간유사도)

둘째, ET분야에 대한 분석 결과, $R^2$값은 .591로 ET분야 특허의 총 피인용횟수의 59.1%를 설명할 수 있는 것으로 나타났다. 이 모형에서는 5개의 독립변수 중 오직 '문헌간유사도'만이 유의한 것으로 나타났으며, 이 변수의 회귀계수는 .666으로 나타났다. 이를 토대로 추정한 ET분야의 예측모형은 다음과 같다.

# 예측모형 2: ET분야

피인용횟수 = 1.577 + (0.666 × 문헌간유사도)

셋째, IT분야에 대한 분석 결과, $R^2$값은 .760으로 IT분야 특허의 총 피인용횟수의 76%를 설명할 수 있는 것으로 나타났다. 이 모형에서 유의한 독립변수는 2개로 나타났는데, 이들 독립변수의 회귀계수를 보면 '서지결합도'가 .364, '문헌간유사도'가 .512로 나타났다. 이 모형에서 유의한 독립변수의 영향력은 '문헌간유사도'(.536), '서지결합도'(.354) 순으로 높은 것으로 나타났다. 이를 토대로 추정한 IT분야의 예측모형은 다음과 같다.

# 예측모형 3: IT분야

피인용횟수 = 2.233 + (0.364 × 서지결합도) + (0.512 × 문헌간유사도)

넷째, NT분야에 대한 분석 결과, $R^2$값은 .896으로 NT분야 특허의 총 피인용횟수의 89.6%를 설명할 수 있는 것으로 나타났다. 이 모형에서도 2개의 독립변수가 유의한 것으로 나타났고, 이들 독립변수의 회귀계수를 보면 '참고문헌 평균 피인용횟수'가 .264, '문헌간유사도'가 .578로 나타났다. 이 모형에서 유의한 독립변수의 영향력은 '문헌간유사도'(.786), '참고문헌 평균 피인용횟수'(.431) 순으로 높은 것으로 나타났다. 이를 토대로 추정한 NT분야의 예측모형은 다음과 같다.

# 예측모형 4: NT분야

피인용횟수 = -5.817 + (0.264 × 참고문헌 평균 피인용횟수) + (0.578 × 문헌간유사도)

다섯째, ST분야에 대한 분석 결과, $R^2$값은 .718로 ST분야 특허의 총 피인용횟수의 71.8%를 설명할 수 있는 것으로 나타났다. 이 모형에서도 2개의 독립변수가 유의한 것으로 나타났고, 이들 독립변수의 회귀계수를 보면 '참고문헌 평균 피인용횟수'가 .504, '문헌간유사도'가 1.144로 나타났다. 이 모형에서 유의한 독립변수의 영향력은 '문헌간유사도'(.746), '참고문헌 평균 피인용횟수'(.361) 순으로 높은 것으로 나타났다. 이를 토대로 추정한 ST분야의 예측모형은 다음과 같다.

# 예측모형 5: ST분야

피인용횟수 = -6.434 + (0.504 × 참고문헌 평균 피인용횟수) + (1.144 × 문헌간유사도)

여섯째, 5개 주제분야를 통합한 전체분야에

대한 분석 결과, 결정계수 $R^2$값은 .583으로 분석 대상 특허의 총 피인용횟수의 58.3%를 설명할 수 있는 것으로 나타났다. 이 모형에서 유의한 독립변수는 3개로 나타났으며, 이들 독립변수의 회귀계수를 보면 '페이지 수' .071, '참고문헌 평균 피인용횟수' -.014, '문헌간유사도'가 .562로 나타났다. 이 모형에서 유의한 독립변수의 영향력은 '문헌간유사도'(.753), '페이지 수'(.069), '참고문헌 평균 피인용횟수'(-.059) 순이며, '참고문헌 평균 피인용횟수'는 음(-)의 영향력을 가진 것으로 나타났다. 이를 토대로 추정한 전체분야의 예측모형은 다음과 같다.

# 예측모형 6: 전체분야

피인용횟수 = 2.542 + (0.071 × 페이지 수) + (-0.014 × 참고문헌 평균 피인용횟수) + (0.562 × 문헌간유사도)

### 3.3 예측값과 실제값의 차이 비교

여기에서는 앞 절에서 추정한 각 주제분야별 예측모형을 통해 산출된 특허 피인용횟수의 예측값과 실제값 간의 차이를 여러 가지 측면에서 비교하였다.

〈표 6〉은 SPSS for windows 12.0을 사용하여 얻은 기술통계량과 잔차통계량을 토대로 예측값과 실제값 간의 차이를 비교한 결과로, 이를 주제분야별로 정리하면 다음과 같다.

참고로, 〈표 6〉에서 유효 특허건수는 총 1,345건으로 이 연구에서의 분석대상 특허인 21,675건의 6.2%에 불과한 것으로 나타났는데, 이는 B-유형 모형 즉 '서지결합도'가 3 이상이거나 '문헌간유사도'가 0.8 이상인 특허만으로 구성되었기 때문이다. 또한 5개 주제분야를 통합한 전체분야의 유효 특허건수가 실제 5개 주제분야의 합계 건수보다 적게 나타났는데, 이는 전체분야의 예측모형에서 유의한 예측변수와 각 주제분야별 예측모형에서 유의한 예측변수들이 서로 다르기 때문이다.

첫째, BT분야의 경우 유의한 변수값이 모두 들어있는 유효 특허건수는 472건이며, 예측값과 실제값의 평균은 각각 10.59(SD=9.56)와 8.36(SD=13.01)으로 약간의 차이가 있는 것으로 나타났다. 예측값에서 실제값을 뺀 값의 절대값인 절대잔차의 평균은 6.16(SD=5.99)으로 나타났으며, 최소값은 0.09, 최대값은 53.79로 나타났다.

〈표 6〉 주제분야별 피인용횟수의 예측값과 실제값 비교

| 주제분야 | 유효특허건수 | 예측값 | | | | 실제값 | | | | 절대잔차 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 최소값 | 최대값 | 평균 | 표준편차 | 최소값 | 최대값 | 평균 | 표준편차 | 최소값 | 최대값 | 평균 | 표준편차 |
| BT | 472 | 4.56 | 124.64 | 10.59 | 9.56 | 0.00 | 285.00 | 8.36 | 13.01 | 0.09 | 53.79 | 6.16 | 5.99 |
| ET | 417 | 1.63 | 51.30 | 7.28 | 6.36 | 0.00 | 108.00 | 7.74 | 9.93 | 0.01 | 37.55 | 4.15 | 5.09 |
| IT | 52 | 3.11 | 76.67 | 21.62 | 19.04 | 0.00 | 491.00 | 19.45 | 26.66 | 0.12 | 38.16 | 7.15 | 7.91 |
| NT | 290 | -5.82 | 82.98 | 13.34 | 16.12 | 0.00 | 174.00 | 12.41 | 16.12 | 0.09 | 69.49 | 7.89 | 9.06 |
| ST | 114 | -6.43 | 48.37 | 6.31 | 7.67 | 0.00 | 157.00 | 6.29 | 6.60 | 0.04 | 38.37 | 4.74 | 4.86 |
| 전체 | 1,249 | -10.14 | 221.98 | 12.87 | 14.31 | 0.00 | 491.00 | 10.49 | 16.41 | 0.00 | 190.98 | 7.11 | 10.80 |

둘째, ET분야의 경우 유효 특허건수는 417건이며, 예측값과 실제값의 평균은 각각 7.28(SD=6.36)과 7.74(SD=9.93)로 비교적 유사한 것으로 나타났다. 예측값과 실제값 간의 절대잔차의 평균은 4.15(SD=5.09)로 나타났으며, 최소값은 0.01, 최대값은 37.55로 나타났다.

셋째, IT분야의 경우 유효 특허건수는 52건이며, 예측값과 실제값의 평균은 각각 21.62(SD=19.04)와 19.45(SD=26.66)로 약간의 차이가 있는 것으로 나타났다. 예측값과 실제값 간의 절대잔차의 평균은 7.15(SD=7.91)로 나타났으며, 최소값은 0.12, 최대값은 38.16으로 나타났다.

넷째, NT분야의 경우 유효 특허건수는 290건이며, 예측값과 실제값의 평균은 각각 13.34(SD=16.12)와 12.41(SD=16.12)로 약간의 차이가 있는 것으로 나타났다. 예측값과 실제값 간의 절대잔차의 평균은 7.89(SD=9.06)로 나타났으며, 최소값은 0.09, 최대값은 69.49로 나타났다.

다섯째, ST분야의 경우 유효 특허건수는 114건이며, 예측값과 실제값의 평균은 각각 6.31(SD=7.67)과 6.29(SD=6.60)로 매우 유사한 것으로 나타났다. 예측값과 실제값 간의 절대잔차의 평균은 4.74(SD=4.86)로 나타났으며, 최소값은 0.04, 최대값은 38.37로 나타났다.

여섯째, 5개 주제분야를 통합한 전체분야의 경우 유효 특허건수가 1,249건이며, 예측값과 실제값의 평균은 각각 12.87(SD=14.31)과 10.49(SD=16.41)로 약간의 차이가 있는 것으로 나타났다. 예측값과 실제값 간의 절대잔차의 평균은 7.11(SD= 10.80)로 나타났으며, 최소값은 0.00, 최대값은 190.98로 나타났다.

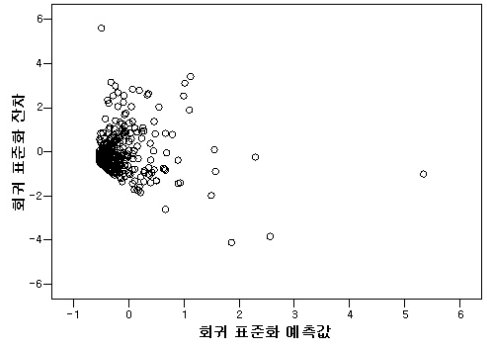
각 주제분야별 예측모형으로부터 산출된 예측값과 실제값 간의 차이를 표준화잔차 산점도를 통해 살펴보면 〈그림 1〉과 같다. 표준화잔차란 원시잔차(raw residual)를 각각의 표준편차로 나누어 표준화한 것으로 주로 이상치 유무를 판단할 때 사용되는데, 통상 표준화잔차의 절대값이 3 이상이면 이상치의 존재가 확실함을 의미한다. 따라서 가능한 한 3 이상의 값이 적을수록 추정된 예측모형이 보다 더 적합하다고 할 수 있다.

〈그림 1〉의 잔차산점도를 살펴보면, 모든 주제분야에 있어서 관찰값들이 특정한 패턴을 보이지 않고 0 주위에 무작위로 고루 퍼져 있어 예측모형의 기본 조건인 선형성과 동일분산성을 충족시키고 있음을 알 수 있다. 또한 관찰값들이 -3에서 3 사이에 집중적으로 분포되어 있는 반면, -3에서 3 사이를 벗어난 이상치 영역에 속하는 관찰값들이 소수에 불과한 점으로 미루어 보아 이 연구에서 추정된 예측모형은 모두 적합하다고 할 수 있다.

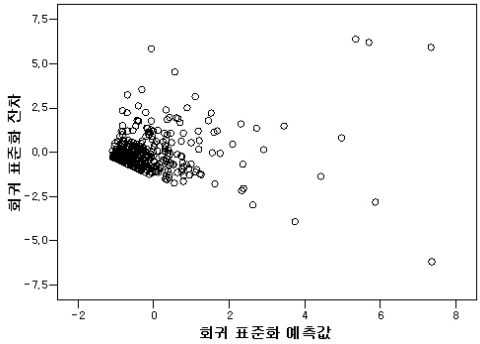
한편 〈그림 2〉는 각 주제분야별 예측모형으로부터 산출한 특허의 피인용횟수에 대한 예측값과 실제값을 실제로 비교한 결과를 그래프로 나타낸 것이다. 여기에서 비교대상 특허는 각 주제분야의 분석대상특허 중에서 유의한 변수값이 모두 들어있는 유효한 특허를 대상으로 최근년도를 기준으로 각각 50건씩 선정하였으며, 5개 주제분야를 통합한 전체분야의 경우에는 각 주제분야별로 최근 년도를 기준으로 10건씩 선정한 것이다.

〈그림 2〉의 그래프를 보면, 각 주제분야별 예측값과 실제값의 추이곡선이 전반적으로 매우 비슷하게 나타난 것을 알 수 있는데, 이로써 보

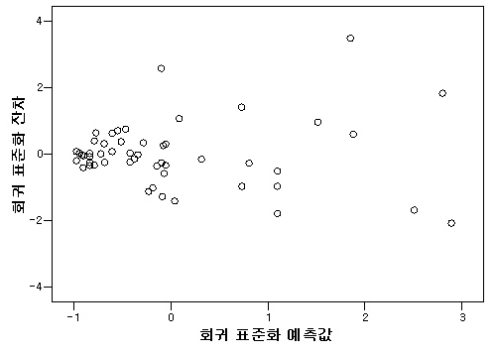
면 이 연구에서 추정된 각 주제분야별 예측모형 은 모두 적합하다고 할 수 있을 것이다.



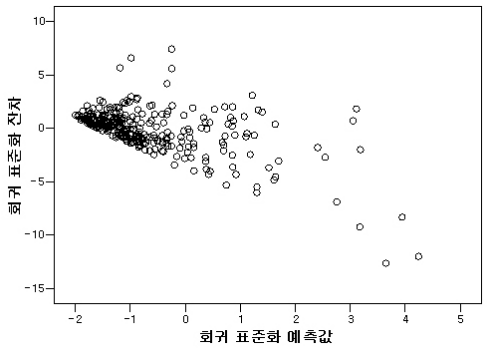

(BT분야)



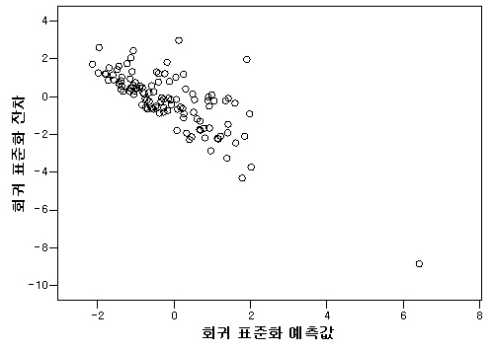

(ET분야)



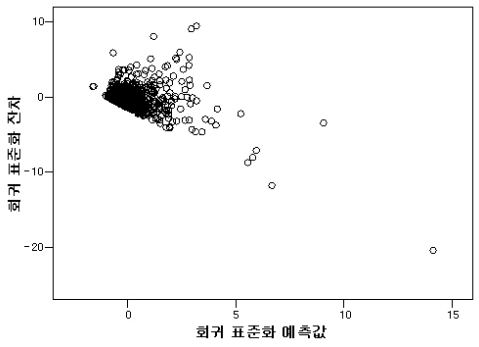

(IT분야)




(NT분야)

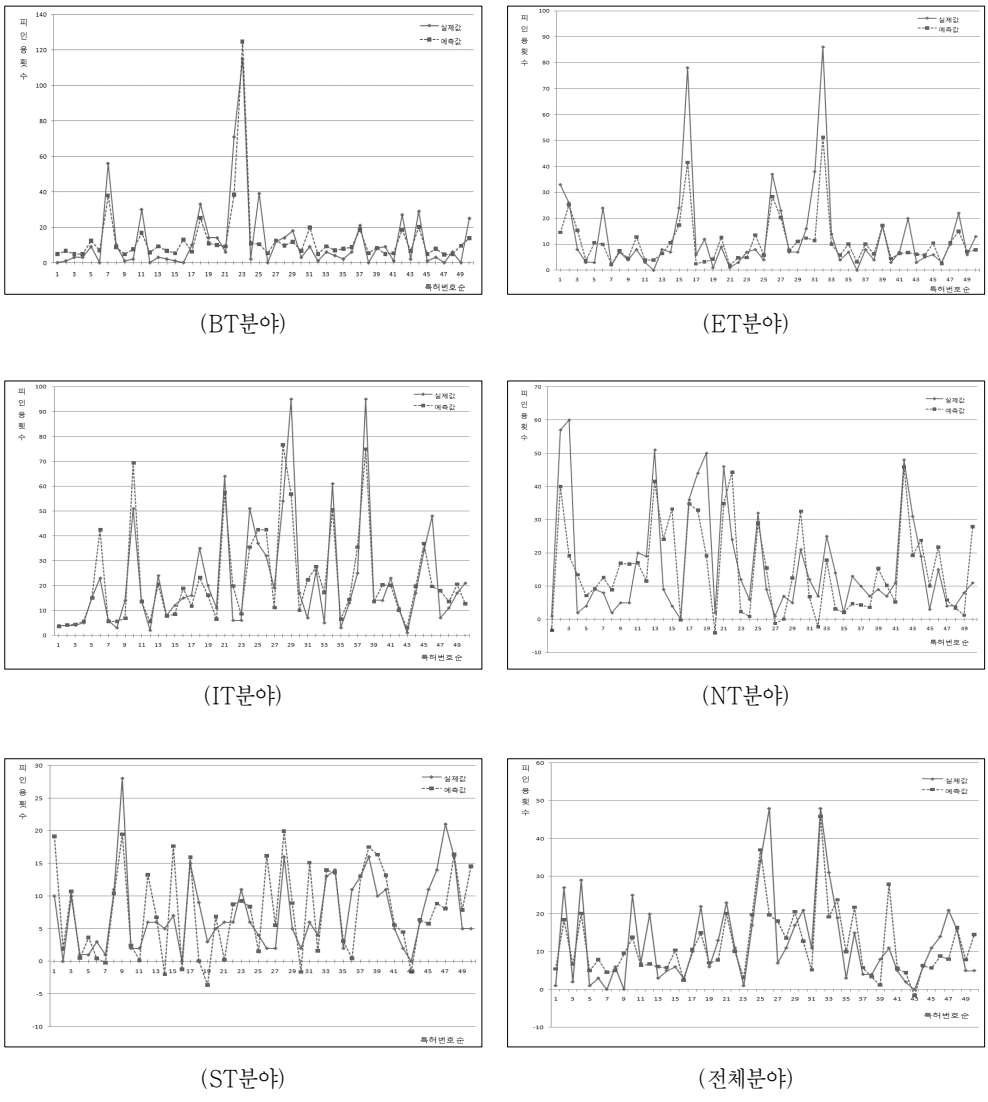



(ST분야)




(전체분야)

〈그림 1〉 주제분야별 잔차산점도




(BT분야) (ET분야)

(IT분야) (NT분야)

(ST분야) (전체분야)

〈그림 2〉 주제분야별 예측값과 실제값의 비교

## 4. 결 론

이 연구에서는 특허의 인용에 일정 수준 이상의 영향을 미치는 것으로 밝혀진 주요 변수들을 대상으로 특허의 피인용횟수를 예측할 수 있는 예측모형을 주제분야별로 추정한 후 이들 예측모형으로부터 산출된 예측값과 실제값을 비교하였는데, 이를 정리하면 다음과 같다.

첫째, '서지결합도'와 '문헌간유사도'에 따른 2가지 유형에 대한 다중회귀분석을 실시한 결

과, '서지결합도'가 1 이상 또는 '문헌간유사도'가 0.3 이상인 경우의 예측모형의 예측력은 주제분야에 따라 26.3%~32.5%로 나타난 반면, '서지결합도'가 3 이상 또는 '문헌간유사도'가 0.8 이상인 경우의 예측모형의 예측력은 주제분야에 따라 58.3%~89.6%로 나타났다. 또한 모든 주제분야에 있어서 전자보다 후자의 예측모형의 예측력이 거의 2배가량 우수한 것으로 나타났다.

둘째, '서지결합도'가 3 이상 또는 '문헌간유사도'가 0.8 이상인 경우의 예측모형에서 예측변수로 사용된 5개의 독립변수 중 '문헌간유사도'가 5개 주제분야 모두에서는 물론 5개 주제분야를 통합한 전체분야에 있어서도 가장 영향력이 높은 것으로 나타났다.

셋째, 이 연구에서 추정된 각 주제분야별 예측모형으로부터 특허 피인용횟수에 대한 예측값과 실제값을 비교한 결과, 표준화잔차 산점도에서는 관찰값들이 특정 패턴을 형성하지 않고 0을 중심으로 -3과 3사이에 무작위로 고루 퍼져 있고, 추이곡선에서도 예측값과 실제값의 그래프가 전반적으로 매우 비슷한 것으로 나타나, 이 연구에서 추정한 각 주제분야별 예측모형은 모두 적합한 것으로 나타났다.

이상의 연구결과는 시간 의존적인 인용의 특성상 최근에 발표된 특허에 대해서는 그 피인용횟수를 제대로 파악하기 어렵다는 한계에도 불구하고 이를 예측할 수 있는 가능성을 보여주었다. 그러나 이 연구에서 통계적으로 신뢰성이 있는 예측모형을 추정하기 위해서는 불가피하게 일정 조건을 갖는 소수의 특허만을 대상으로 그 피인용횟수를 예측할 수밖에 없다는 제한된 예측모형이라는 한계성을 지니고 있다.

한편 이 연구결과를 토대로 향후 특허인용 예측모형을 설계함에 있어서 시사하는 바를 정리하면 다음과 같다.

첫째, 5개 주제분야에서 상관관계와 영향력이 다른 변수들에 비해 상대적으로 높은 것으로 판명된 '서지결합도'와 '문헌간유사도'의 경우 해당 변수값에 따라서는 물론이고 주제분야마다 상관관계 및 예측모형의 결정계수가 크게 달라지므로 이러한 점을 고려하여 예측모형을 설계하도록 해야 할 것이다. 아울러 특히 이들 2개의 변수는 일부 주제분야에 있어서 다중공선성 문제를 일으킬 가능성이 높다는 점도 감안하여 예측모형을 설계해야 할 것으로 보인다.

둘째, 각 주제분야별로 각종 변수의 변수값에 대한 평균의 차이가 있을 뿐만 아니라 각 주제분야마다 회귀모형에서 유의한 변수가 다른 경우가 많다는 점을 고려할 때, 모든 주제분야를 아우르는 통합적인 예측모형보다는 각 주제분야별로 예측모형을 설계하는 것이 더 바람직할 것으로 보인다.

셋째, 이 연구에서 추정된 예측모형에서 특허의 피인용횟수에 가장 영향을 미치는 변수로는 모든 주제분야에 걸쳐 '문헌간유사도'로 나타난 만큼, '문헌간유사도'를 토대로 특허의 피인용횟수에 대한 예측력을 좀 더 높일 수 있는 다양하고 심도있는 연구가 필요할 것으로 보인다. 각 특허문헌에 부여된 색인어나 특허의 권리범위인 청구항을 활용하거나 발명의 명칭, 즉 제목에 출현한 용어에 가중치를 부여하는 방법도 그 고려대상이 될 수 있을 것이다.

# 참 고 문 헌

노경란. 2006.『특허분석을 통한 과학기술자의 과학논문 인용행태에 관한 연구』. 박사학위논문, 연세대학교, 문헌정보학과.

박종용. 2008.『특허인용 관계를 이용한 기술지식 흐름의 실증분석: 산업수준의 네트워크 분석과 클러스터 분석』. 박사학위논문, 서울대학교, 기술경영경제정책대학원.

유재복, 정영미. 2010. 특허 인용에 영향을 미치는 요인 분석.『정보관리학회지』, 27(1): 103-118.

Albert, M. B., D. Avery, F. Narin, and P. McAllister. 1991. "Direct validation of citation counts as indicators of industrially import patents." *Research Policy*, 20(3): 251-259.

Ashton, W. B. and R. K. Sen. 1989. "Using patent information in technology business planning." *Research Technology Management*, 32(1): 36-42.

Carpenter, M. P., F. Narin, and P. Woolf. 1981. "Citation rates to technologically important patents." *World Patent Information*, 3(4): 160-163.

Carpenter, M. P. and F. Narin. 1983. "Validation study: patent citations as indicators of science and foreign dependence." *World Patent Information*, 5(3): 180-185.

Chakrabarti, A. K. 1991. "Competition in high technology: analysis of problems of US, Japan, UK, France, West Germany and Canada." *IEEE Transactions on Engineering Management*, 38(1): 78-84.

Fung, M. K. and W. W. Chow. 2001. "Measuring the intensity of knowledge flow with patent statistics." *Economic Letters*, 74: 353-358.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg. 2000. *Market value and patent citations: a first look*. NBER Working Paper No. 7741, Cambridge: National Bureau of Economic Research.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg. 2001. *The NBER patent citation data file: lessons, insights and methodological tools*. NBER Working Paper No. 8498, Cambridge: National Bureau of Economic Research.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg. 2005. "Market value and patent citations." *RAND Journal of Economics*, 36(1): 16-38.

Harhoff, D., F. Narin, F. M. Scherer, and K. Vopel. 1999. "Citation frequency and the value of patented inventions." *Review of Economics & Statistics*, 81(3): 511-515.

Huang, M. X., D. R. Chen, and H. W. Zhang. 2003. "Evaluation of national competitiveness from the view of patent technology." *Paper of China Society of Li-*

*brary*, 70: 18-30.

Iversen, E. J. 2000. "An excursion into the patent-bibliometrics of Norwegian patenting." *Scientometrics*, 49(1): 63-80.

Jaffe, A. B., M. Trajtenberg, and R. Henderson. 1993. "Geographic localization of knowledge spillovers as evidenced by patent citations." *Quarterly Journal of Economics*, 108(3): 577-598.

Jaffe, A. B. and M. Trajtenberg. 1996. *Flows of knowledge from universities and federal, abs: modeling the flow of patent citations over time and across institutional and geographic boundaries.* NBER Working Paper No. 5712. Cambridge: National Bureau of Economic Research.

Jaffe, A. B. and M. Trajtenberg. 1999. "International knowledge flows: evidence from patent citations." *Economics of Innovation and New Technology*, 8(1-2): 105-136.

Jaffe, A. B., M. Trajtenberg, and M. S. Fogarty. 2000. "Knowledge spillovers and patent citation: evidence from a survey of inventors." *American Economic Review*, 90(2): 215-218.

Karki, M. M. S. 1997. "Patent citation analysis: a policy analysis tool." *World Patent Information*, 33(4): 269-272.

Lanjouw, J. O., A. Pakes, and J. Putnam. 1998. "How to count patents and value intellectual property: uses of patent renewal and application data." *The Journal of Industrial Economics*, 46(4): 405-432.

Lanjouw, J. O. and M. Schankerman. 2004. "Patent quality and research productivity: measuring innovation with multiple indicators." *The Economic Journal*, 114 (495): 441-465.

Lee, Y. G., J. D. Lee, and Y. I. Song. 2006. "An Analysis of citation counts of ETRI-invented US patents." *ETRI Journal*, 28(4): 541-544.

Lin, B. W., C. J. Chen, and H. L. Wu. 2007. "Predicting citations to biotechnology patents based on the information from the patent documents." *International Journal of Technology Management*, 40(1-3): 87-100.

Narin, F., M. P. Carpenter, and P. Woolf. 1984. "Technological performance assessments based on patents and patent citations." *IEEE Transactions on Engineering Management*, 36(2): 172-183.

Narin, F., K. S. Hamilton, and D. Olivastro. 1997. "The increasing linkage between US technology and public science." *Research Policy*, 26(3): 317-330.

Narin, F., E. Noma, and R. Perry. 1987. "Patents as indicators of corporate technological strength." *Research Policy*, 16(2-4): 143-155.

Porter, M. F. 1980. "An algorithm for suffix stripping." *Program*, 14: 130-137.

Trajtenberg, M. 1990. "A penny for your quotes: Patent citations and the value of innovations." *The Rand Journal of Economics*, 21(1): 172-187.

Verspagen, B. 1997. "Measuring intersectoral spillovers: estimates from the European and US patent office databases." *Economic Systems Research*, 9(1): 47-66.

Yang, Z. K., Q. N. Lie, and Z. Y. Liu. 2008. "Top ten highly cited patents in USPTO." *Proceedings of WIS 2008*. [online]. [cited 2009.2.9]. 〈http://creativecommons.org/licenses/by/2.0〉.