An Extraction Method of Bibliographic Information from the US Patents: Using an HTML Parsing Technique

미국 특허 서지정보 추출 방법에 대한 연구: HTML 파싱 기법의 활용을 중심으로

Yoo-Jin Han*
Seung-Woo Oh**

ABSTRACT

This study aims to provide a method of extracting the most recent information on US patent documents. An HTML paring technique that can directly connect to the US Patent and Trademark Office (USPTO) Web page is adopted. After obtaining a list of 50 documents through a keyword searching method, this study suggested an algorithm, using HTML parsing techniques, which can extract a patent number, an applicant, and the US patent class information. The study also revealed an algorithm by which we can extract both patents and subsequent patents using their closely connected relationship, that is a very distinctive characteristic of US patent documents. Although the proposed method has several limitations, it can supplement existing databases effectively in terms of timeliness and comprehensiveness.

초 록

본 연구는 미국 특허 문서에서 가장 최신의 정보를 추출할 수 있는 방법을 제시하였다. 이를 위해 미국특허청 웹페이지에 직접 접속하여, HTML 문서를 파싱하는 방법을 제시하였다. 먼저 관심 있는 키워드로 검색을 한 후 50개로 이루어진 리스트가 출력되면, HTML 파싱 기법을 이용하여 여기서 직접 특허번호, 출원인, 미국 특허 클래스와 같은 주요 서지정보를 추출할 수 있는 알고리즘을 제안하였다. 또한 미국 특허문서에서 특수하게 제공되는 선·후행 특허간의 관계를 활용해 본 특허와 후행 특허의 미국 특허 클래스를 동시에 추출 할 수 있는 알고리즘도 보여주었다. 본 연구에서 제시한 방법은 몇 가지 한계를 가지지만, 적시성·포괄성 측면에서 이미 존재하는 데이터베이스를 보완할 수 있을 것이다.

Keywords: US patents, bibliographic information, extraction, HTML parsing
미국 특허, 서지정보, 추출, HTML 파싱

---

* Assistant Professor, School of Global Service, Sookmyung Women's Unversity (yjhan@sookmyung.ac.kr)
** Ph.D. Candidate, Technology Management, Economics and Policy Program, Seoul National University(lovision@temep.snu.ac.kr)

■ Received : 16 April 2010 ■ Revised : 4 June 2010 ■ Accepted : 13 June 2010
■ Journal of the Korean Society for Information Management, 27(2): 7-20, 2010.
〔DOI:10.3743/KOSIM.2010.27.2.007〕

1. Introduction

The US patent database is considered the most comprehensive and valuable in terms of the progressivity and applicability of its stored technologies, which, in turn, allows many users to exploit it (Hall et al. 2001).[1] To tap into the bibliographic information in the US patent database, however, every HTML document must be opened one by one after searching for the patents of interest. Alternatively, commercial databases such as the Derwent Patent Database have to be used to receive a large volume of bibliographic information (Simmons 2004).[2] Although this database is widely used depending on various purposes (Calcagno 2008), it still offers insufficient information from the perspectives of timeliness and comprehensiveness compared to obtaining data directly from the US patent database. In addition, although the National Bureau of Economic Research (NBER) built a large-scale database with an available time period of patent documents ranging from 1975 to 2006, this database has had, thus far, limitations similar to those of the US patent database.[3] Therefore, in this research, we aim to propose a new method by which necessary bibliographic information can be extracted directly from the US patent database. In this way, we can extract the most up-to-date information about the corresponding fields according to our research purposes and, therefore, reduce dependency on the format, amount and time-period of patent data provided by various readily available secondary databases such as the ones enumerated above.

There are various ways to extract necessary data from the Web, such as Natural Language Processing (NLP) and Information Retrieval (IR). These two algorithms are based on the logic whereby users can obtain the exact data of their interest - representative examples are WESTLAW, which offers legal information, and NEXIS and Dow Jones/Retrieval, which provide journalistic information. There is also a method that resizes fonts and graphics for mobile phones and personal digital assistants (PDAs). Finally, there exist screen reader programs such as the Screen Reader and Microsoft's Narrator which help the blind by reading aloud after the necessary information is retrieved. Although there are numerous ways to extract data from the Web, we propose a method to extract necessary fields and contents from a large database composed of a plethora of HTML docu-

1) There are plenty of studies employing the US patent database, including the impact of technological progress on one country's economy and the priority selection of R&D using basic statistical data such as the number of patent applications or the number of patent registrations; and the technological impact analysis or technological knowledge flows using citation information ("citation" is a very unique set of information provided by the US patent database), which shows the relationship between one patent and those that influenced it and that between one patent and those influenced by it (Ernst 2003; Yoon and Park 2003; Yoo and Chung 2010).

2) A commercial patent database offered by Thomson.

3) The initial attempt was made by Hall et al. (2001), who offered the major bibliographic information from 1975 to 1999; the subsequent dataset revealed covers the period 1975 to 2000; and the final version was open to the public showing information between 1975 and 2006.

ments using an HTML parsing technique.[4] In their research, they showed a way to extract necessary parts by deleting those that users do not want from the HTML documents, namely, HTML parsing. A similar algorithm is adopted in this research because this method can be effectively applied to the process of extracting required data from the US patent database.

There are multifarious fields in a patent database, such as those that contain basic information embracing patent number, application date, and class[5], as well as those that include an abstract and description about an invention. In addition, depending on the fields users initially want to extract, the number of patents retrieved broadly ranges from a single digit to a million.[6] In most cases, however, we first search for the patents at the technological level when we intend to develop a new technology, and then we extract information about class and applicant[7] (Yoon and Park 2004; Lichtenthaler 2009). Therefore, in this research, an algorithm that extracts the corresponding class and applicant after searching for a specific "technology" is proposed. As an extended form, we also propose an advanced algorithm that can be employed when two patent documents are linked by a hyperlink. As an exemplary keyword for technology, we select the term "semiconductor."

2. Characteristics of US Patent Documents

There are over 3.5 million patents registered in the US patent database (see Table 1), and users vary from inventors and scientists to engineers worldwide. US patent documents contain various bibliographic information, including descriptions about a certain invention, in order to grant an inventor a right of appropriability (see Table 2), and it is possible to search for each patent by accessing the US Patent and Trademark Office (USPTO) website (www.usp to.gov).

The primary set of information covered by the US patent database is composed of two parts - regis

4) At present, we cannot estimate since when the researchers started to use the way to extract necessary data from the web using this method. Despite this non-availability of the exact reference source, there are many articles which make applications of this method such as Gupta et al. (2005).

5) There are many types of technology that can be filed in a "patent" form, which can be categorized by a specific scheme called a "class." There are bifurcated ways to provide information about a "class" in US patent documents: US Patent Classification (USC) and International Patent Classification (IPC). The USC has a two-level structure consisting of a "class" and a "sub-class" (e.g., in the case of 100/000, 100 indicates a class and 000 shows a subclass), while the IPC has a five-level anatomy composed of a "section," a "class," a "subclass," a "group," and a "'subgroup" (e.g., in the case of A01B 00/01, the "section" refers to A; the "class" 01; the "subclass" B; the "group" 00; and the "subgroup" 01).

6) There are over 3.5 million patents registered in the US patent database.

7) The largest portion of applicants is taken up by firms, and since universities and public research institutes expect revenue through licensing, they can become the "applicants." If central or local governments file patents, they can become the "applicants."

〈Table 1〉 Number of US patent registrations

Year	1976	1977	1978	1979	1980	1981	1982	1983	1984
No. of patent regist-rations	70,236	65,269	66,102	48,853	61,827	65,770	57,889	56,862	67,201
Year	1985	1986	1987	1988	1989	1990	1991	1992	1993
No. of patent regist-rations	71,661	70,860	82,952	77,924	95,539	90,366	96,514	97,443	98,344
Year	1994	1995	1996	1997	1998	1999	2000	2001	2002
No. of patent regist-rations	101,676	101,419	109,646	111,984	147,520	153,487	157,496	166,038	163,518
Year	2003	2004	2005	2006	2007	2008	Total		
No. of patent regist-rations	169,035	164,291	143,806	173,770	157,283	157,772	3,520,353		

Source: WIPO (2010)

〈Table 2〉 Bibliographic information included in the US patent database

Code	Field Name	Code	Field Name
PN	Patent Number	IN	Inventor Name
ISD	Issue Date	IC	Inventor City
TTL	Title	IS	Inventor State
ABST	Abstract	ICN	Inventor Country
ACLM	Claim(s)	LREP	Attorney or Agent
SPEC	Specification[8]	AN	Assignee Name
CCL	Current U.S. Classification	AC	Assignee City
ICL	International Classification	AS	Assignee State
APN	Application Serial Number	ACN	Assignee Country
APD	Application Date	EXP	Primary Examiner
PARN	Parent Case Information[9]	EXA	Assistant Examiner
RLAP	Related US App. Data	REF	Referenced By
REIS	Reissue Data	FREF	Foreign References
PRIR	Foreign Priority	OREF	Other References
PCT	PCT[10] Information	GOVT	Government Interest
APT	Application Type		

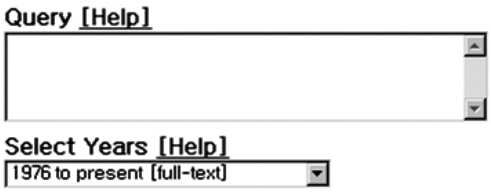
8) Detailed explanation about the usage of an invention.
9) Information about prior patents is included.
10) This is the acronym for the Patent Cooperation Treaty, according to which a patent filed in one country can be simultaneously acknowledged in the countries that have joined the treaty.

tered patents and applied ones pending for examination - from 1976 to present.[11] Among these two parts, most research has been conducted by employing the former set of data because a patent should be registered for it to be effective. We can search data in need through the "Advanced Search" menu by accessing the USPTO Web site. In this menu, we can see the window (Figure 1) through which we can type a search word and obtain the corresponding results.

Specifically, patent documents[12] can be searched though the window shown in Figure 1 by selecting the field of interest, and when the field and the word are connected with a slash (/) (e.g., if we intend to find a document whose patent number is 100, we have to type the search word, PN/100). In addition, if there are more than two fields of interest, logic operators such as AND and OR can be adopted (e.g., if we have the word "bicycle" in the patent title and the residence of an applicant is "Korea," we can use a search query like TTL/bicycle and ACN/KR). Next, we search for a patent document by inputting a keyword of interest in a certain field. A list of 50 patent documents is then displayed, as shown in Figure 2. The parts connected by hyperlinks are patent numbers and titles.

In this way, all the information that one US patent document contains can be seen in one HTML document by clicking the hyperlinks of patent numbers or those of patent titles shown in Figure 3. We can also draw an algorithm that can be applied according to the same procedure because all US patent documents follow the stylized format shown in Figure 3.

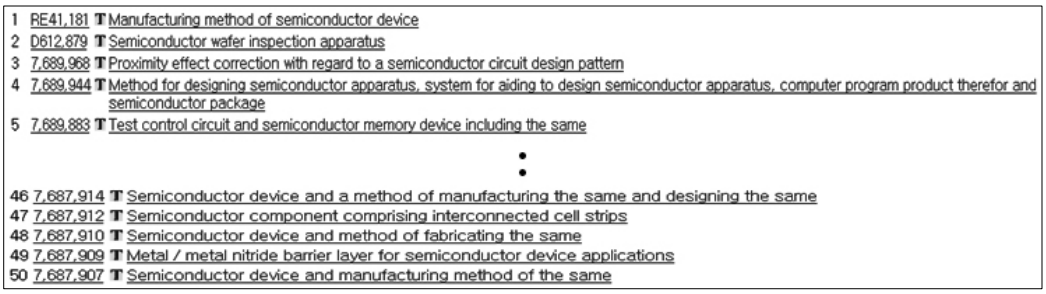
Finally, we can refer to a citation relationship between patents in the US patent documents which occurs in two directions. For example, if there exists a patent document A in this database, there also exist patent documents that patent A refers to. This is called "prior art." If a technology is complex, one patent can sometimes refer to over a hundred; and if a technology is simple, one patent can be filed based on one or two patents. If we draw a schematic diagram, we can obtain one similar to



〈Figure 1〉 Input window of a search query in the USPTO Web site with the time-period covered

11) A patent system is designed to grant a monopolistic right to a holder, and since it has to be given by the government, an "examination" process accompanies it. Patents before an "examination" are called "applied patents" while those after an "examination" are referred to as "registered patents."

12) The window shows the upper part of a patent document; in its lower part, additional fields such as "Referenced By," "Claims," and "Specification" can be seen.

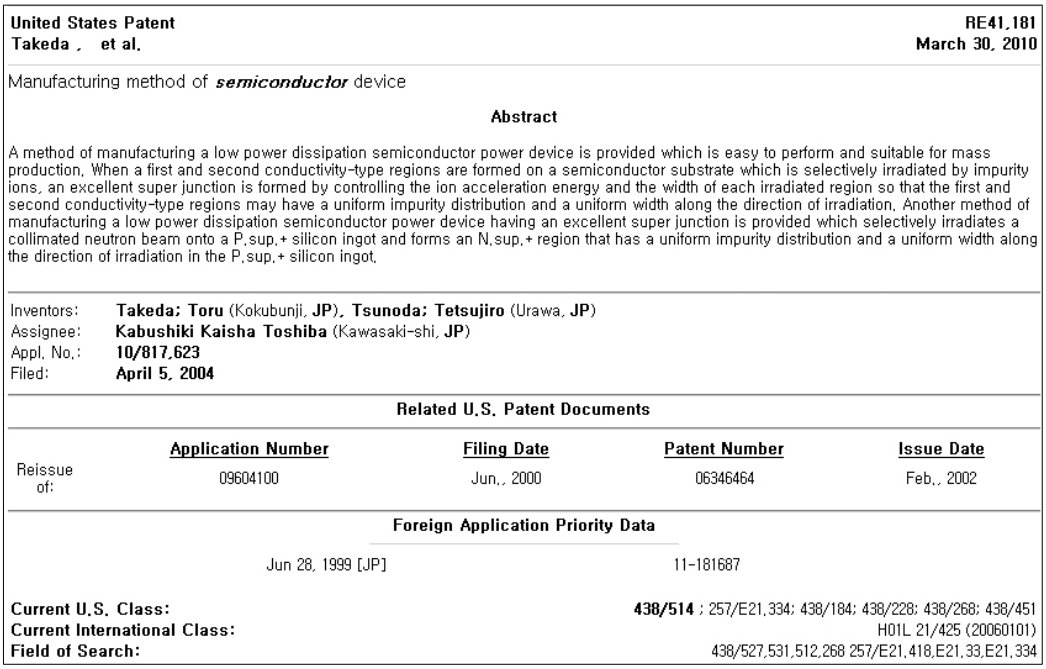


1 RE41,181 T Manufacturing method of semiconductor device
2 D612,879 T Semiconductor wafer inspection apparatus
3 7,689,968 T Proximity effect correction with regard to a semiconductor circuit design pattern
4 7,689,944 T Method for designing semiconductor apparatus, system for aiding to design semiconductor apparatus, computer program product therefor and semiconductor package
5 7,689,883 T Test control circuit and semiconductor memory device including the same

⋮

46 7,687,914 T Semiconductor device and a method of manufacturing the same and designing the same
47 7,687,912 T Semiconductor component comprising interconnected cell strips
48 7,687,910 T Semiconductor device and method of fabricating the same
49 7,687,909 T Metal / metal nitride barrier layer for semiconductor device applications
50 7,687,907 T Semiconductor device and manufacturing method of the same

〈Figure 2〉 A list of searched US patents



United States Patent **RE41,181**
Takeda , et al. **March 30, 2010**

Manufacturing method of ***semiconductor*** device

Abstract

A method of manufacturing a low power dissipation semiconductor power device is provided which is easy to perform and suitable for mass production. When a first and second conductivity-type regions are formed on a semiconductor substrate which is selectively irradiated by impurity ions, an excellent super junction is formed by controlling the ion acceleration energy and the width of each irradiated region so that the first and second conductivity-type regions may have a uniform impurity distribution and a uniform width along the direction of irradiation. Another method of manufacturing a low power dissipation semiconductor power device having an excellent super junction is provided which selectively irradiates a collimated neutron beam onto a P,sup,+ silicon ingot and forms an N,sup,+ region that has a uniform impurity distribution and a uniform width along the direction of irradiation in the P,sup,+ silicon ingot.

Inventors: **Takeda; Toru** (Kokubunji, **JP**), **Tsunoda; Tetsujiro** (Urawa, **JP**)
Assignee: **Kabushiki Kaisha Toshiba** (Kawasaki-shi, **JP**)
Appl. No.: **10/817,623**
Filed: **April 5, 2004**

Related U.S. Patent Documents

	Application Number	Filing Date	Patent Number	Issue Date
Reissue of:	09604100	Jun., 2000	06346464	Feb., 2002

Foreign Application Priority Data

Jun 28, 1999 [JP] 11-181687

Current U.S. Class: **438/514** ; 257/E21.334; 438/184; 438/228; 438/268; 438/451
Current International Class: H01L 21/425 (20060101)
Field of Search: 438/527,531,512,268 257/E21.418,E21.33,E21.334

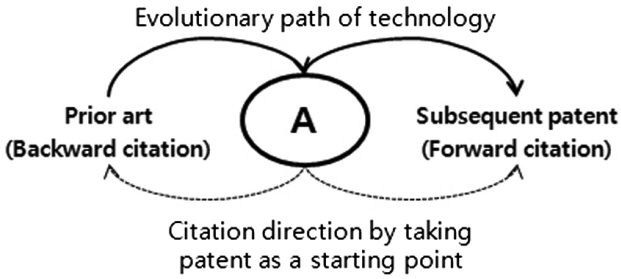
〈Figure 3〉 US patent document[13)]

the diagram shown in Figure 4.

As shown in this diagram, a technology steps on an evolutionary path it develops from a “prior art,” then to patent A, and finally to a “subsequent patent[14)]”, which is referred to as “backward citation.” It is termed “forward citation” when patent A is

13) It shows the upper part of on patent document, and in its lower part, we can see additional fields such as ‘Referenced By’, ‘Claims’, and ‘Specification.’

14) There exists a certain fixed word for previously issued patents, that is, “prior art”; however, there is no fixed term for subsequently issued patents.



〈Figure 4〉 Evolutionary path of one technology by taking patent A as a starting point and directions of citations

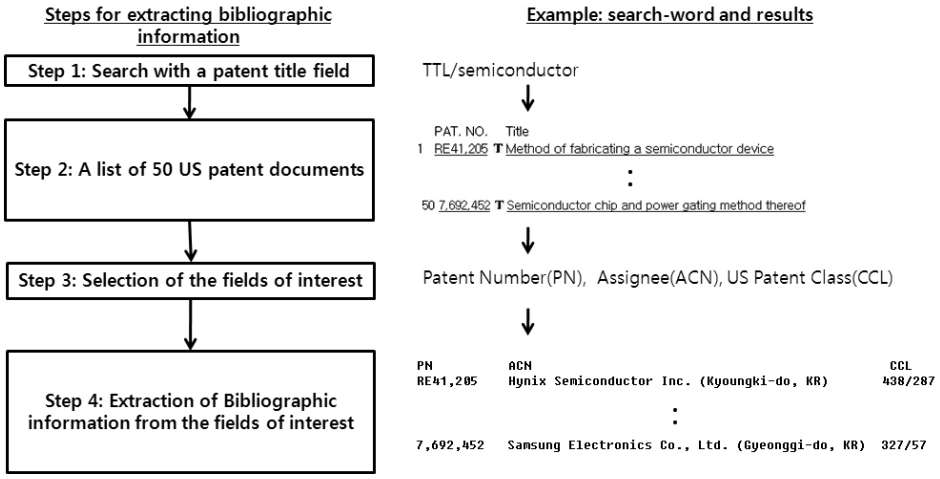
taken as an axis.[15] In the case of backward citation, US patent documents can be seen by clicking the corresponding patent numbers of titles. In the case of forward citation, however, we can refer to the US patent documents by clicking the hyperlink named "REF(Referenced By)." Therefore, in the latter case, we have to see the analogous screen, as shown in Figure 3, and click each patent number or title one by one.

3. A Method to Extract Specific Fields and Contents

As previously mentioned, in this research, we present extracting methods by exemplifying a method after selecting a specific technology of interest. To do this, we input a relevant keyword in the title field (TTL) and press the search button (Search), obtaining a list of 50 US patent documents. However, because this result simply shows patent numbers and title, as shown in Figure 2, the hyperlinks should be clicked on one by one in order to refer to the additional information on one patent document. To solve this problem, an algorithm by which we can extract the necessary fields directly from the US patent database is used as shown in Figure 5.

A procedural flow is established to convert this scheme into an algorithm form for programming, as shown in Table 3. That is, we can save the Web pages using a certain keyword and extract the required information from the HTML documents composed according to certain rules conducted in five steps. In the first step, we save the HTML documents of interest; in the second step, the patent numbers; in the third step, the assignees; and in the fourth step, US patent classes. Finally, we repeat this procedure 50 times because we only obtain a list of 50 US patent documents at one time. If we want to obtain a list of over 50 US patent documents, we shoul change the Web address in the first step until we finish the complete list.

15) If we take patent A as a base, it exhibits a reverse tracking pattern when we want to search for the previously issued patents. It also shows a forward tracking pattern when we look for the subsequently issued ones.



〈Figure 5〉 Steps for extracting bibliographic information in one patent document

〈Table 3〉 Extraction of bibliographic information form one patent: stepwise algorithm

Step	Algorithm
1	• Save the HTML documents of the US patents after searching a technology of interest
2	• By reading the document from the top, search the word "Title" and save the patent number following it
3	• By reading the document again, save the string following the word "Assignee"
4	• By continuously reading the document, save the class following the word "Current U.S. Class"
5	• Repeat 50 times by applying the same procedure in the next patent document with the pointer saved

Finally, the step-wise PHP[16] codes can be programmed as shown in Table 4. That is, in the first step, the searched Web address is stored as an array. In the second step, we allocate an array with a size of 10 by assuming that the US patent numbers, which range from 6 – 7, are inconsistent. In the third step, the string of the field "Assignee" in the newly created "AN" column is saved. In the fourth step, as the US patents documents are composed of "classes/subclasses," the three-digit numbers both from the left and right sides of the slash (/) are extracted. In the last step, steps 2 – 4 are repeated until no more pages exist.

However, the method proposed in Figure 5 has its limitation: it can only facilitate the extraction of the required information from the HTML documents using the hyperlinks only once. To extract information on subsequent patents simultaneously,

16) The PHP is an acronym for the Professional HTML Preprocessor contrived by Rasmus Lerrdorf and is widely used for the Web programming in that it enables an easy access to the Web databases and is distributed free (Lerdorf et al. 2006).

the procedure shown in Figure 6 should be conducted.

Next, a procedure can be developed by converting this scheme into an algorithm for programming. This algorithm involves more procedures than those in Table 5, with eight steps to follow. Steps 1 – 3 follow the same procedures as shown in Table 3. Steps 4 – 7 show the repetitive procedure required to extract the fields and contents of interest. Specifically, Step 4 represents the phase where the subsequent patent documents hyperlinked using the phrase "Referenced By" is saved. Steps 5 and 6 show the stages where the fields of interest, namely, patent numbers and the US patent classes, are extracted. Step 7 exhibits the level where the loop that deals with the subsequent patents is avoided, moving on to the original patent. Finally, in step 8, steps 2 – 7 are repeated based on a list of 50 US patent documents. The same process of changing the Web address should also be applied, as shown in Table 3.

Again, the step-wise PHP codes can be programmed as in Table 6. Steps 1 and 2 are the same as those shown in Table 2. In Step 3, the US patent classes are extracted using the same method applied in Table 5. However, in Step 4, we go one step further by extracting the required information from the subsequent patent documents using the HTML links. Next, similar to steps 2 and 3, the patent numbers of the subsequent patent documents in Step 5 and their US patent classes in Step 6 are extracted, respectively. In Step 7, after repeating this procedure based

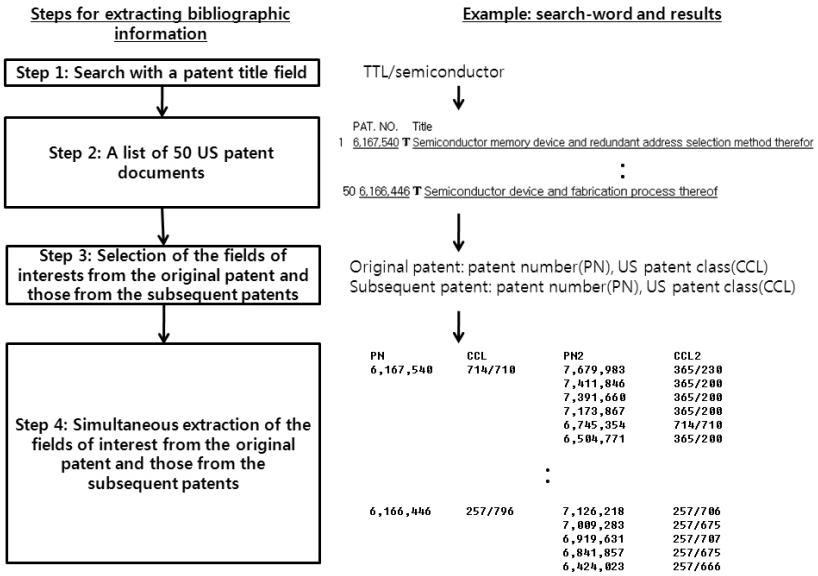
〈Table 4〉 Extraction of bibliographic information from one patent: step-wise PHP code

Step	PHP Code
Step 1	$page1=file("http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=0&p=1&f=S&l=50&Query=ttl%2Fsemiconductor%0D%0A&d=PTXT");
Step 2	$PN=substr($page1[$i],$end_idx-10,10);

Step 3	if(strpos($page2[$k],"Assignee:")){
	$k+=2;
	$st_an=strrpos($page2[$k],"〈B〉");
	$end_an=strrpos($page2[$k],")");
	$AN=substr($page2[$k],5,$end_an-4);
	$AN=str_replace("〈/B〉","",$AN);
	}

Step 4	if(!$pass && strpos($page2[$k],"Current U.S.Class:")){
	$k++;
	$st_ccl=strpos($page2[$k],"〈B〉");
	$end_ccl=strpos($page2[$k],"〈/B〉");
	$CCL=substr($page2[$k],$st_ccl+3,end_ccl-st_ccl-3);
	$pass=true;
	}

Step 5	$page1_end=count($page1);
	for($i=0; $i< $page1_end; $i++) {
	}
	$i++;



〈Figure 6〉 Steps for simultaneous extraction of the fields of interest from the original patent and those from the subsequent patents

on the number of original patents, we escape the loop and proceed to the list of original patents if there are no more subsequent patents left. Finally, we can obtain the final result after repeating the procedure above depending on the number of original patents.

〈Table 5〉 Simultaneous extraction of the bibliographic information from the original patent and subsequent patents: step-wise algorithm

Step	Algorithm
1	• Save the HTML documents of the US patents after searching a technology of interest
2	• By reading the document from the top, search the word "Title" and save the patent number following it
3	• By continuously reading the document, search the word "Current U.S. Class" and save the 'class' following it .
4	• Save the HTML documents of the subsequent patents
5	• By reading the documents of the subsequent patents from the top, search the word "Title" and save the patent number following it
6	• By continuously reading the documents of the subsequent patents, search the word "Current U.S. Class" and save the 'class' following it
7	• Repeat the same procedure depending on the number of the subsequent patents. If there is no more subsequent patents, go to the next stage
8	• Repeat 50 times by applying the same procedure in the next patent document with the pointer saved

〈Table 6〉 Simultaneous extraction of the bibliographic information from the original patent and the subsequent patents: step-wise PHP code

Step	PHP Code
Step 1	$page1=file("http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=0&p=1&f=S&l=50&Query=ttl%2Fsemiconductor%0D%0A&d=PTXT");
Step 2	$PN=substr($page1[$i],$end_idx-10,10);

Step 3	if(!$pass && strpos($page2[$k],"Current U.S. class:")){
	$k++;
	$st_ccl=strpos($page2[$k],"〈B〉");
	$end_ccl=strpos($page2[$k],"〈/B〉");
	$CCL=substr($page2[$k],$st_ccl+3,end_ccl-st_ccl-3);
	$pass=true;
	}

Step 4	if(strpos($page2[$k],"Referenced by"){
	$st_idx2=strpos($page2[$k],"Referenced by");
	$end_idx2=strpos($page2[$k],"[");
	$url2=substr($page2[$k],$st_idx2+27, $end_idx2-44);
	$page3=file("http://patft.uspto.gov".specialcharshtml($url2));
	$page3_end=count($page3);
	//echo "page3):http://patft.uspto.gov".$url2.": $page3_end〈br〉";
	for($1=0;$1〈$page3_end;$1++){
	$n=0;
	//echo $1.htmlentities($page3[$1])."〈br〉";
	if(strpos($page3[$1],"E〉Sin")){
	//echo "here";
	$1++;
	$st_idx3=strpos($page3[$1],"URL");
	$end_idx3=strrpos($page3[$1],"〉");
	$url3=substr($page3[$1],$st_idx3+4,end_idx3-st_idx3-5);
	$page4=file("http://patft.uspto.gov".specialcharshtml($url3));
	$page4_end=count($page4);
	//echo "page4):".$url3."〈br〉";

Step 5	$PN1=substr($page3[$m],$end_idx3-10,10);

Step 6	if(strpos($page4[$q],"Current U.S. Class:")){
	$q++;
	$st_ccl2=strpos($page4[$q],"〈B〉");
	$end_ccl2=strpos($page4[$q],"〈/B〉");
	$CCL2=substr($page4[$q],$st_ccl2+3,end_ccl2-st_ccl2-3);
	break;
	}

Step 7	else if(strpos($page3[$1],"PAT. NO.")){ }

Step 8	$page1_end=count($page1);
	for($i=0; $i〈 $page1_end; $i++) {
	}
	$i++;

4. Conclusion and Discussion

In this research, we aimed to propose a method to extract the timeliest bibliographic information from US patent documents, which contain the highest value among most technological documents. To this end, we demonstrated a method to extract the required bibliographic information directly from the USPTO Web site aside from using the Derwent Patent Database or the NBER database. The distinctive characteristic of the US patent documents is that they offer documented information through one HTML document composed of the corresponding hyperlinks, enabling us to extract the necessary information by implementing a simple algorithm. Therefore, in this study, we presented a method to extract the bibliographic information of interest by applying an HTML parsing technique.

The retrieved bibliographic information can be used in various ways. First, we can answer such questions as "What applicants (e.g., firms, universities, research institutes, government, etc.) hold what kinds of patents related to a certain technology?" with information on the "applicants" and "How many technologies are developing in specific technological areas (e.g., electronics, electrics, machines, etc.)?" with information on "US patent classes (see Table 7)."

By referring to the citation relationships of each patent's class, which has been drawn from the original patent and subsequent patents, we can grasp a technological trajectory that shows one technology affecting subsequent technologies (see No and Park 2010).

The method proposed in this research can be extended in various ways. First, analyzing the abstracts and claims with the "assignee" fields extracted is worthwhile; the former gives us information on inven-

〈Table 7〉 Examples of analysis based on the "applicants(left)" and "technological areas(right)"

Name	No. of patents	Name	No. of patents
Hynix Semiconductor	6	365: Static information storage and retrieval	17
Renesas Technology	5	257: Active solid-state devices	11
Kabushiki Kaisha Toshiba	5	324: Electricity: measuring and testing	4
NEC Electronics Corporation	4	326: Electronic digital logic circuitry	3
Elpida Memory	4	327: Miscellaneous active electrical nonlinear devices, circuits, and systems	3
Samsung Electronics	3	455: Telecommunications	2
Panasonic Corporation	3	716: Data processing: design and analysis of circuit or semiconductor mask	2
Others	20	Others	7
Total	50	Total	50

tions, whereas the latter encourages us to understand the mechanism of proprietary rights. As the US patent classes are analogous to industrial classification, we can measure the number of patents in their respective industries by extracting a large volume of the US patent classes and by making a concordance table about them. To develop this further, we can follow the pattern of inter-industrial knowledge flows by employing the citation relationship.

Although this paper contributes to research by presenting a method to extract the required information by directly connecting to the USPTO Web site, it has to be complemented from several perspectives. First, because the algorithm proposed above is based on the list of 50 US patent documents, we have to change the Web address for every list of 50 US patent documents. Second, we can only extract relatively small-sized bibliographic information, such as patent numbers, assignees, and US patent class through this algorithm. Extracting large-sized information, such as abstracts or claims, using this algorithm may be too time-consuming. Therefore, additional efforts to solve these intrinsic problems should be conducted in the future.

References

Calcagno, M. 2008. "An investigation into analyzing patents by chemical structure using Thomson's Derwent World Patent Index codes." *World Patent Information*, 30(3): 188-198.

Ernst, H. 2003. "Patent Information for Strategic Technology Management." *World Patent Information*, 25(3): 233-242.

Gupta, S., G. E. Kaiser, P. Grimm, M. F. Chiang, and J. Starren. 2005. "Automating Content Extraction of HTML Documents." *World Wide Web*, 8(2): 179-224.

Hall, B., A. B. Jaffe, and M. Trajtenberg. 2001. *The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools*. NBER Working Paper 8498.

Lerdorf, R., K. Tatroe, and P. MacIntyre. 2006. *Programming PHP* (2nd ed.). O'Reilly Media: Sebastopol, CA.

Lichtenthaler, U. 2009. "The role of corporate technology strategy and patent portfolios in low-, medium- and high-technology firms." *Research Policy*, 38(3): 559-569.

No, H. J. and Y. Park. 2010. "Trajectory patterns of technology fusion: Trend analysis and taxonomical grouping in nanobiotechnology." *Technological Forecasting and Social Change*, 77(1): 63-75.

Simmons, E. S. 2004. "The online divide: a professional user's perspective on Derwent database development in the online era." *World Patent Information*, 26(1): 45-47.

World Intellectual Property Organization (WIPO,

2010) *IP Statistics*.

Yoo, J. B. and Y. M. Chung. 2010. "Analysis of factors influencing patent citations." *Journal of the Korean Society for Information Management*, 27(1): 103-118.

Yoon, B. U. and Y. Park. 2004. "A text-mining-based patent network: Analytical tool for high-technology trend." *The Journal of High Technology Management Research*, 15(1): 37-50.