International Journal on Digital Libraries manuscript No.
(will be inserted by the editor)

Hidetsugu Nanba · Natsumi Anzen · Manabu Okumura

Automatic Extraction of Citation Information in Japanese Patent Applications

Abstract The need for academic researchers to retrieve patents and research papers is increasing, because applying for patents is now considered an important research activity. However, retrieving patents using keywords is a laborious task for researchers, because the terms used in patents for the purpose of enlarging the scope of the claims are generally more abstract than those used in research papers. Therefore, we have constructed a framework that facilitates patent retrieval for researchers, and have integrated research papers and patents by analysing the citation relationships between them. We obtained cited research papers in patents using two steps: (1) detection of sentences containing bibliographic information, and (2) extraction of bibliographic information from those sentences. To investigate the effectiveness of our method, we conducted two experiments. In the experiment involving Step 1, we prepared 42,073 sentences, among which a human subject manually identified 1,476 sentences containing citations of papers. For Step 2, we prepared 3,000 sentences, in which the titles, authors, and other bibliographic information were manually identified. We obtained a precision of 91.6%, and a recall of 86.9% in Step 1, and a precision of 86.2% and a recall of 85.1% in Step 2. Finally, we constructed an information retrieval system that provided two methods of retrieving research papers and patents. One method was retrieval by query, and another was from the citation relationships between research papers and patents.

Keywords Citation relationships · Information retrieval · Invalidity search · Scientometrics · Research paper · patent

1 Introduction

To appreciate the scope of a particular research field, retrieving both research papers and patents has become important for researchers in research fields with a high industrial relevance, such as bioscience, medical science, computer science, and materials science. However, retrieving patents using keywords is a laborious task for researchers, because the terms used in patents for the purpose of enlarging the scope of the claims are generally more abstract and more creative than those used in research papers. As a result, different patents tend to contain different terms, even though these terms refer to the same things. Moreover, it is often necessary for researchers to use patent classification codes, such as International Patent Classification (IPC) codes and F terms, for effective patent retrieval, but professional skills and abundant experience are also required. Therefore, we propose a method that enables researchers to retrieve patents without the need for any professional skills.

We have integrated research papers and patents by extracting the citation information from patents. For a given retrieved research paper, any related patents can also be found by tracing the citation relationships between the paper and any patents. Therefore, it is possible for researchers to retrieve patents, even though they may not have any special skills for retrieving patents.

Our integrated database is also useful for examiners in government patent offices, and for searchers in the intellectual property divisions of private companies. Their particular purpose is to carry out an "invalidity search" on existing patents or on research papers that can invalidate the patents of rival companies or patents under

Hidetsugu Nanba
Faculty of Information Sciences, Hiroshima City University
3-4-1 Ozukahigashi, Asaminamiku, Hiroshima 731-3194
JAPAN
Tel. & Fax: +81-82-830-1584
E-mail: nanba@its.hiroshima-cu.ac.jp

Natsumi Anzen
NEC System Technologies
1-40-1 Tomo-minami, Asaminamiku, Hiroshima 731-3168
JAPAN
E-mail: anzen-nxa@necst.nec.co.jp

Manabu Okumura
Precision and Intelligence Laboratory, Tokyo Institute of Technology
4259 Nagatsuta, Yokohama 226-8503 JAPAN
E-mail: oku@pi.titech.ac.jp

application in a national patent office. Patents that have many citations of papers are considered relevant to invalidity search, especially in detecting research papers that can invalidate the other patents.

Our system is also useful for researchers in scientometrics. In this research field, the citations between patents and research papers are used to analyse the influence of science to technology [13][18]. In general, if basic research in a domain strongly affects a technology, then it is assumed that the patents in that specific domain will cite many papers. Therefore, researchers measure the relevance of basic research to an industry by counting the number of cited research papers in patents in a given domain. Manually constructed citation databases are used in this type of analysis, but our automatically integrated database allows this type of analysis to be carried out efficiently.

The following four points need to be considered to integrate a patent and a research paper database,

1. Extraction of cited papers in a research paper.
2. Extraction of cited patents in a research paper.
3. Extraction of cited patents in a patent.
4. Extraction of cited papers in a patent.

In this work, we have integrated a Japanese patent database and a multilingual research paper database, "PRESRI"(**P**aper **RE**trieval **S**ystem using **R**eference **I**nformation)[1][14–16], which was constructed by collecting more than 78,000 PostScript and PDF files found on the Internet and extracting the bibliographic information from these files. This database also contains bibliographic from more than 346,000 cited papers that was extracted from the files. For each cited paper, one of the following citation types were automatically determined by analysing the context of the citations in the research papers using several cue phrases [14,15].

- Type B: Citations to other researchers' theories or methods.
- Type C: Citations comparing related work or to point out problems.
- Type O: Citations other than Types B and C above.

Using this database, Points 1 and 2 above are already resolved. However, it is impossible to apply their method to extract cited patents and papers in a patent, because citation styles in patents are much different from those in research papers, which we will describe in Section 3.2.2 in detail. Therefore, we focused on Points 3 and 4.

In general, Japanese patent specifications have the following structure:

- Invention title.
- Claims.
- Detailed description.
  - Field of the invention.
  - Prior art.
  - Means of solving the problems.
  - Embodiments of the invention.
  - Effects of the invention.
- Brief explanation of drawings.

In the "prior art" field, the author of the patent cites other related patents and/or papers. Therefore, we extracted citation information from the "prior art" field.

The remainder of this paper is organized as follows. Section 2 describes some related works. Section 3 explains the procedure used to integrate the research paper and patent database, and discusses how we investigated the effectiveness of our method by conducting some examinations. Section 4 discusses our experimental results, and Section 5 discusses the behaviour of our system. Finally, we provide our conclusions in Section 6.

2 Related Work

In this section, we describe some related works in "invalidity searches", "cross-genre information retrieval", and "scientometrics".

Invalidity searches

An invalidity search task was performed in the Patent Retrieval Task of the Fourth [3], the Fifth [4], and the Sixth [5] NII Test Collection for Information Retrieval (NTCIR) workshops. The goal of this task was to retrieve patents that could invalidate existing claims. Five groups with 21 systems participated in the Japanese retrieval subtask in the Sixth NTCIR, and the systems were evaluated using the Mean Average Precision (MAP). The best system obtained a MAP of 0.0815 [12]. The system analysed the structure of queries, and weighted terms in particular essential parts of the queries, using several weighted methods, such as the inverse document frequency (IDF) without a term frequency (TF) method.

In contrast to the task at the NTCIR, we aimed to retrieve both patents and research papers that could invalidate existing claims. By integrating both patent and research paper databases, we were able to construct a cross-genre retrieval environment.

Cross-genre information retrieval

There has been much research in the field of cross-genre information retrieval, such as that discussed in the technical survey task of the Patent Retrieval Task of the third NTCIR workshop [11]. This task aimed to retrieve patents relevant to a given newspaper article. In this task, Itoh et al. focused on "Term Distillation" [10]. The distribution of the frequency of the occurrence of words was considered to be different between heterogeneous databases. Therefore, unimportant words were assigned high scores when using TFIDF to weight words. Term Distillation is a technique that can prevent such cases by filtering out words that can be assigned incorrect weights. This idea was also used to link news articles and blog entries [9]. This is considered to be useful for integrating Japanese patent and research paper

[1] http://www.presri.com

databases. However, a machine translation technique is also required in our case, because most of the papers in PRESRI are written in English. Therefore, we integrated our patent database and PRESRI by analysing the citation relationships between these databases.

There have been several reports on the quality of citations between research papers and patents [1][21]. According to Schmoch [21], patent citations can be divided into two types: (1) documents of particular relevance, and (2) references concerning the general background. Schmoch reported that 29% of all citations can be classified as Type 1. He also showed in another paper that a large number of references can be linked to citing patents in the field of space technology in a very broad sense [20]. Although the ratio of citations in Type 1 is different for each domain, searchers are required to select relevant citations to some extent, when they collect documents by tracing citations. As it is difficult to identify the category of each citation automatically, the problem of cross-genre information retrieval is not fully resolved by analysing citation relationships alone. However, we still take an optimistic view of using citations, because they have been used successfully in scientometrics research, which is described below.

Scientometrics

Scientometrics is “the study of the measurement of scientific and technological progress”. One of the typical techniques used in scientometrics is to evaluate the productivity of individual researchers, organizations, and countries using bibliometric techniques, such as an impact factor. Another typical method used in scientometrics is to measure how basic science affects technologies in a given research field using the citation relationships between research papers and patents. These results can be used by governments or by private companies for allocating research funds to research areas that have high industrial relevance.

Many studies focusing on citations between papers and patents have been carried out. To observe the links between industry and basic science, Narin et al. traced the citations between patents and research papers, published in the USA, UK, Germany, Japan, and France, and showed both the domestic and international effect of science on technology [18]. Other related works in this field has been summarized by Meyer in a review article [13]. The citation databases in these analyses were constructed manually, but we have integrated patent and research paper databases automatically, which makes it possible to conduct this type of analysis more effectively.

3 Integration of Patent and Research Paper Databases

3.1 Procedure for the Integration of Patent and Research Paper Databases

We integrated the patent and research paper databases by citation, as research papers are usually cited in the prior art fields. Figure 1 shows an example of a citation of a research paper in the prior art field. Here, the serial number is shown at the beginning of each sentence as a reference. Among the three sentences shown in Figure 1, the paper is only cited in Sentence 3. To extract any bibliographic information from the cited papers in the prior art field, we must first extract such sentences (Step 1), and then identify any bibliographic information from them, such as the title or author (Step 2).

In the final step, we identified any duplicate publication number of the patent that had been extracted in the previous step. Then, the data from PRESRI and the extracted bibliographic information of the research papers carried out in the first step were gathered and identified as duplicate bibliographic information. Unlike patents, a research paper does not have an identification (i.e., publication) number. For this reason, it is necessary to compare titles, publication year, and other fields between two papers.

There are many related works for this task [6,8,7], but we used a method based on Nanba’s work to identify duplicate bibliographic information in papers [16]. The procedure used was as follows.

1. Eliminate particular marks, such as punctuation and hyphens, from both titles.
2. Compare the year of publication.
3. Calculate how similar two strings are using a technique called dynamic programming (also known as “edit distance”) [19].
4. Identify bibliographic information if the ratio exceeds a threshold value, and if the publication years match (if one or both publication years are not extracted in

[original]

(1) 従来，この種のオンライン文字認識装置は，高精度に文字を認識するために，入力パタンと標準パタンとの間でストロークまたは特徴点間の対応付けを行う必要がある．(2) この対応付けでは，手書きの変動に対しても頑健に認識するために，筆順や画数の制約を課したり，標準パタンを入力パタンへの重なりが最大となるように適応的整形を加えたりしている．(3) 従来のオンライン文字認識方法の**一例**が，1995年，若原徹他，ストローク単位のアフィン変換を用いたオンライン手書き漢字認識（電子情報通信学会技術報告書 PRU95-111 pp. 49-54）に記載されている．

[translation]

(1) Traditionally, this type of online character recognition system requires the assignment of strokes or feature points between an input pattern and a standard pattern for high accuracy character recognition. (2) In this assignment, stroke order and stroke count are used as the constraints for a robust recognition of various types of handwriting, and standard patterns are transformed so as to match an input pattern. (3) An **example** of a traditional online character recognition method is described in the following paper. 1995, Wakahara, et al. Online handwriting Kanji character recognition using affine transformation for each stroke (IEICE Technical Report PRU95-111, pp. 49-54).

Fig. 1 An example of a citation of a research paper in a prior art field

the extraction stage, then we considered that both years matched).

In the following subsection, we describe the first two steps.

3.2 Extraction of Cited Research Papers

3.2.1 Detection of Sentences Containing Bibliographic Information

In general, this type of sentence contains several useful cue phrases. For example, the citation of a paper often follows the phrase, “一例 (one example)”, such as in Sentence 3 in Figure 1, while “に記載 (described in)” follows a citation of a paper. The terms “Vol. (volume)”, “No. (number)”, and “pp. (page range)” are also useful. On the other hand, some phrases, such as “新聞 (newspaper article)” or “特許 (patent)”, seldom appear in this type of sentences. Therefore, we used such cue phrases to detect sentences that contained bibliographic information.

We manually analysed hundreds of prior art fields that were randomly selected from a patent database, and found that there were three types of cue phrases:

- **Positive cues**
  - **External cues** appeared before, or after a citation of a paper, e.g., “例えば (for example)”, “一例 (one such example)”, and “に記載 (described in)”.
  - **Internal cues** appeared in the citation of a paper, e.g., “pp.”, “Vol”, and “No.”
- **Negative cues** did not appear in sentences that cited research papers,
  e.g., “新聞 (newspaper article)” and “特許 (patent)”.

However, it is costly to form an exhaustive list of cue phrases manually, because the ratio of cited papers in a prior art field is not as high as that of cited patents. Nevertheless, the style of citations is different in patents compared with research papers. For example, the expressions, “pp.”, “p.”, and “pages” are used when the pages of cited patents are expressed in research papers. In addition to these expressions, “頁” (page) and “ページ” (page) are also used in patents. Unfortunatelly, we could not cover enough cue phrases to detect sentences containing bibliographic information from hundreds of prior art fields. Therefore, we collected cue phrases semi-automatically using three corpora: “PRESRI”, “PATENT”, and “PRIOR ART 1.”[2] The detail of these corpora are shown in Table 1.

To collect cue phrases efficiently, we carried out the following procedure using these corpora.

Table 1 Three corpora for collecting cue phrases

Name	Genre	Description	Language
PRESRI	Paper	346,000 bibliographic information	English / Japanese
PATENT	Patent	3,747,000 full text applications published in the 10 years between 1993 and 2002 (100GB)	Japanese
PRIOR ART 1	Patent	42,073 sentences with manually annotated tags.	Japanese
PRIOR ART 2	Patent	3,000 sentences with manually annotated tags	Japanese

- PRIOR ART 1 is the corpus for detection of sentences containing biliographic information.
- PRIOR ART 2 is the corpus for extraction of biliographic information.

[1. Initial selection step]

We extracted arbitrary length character strings in a “source of paper” item (usually a conference name or the title of a journal) in the PRESRI corpus to obtain candidates for the cue phrases to be included. Then, we manually selected the included cue phrases, such as “Proceedings of” and “Workshop of”, from these candidates.

[2. ML step]

We applied the Support Vector Machine (SVM) method to the “PRIOR ART 1” corpus. We extracted features from each sentence in the PRIOR ART 1 corpus, which included information on the existence of each cue phrase. We used a polynomial kernel with a degree = 2, and obtained a classifier that identified sentences containing cited papers. Here, we used the existence of cue phrases as a feature of our machine learning.

[3. Iterative selection step]

We extracted sentences from the PATENT corpus using the classifier. Then, we looked for other cue phrases from the sentences. When we found new cue phrases, we returned to **the ML step** discussed above.

By repeating the above steps until no more cue phrases were obtained, we finally obtained 14 external cues, 22 internal cues, and two negative cues. We also obtained an SVM-based sentence extractor using these cue phrases. The Appendix shows a list of these cue phrases.

3.2.2 Extraction of Bibliographic Information

We used information extraction based on machine learning to extract any bibliographic information, such as the title or authors, from the sentences extracted in the previous step discussed above. In the following subsection, we define the tags used in our examination, and then describe our extraction method.

[2] We released these data to the research groups that had entered into a license agreement with the NTCIR-6 test collection of the National Institute of Informatics (NII), Japan. Please refer to the NTCIR Web Page for details (http://ntcir.nii.ac.jp). In this data set, the tags were annotated by a single human analyst, and were checked a different analyst.

Tag Definition

We defined a tag set for the bibliographic information in the prior art field as follows.

- **E-AUTHORS** and **J-AUTHORS** included the authors' names written in English and Japanese, respectively. If a paper had multiple authors, then the AUTHORS tags were also added before the first author and after the last author.
- **E-TITLE** and **J-TITLE** included the title of the paper written in English and Japanese, respectively. The TITLE tag excluded any quotation marks.
- **E-SOURCE** and **J-SOURCE** included the source of the paper written in English and Japanese, respectively. The tag included the title of any conference proceedings, volume, number, publisher, or URL.
- **DATE** included the publication year. The DATE tags could also include the month or day (e.g., "September 2003").
- **PAGE** included the page range of a research paper (e.g., "pp. 1–8", "p. 23", "2138–2152").
- **OTHER** included other letter strings (e.g., "to appear").

A tagged example is given below.

[original]
<OTHER> この論理マクロ展開方法は </OTHER><J-SOURCE> 第２３回デザイン オートメーション カンファレンス予稿集 </J-SOURCE><PAGE> 第５９４頁から第６００頁 </PAGE><OTHER>（</OTHER><DATE> １９８６年 </DATE><OTHER>）(</OTHER><E-SOURCE>Proc. of ２３ rd DAC</E-SOURCE><OTHER>, </OTHER> <PAGE>pp. ５９４－６００ </PAGE><OTHER>）において記載されている。</OTHER>
[translation]
<OTHER>This logic macro expansion method was described in </OTHER> <J-SOURCE>proceedings of the 23^{th} Design Automation Conference </J-SOURCE> <PAGE> pages 594-600</PAGE> <OTHER>(</OTHER> <DATE>1986</DATE> <OTHER>) (</OTHER> <E-SOURCE> Proc. of 23^{rd} DAC </E-SOURCE> <OTHER>, </OTHER> <PAGE> pp. 594-600 </PAGE> <OTHER>).</OTHER>

Fig. 2 An example of a manually tagged prior art field

Information extraction based on machine learning

There have been several studies on extracting bibliographic information from lists of references in research papers. Most of these were based on Hidden Markov Models (HMMs) [2] [22]. However, we did not use HMMs in our study, because the state transition model of HMMs for patent applications is more complex than that for research papers. For example, a title or an author name(s) is usually written at the beginning of each bibliographic information string. On the other hand, patent applications do not follow this pattern, as shown in Figure 1. To confirm that citations in patents are more complex than those in papers, we applied the state transition model for papers [16] shown in Figure 3 to citations in patents. Here, we ignored DATE and PAGE tags, because these were not contained in the transition model. As a result, we found that 62.2% of citations in patents were not accepted using this model. [3]

AUTHORS-TITLE-SOURCE
AUTHORS-SOURCE
AUTHORS-TITLE
TITLE-SOURCE
SOURCE
TITLE

Fig. 3 State transition models for research papers [16]

Nanba et al. [16] proposed a method for extracting bibliographic information from lists of references in research papers based on the SVM method. They compared the SVM-based method with the HMM-based method, and confirmed experimentally that the SVM-based method was superior to the HMM-based method. In the same way, we also used an SVM method. As a machine learning method, we also examined the Conditional Random Fields (CRF) method, whose empirical success has been reported recently in the field of natural language processing.

Both the SVM- and CRF-based methods assign a class to each word. Features and tags are given in the SVM and CRF methods as follows: (1) the k tags occur before a target word, (2) k features occur before a target word, and (3) k features follow a target word (Figure 4). We used values of $k = 2$ and $k = 5$ for the CRF and SVM methods, respectively, which were determined in a pilot study. Here, we use the following features for machine learning: a word, its part of speech, and whether the word was an internal cue, as described in Section 3.2.1.

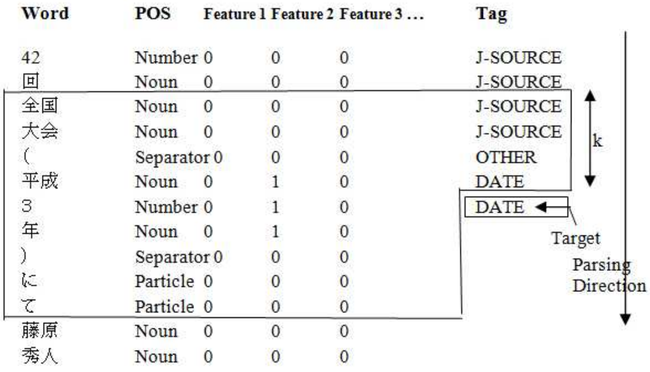

Word	POS	Feature 1	Feature 2	Feature 3 …	Tag
42	Number	0	0	0	J-SOURCE
回	Noun	0	0	0	J-SOURCE
全国	Noun	0	0	0	J-SOURCE
大会	Noun	0	0	0	J-SOURCE
(	Separator	0	0	0	OTHER
平成	Noun	0	1	0	DATE
3	Number	0	1	0	DATE
年	Noun	0	1	0	
)	Separator	0	0	0	
に	Particle	0	0	0	
て	Particle	0	0	0	
藤原	Noun	0	0	0	
秀人	Noun	0	0	0	

Fig. 4 Features and tags given to the SVM and the CRF

[3] There were 65 state transition patterns in patents.

4 Experiments

To investigate the effectiveness of our method, we conducted the following two examinations: (1) detection of sentences containing bibliographic information, and (2) extraction of bibliographic information.

4.1 Detection of Sentences Containing Bibliographic Information

Data sets

We used the “PIRIOR ART 1” corpus, which was described in Table 1 in Section 3.2.1. We manually assigned “PAPER” tags to 42,073 sentences that contained bibliographic information. Among these sentences, we used 32,537 sentences for training (among these, the “PAPER” tags were assigned to 1,186 sentences), and 9,536 sentences for testing (among these, the “PAPER” tags were assigned to 290 sentences).

Alternatives

We conducted experiments using the following three methods:

- **External cue (a baseline method)**, for extracting all sentences containing an external cue.
- **Internal cue (a baseline method)**, for extracting all sentences containing an internal cue.
- **Internal and external cues (a baseline method)**, for extracting all sentences containing both internal and external cues.
- **Our method**, for extracting sentences based on the SVM, which used external and internal cues as features.

Machine learning

We used the TinySVM[4] software package as our SVM learning package. We employed a polynomial kernel with a degree = 2, which was defined by the following equation:

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d \qquad (1)$$

Evaluation method

We used the following measures for evaluation.

$$Recall = \frac{\begin{array}{c}\textit{The number of correctly}\\ \textit{extracted sentences}\end{array}}{\begin{array}{c}\textit{The number of sentences}\\ \textit{that should be extracted}\end{array}} \qquad (2)$$

$$Precision = \frac{\begin{array}{c}\textit{The number of correctly}\\ \textit{extracted sentences}\end{array}}{\begin{array}{c}\textit{The number of sentences}\\ \textit{that the system extracts}\end{array}} \qquad (3)$$

[original]
従来，カラー画像処理装置における，階調表現に用いられる誤差拡散処理の例としては，Floyd-Steinbergによる文献“An Adaptive for Spatial Grey Scale”が知られている．
[translation]
An article “An Adaptive for Spatial Grey Scale” by Floyd-Steinberg is known as an example of Error Diffusion Method in colour image processing devices.

Fig. 5 An example sentence that our method could not detect

Experimental results and discussion

Table 2 Experimental results for detecting sentences containing bibliographic information

	Precision	Recall
External cue (baseline)	0.079 (282/3555)	0.972 (282/290)
Internal cue (baseline)	0.239 (267/1119)	0.921 (267/290)
External and internal cues (baseline)	0.261 (260/998)	0.897 (260/290)
Our method	0.916 (252/275)	0.869 (252/290)

Our results are shown in Table 2. The recall scores were almost the same among the three methods. The precision score of the “Internal cue” was better than that of the “External cue”. This indicates that internal cues are more useful than external cues. A precision score of “Internal and external cues” is better than “Internal cue” and “External cue”. Our method obtained the best precision score. It is considered that the SVM method can be used to optimize the combination of internal and external cues.

In Figure 5, we show a typical example of a sentence that our method could not detect. Most of such sentences did not contain any sources of papers, because most of the internal cues were related to the sources of the papers, such as “Vol.”, “No.”, and “pp.” However, this is not a serious problem, because such bibliographic information is not integrated in PRESRI due to the lack of bibliographic information.

4.2 Extraction of Bibliographic Information

Machine learning

As shown in Figure 4, both the SVM- and CRF-based methods assign one of the nine tags (classes) described in Section 3.2.2 to each word. We used the pairwise classification method and a polynomial kernel with a degree = 2 for classifying multiclasses using the SVM binary classifier. We used the YamCha[5] software package, which specializes in text chunking based on TinySVM software. As another example of a machine learning method, we also used the CRF++[6] software as a CRF learning package.

[4] http://chasen.org/˜taku/software/TinySVM

[5] http://cl.aist-nara.ac.jp/˜taku/software/yamcha/

[6] http://www.chasen.org/˜taku/software/CRF++/

Data sets

We used the “PIRIOR ART 2” corpus shown in Table 1 described in Section 3.2.1. We manually assigned tags, using the process described in Section 3.2.2, and obtained 3,000 tagged sentences. Then, we performed a ten-fold cross validation test.

Evaluation method

We used Recall and Precision for evaluation. We considered an entry correct if the alphabet, number, and hiragana, katakana, and kanji characters (Japanese text only) in the system output matched those in the correct data. An example of our evaluation is shown in Table 3. In this example, both E-SOURCE and DATE data in the system output are correct, because the difference between the correct data and the system output were the separators ‘ “ ’, ‘ ” ’, and ‘)’.

Experimental results and discussion

Table 4 Experimental results obtained by extracting bibliographic information from citations of prior art

		YamCha		CRF++	
		Precision	Recall	Precision	Recall
English	Author	0.720	0.787	0.872	0.857
	Source	0.752	0.787	0.805	0.799
	Title	0.763	0.901	0.746	0.903
Japanese	Author	0.742	0.715	0.885	0.765
	Source	0.733	0.662	0.834	0.736
	Title	0.868	0.880	0.848	0.881
Page		0.941	0.932	0.973	0.973
Date		0.897	0.897	0.932	0.921
Average		0.802	0.822	0.862	0.854

The results obtained are shown in Table 4. As can be seen from the data in Table 4, the performance of the CRF++ package was better than that of the YamCha package. Among the eight types of field, “Source” and “Title” performed worse using the CRF++ method.

Among the results of the CRF++ technique, the extraction of English titles performed the worst. An example of such cases is shown in Figure 6. In this example, the title of the paper is not shown. Instead, both the conference name (Automated IC Manufacturing) and the source (“ECS Fall Meeting”, Vol. 88-2, p. 566(1988)) are shown in different parts. In this case, our method mistook “Automated IC Manufacturing” as a title. We observed many other similar examples, when expressions, such as “Society of” or “Association” were not contained in the sources.

We also show other typical errors in Figure 7. In this example, both the title (“光コンピューティング (optical computing)”) and the source (“１９８９年秋期，第５０回応用物理学会学術講演会 (the 50th Annual Meeting of Japan Society of Applied Physics, 1989 in the fall)” and “ＪＳＡ ｃａｔ－ｎｏ： ＡＰ８９１２３２”) are shown, but our method mistook the “Source” field (underlined in Figure 6) for the “Title” field. This is because both the “Source” and “Title” fields consisted of long character strings without punctuation and symbols. As such, both fields superficially resembled each other.

[original]
例えば，ＬＳＩの製造ラインの自動化について，米国電気化学会（Electrochemical Society）主催の国際会議「Automated IC Manufacturing」において報告されている（“ECS Fall Meeting”, Vol. 88-2, p. 566(1988)）
[translation]
For example, there is an report on the automation of LSI manufacturing lines (“ECS Fall Meeting”, Vol. 88-2, p. 566(1988)) in the international conference “Automated IC Manufacturing” hosted by the Electrochemical Society (Electrochemical Society)

Fig. 6 An example sentence where our method mistakenly extracted a “Source” field as a “Title” field (Type 1)

[original]
光コンピュータ素子や光ニューロ素子が研究開発されており，その詳細については下記に示す１９８９年秋期，第５０回応用物理学会学術講演会 シンポジウムダイジェスト,「光コンピューティング」（ＪＳＡ ｃａｔ－ｎｏ： ＡＰ８９１２３２）に述べられている
[translation]
There has been much research into optical computer devices and optical neuro devices, the details of which will be reported in the 50th Annual Meeting of Japan Society of Applied Physics, 1989 in the fall symposium digest “optical computing” (JSA cat-no: AP891232))

Fig. 7 An example sentence where our method mistakenly extracted the “Source” field as the “Title” field (Type 2)

To improve these errors, we are now considering using lists of journal titles and conference names as one of the features for machine learning.

4.3 Discussion

Effects of two-steps bibliographic information extraction

We detected cited research papers in patents in two stages: (1) detection of sentences containing bibliographic information, and (2) extraction of bibliographic information from those sentences. To confirm the effectiveness of our two step extraction method, we compared it with a method that did not contain Step 1 (a one-step method).

In the two-step method, we applied two machine learning methods, SVM and CRF, to the sentences containing cited papers. In the one-step method, instead of using only these sentences containing cited papers, we applied the machine learning methods to all the sentences in the PRIOR ART 1 corpus, which we described in Section 3.2.1.

The results are shown in Table 5. As can be seen from the data in Table 5, the results using our method were much better than those obtained using the one-step method. This indicates that our two-step method is effective for detecting cited research papers in patents.

Processing speed

To extract the full bibliographic information from the 100 GB of the PATENT corpus, we needed to operate

Table 3 Example of evaluation

Correct			E-SOURCE		DATE		PAGE	
String (Translation)	例えば， (For example)	“	Solid state technology	”	(May 1990	)	P149-154	がある。 (exists)
System		E-SOURCE			DATE		PAGE	

Table 5 Comparison of the two-step and one-step extraction methods

		our method (Step 1 and Step 2)		One-step method (Step 2)	
		Precision	Recall	Precision	Recall
English	Author	0.872	0.857	0.539	0.333
	Source	0.805	0.799	0.522	0.354
	Title	0.746	0.903	0.571	0.500
Japanese	Author	0.885	0.765	0.350	0.389
	Source	0.834	0.736	0.488	0.406
	Title	0.848	0.881	0.474	0.450
Page		0.973	0.973	0.760	0.702
Date		0.932	0.921	0.606	0.664
Average		0.862	0.854	0.539	0.475

using the full sentences in the corpus. To conduct this process quickly, we divided the corpus into 10 parts, and ran the assessment using 10 processors. As a result, we extracted the entire bibliographic information within a period of one hour.

Experimental results of our bibliographic information extraction

On considering the complexity and the difficulty of our task, our experimental results are encouraging. In a previous stidy, Nanba et al. [16] obtained an accuracy of 87.4% in experiments extracting bibliographic information from lists of references. Compared with their task, our task was more complex, because the bibliographic information was embedded in natural language sentences in our task. Nevertheless, our results are close to Nanba's figures.

5 System Overview

In this section, we will introduce the overview of our system.

5.1 Data of the Citation Relationships between Patents and Research Papers

Using our method, we extracted 86,415 cited papers from 3,496,253 Japanese patent applications published in the 10 years between 1993 and 2002. To grasp the outline of the data, we classified these cited papers using IPC codes of each citing patent. The IPC system is a global standard hierarchical patent classification system that contains 7,314 main groups, and at least one IPC code is manually assigned to each patent application.

By counting the number of cited papers for each IPC code, we can measure the relevance of basic research to the technology in each domain. The numbers of cited papers for each IPC code are shown in Table 6. Our system was used effectively in the fields shown in Table 6.

Table 6 The number of cited papers for each domain

IPC code	description	the number of cited papers
C12N	enzyme, biogenetics	12160
H01L	semiconductor, superconduction	10442
A61K	medicine, organic compound	8834
C12R	alcohol, microorganism	8656
C12P	microbial production of compounds	8282
G06F	computer, memory, programming	7604
C07D	penicillin compound	7434
C07C	low-molecular-weight compound	6825
G01N	material science	5553
H04N	TV, fax, video tape	4841

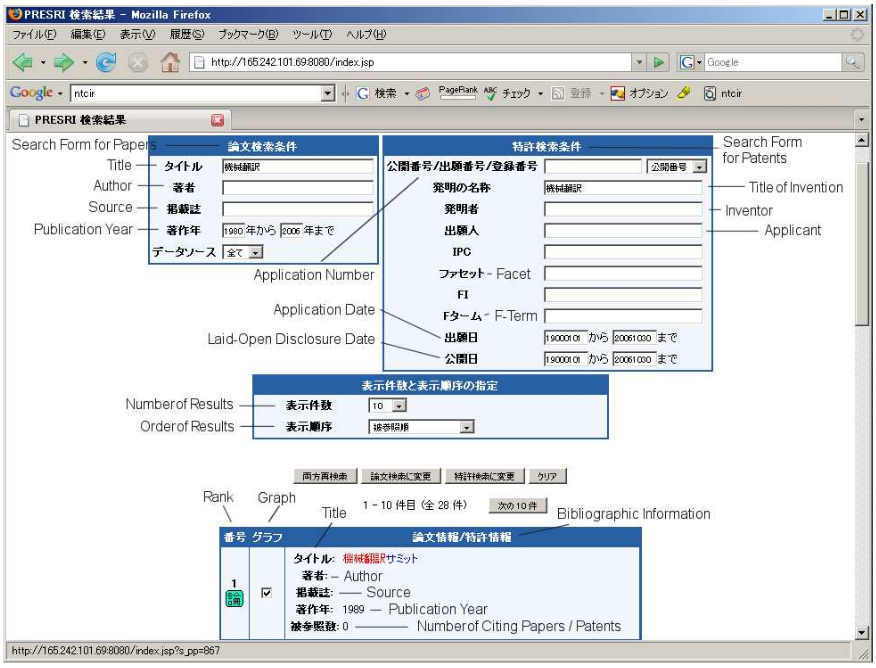
5.2 System Behaviour

We integrated cited papers described in Section 5.1 with PRESRI using the procedure outlined in Section 3.1, and constructed a system that enables us to retrieve both research papers and patents by key phrases and by citations.

Figure 8 shows the search results using the key phrase “機械翻訳 (machine translation)”. The checked boxes are shown at the head of each result. If the user checked the boxes of relevant papers and selected the “display a citation graph” command at the bottom of the page, then PRESRI showed the selected papers along with some related papers and patents as a visual output, as shown in Figure 9.

In Figure 9, the dots, squares, and arrows denote the papers, patents, and citation relationship between documents, respectively. The title of the paper was shown in a pop-up window [17] if the user placed the cursor over a particular dot (i.e., paper) or a square (i.e., patent). If the user paused the cursor for a period of more than a second, then the author(s) and an abstract of the paper were shown along with the title. The citation area was shown in a pop-up window if the user placed the cursor over an arrow. From the citation graph of research papers and patents shown in Figure 9, the user is able to understand the progress and transition of specific field studies at a glance. This type of information is helpful for an efficient survey of a field of study.

Our system may also be useful for various scientometrics studies. For example, the data in the “number of cited papers for each domain” shown in Table 8 indicate the research fields that have high industrial relevance. These data can be used when a government or a private company makes decision for the allocation of research funds to a particular field. Another example is in identifying important fundamental research from an industrial viewpoint using bibliometric techniques.



論文検索条件

タイトル 機械翻訳
著者
掲載誌
著作年 1980 年から 2006 年まで
データソース 全て

特許検索条件

公開番号/出願番号/登録番号 公開番号
発明の名称 機械翻訳
発明者
出願人
IPC
ファセット
FI
Fターム
出願日 19000101 から 20061030 まで
公開日 19000101 から 20061030 まで

表示件数と表示順序の指定

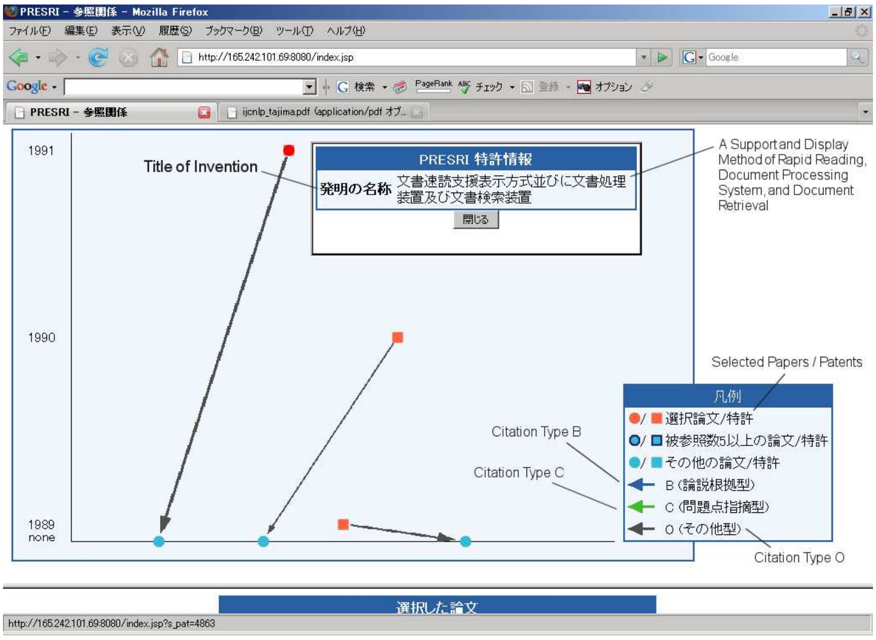
表示件数 10
表示順序 被参照順

1 – 10 件目 (全 28 件)

番号	グラフ	論文情報/特許情報
1 論	☑	タイトル: 機械翻訳サミット 著者: 掲載誌: 著作年: 1989 被参照数: 0

http://165.242.101.69:8080/index.jsp?s_pp=867

Fig. 8 Search results using a key phrase


PRESRI – 参照関係 – Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://165.242.101.69:8080/index.jsp
Google
検索 PageRank チェック 登録 オプション
PRESRI – 参照関係
ijcnlp_tajima.pdf (application/pdf オブ...
1991
1990
1989
none
Title of Invention
A Support and Display Method of Rapid Reading, Document Processing System, and Document Retrieval
閉じる
Selected Papers / Patents
Citation Type B
Citation Type C
Citation Type O
凡例
●/■ 選択論文/特許
○/□ 被参照数5以上の論文/特許
●/■ その他の論文/特許
← B (論説根拠型)
← C (問題点指摘型)
← O (その他型)
選択した論文

PRESRI 特許情報

発明の名称 文書速読支援表示方式並びに文書処理装置及び文書検索装置

http://165.242.101.69:8080/index.jsp?s_pat=4863

Fig. 9 Citation relationships between research papers and patents

6 Conclusions

We have integrated research papers and patents by analysing the citation relationships between them. In this work, we focused on the detection of cited research papers in patents in two ways: (1) detection of sentences containing bibliographic information, and (2) extraction of bibliographic information from those sentences. We obtained a precision of 91.6% and a recall of 86.9% in Step 1, and a precision of 86.2% and a recall of 85.1% in Step 2.

Our task was more complex than Nanba's task of extracting bibliographic information from lists of references [16], because the bibliographic information was embedded in natural language sentences in our task. Nevertheless, our experimental results almost reached the 87.4%, accuracy score obtained in Nanba's experiments. Therefore, we consider that our results are encouraging.

In future work, we need to improve our method of identifying duplicate bibliographic information between research papers and cited papers in patents. As we described in Section 3, there are many related works or systems for this task [6,8,7].

We will also need to study identification of the types of patent citations. As Schmoch reported [21], there are at least two types of patent citations: (1) documents of particular relevance, and (2) references concerning the general background. Automatically identifying the former type of citations is required if our system is to be used for invalidity searches and scientometric research. There has been related research into identifying types of citations in research papers [14][23]. This research may provide clues to identify types of patent citations.

Appendix

Here, we show the list of cue phrases used for the detection of sentences containing bibliographic information.

Positive cues

– **External cues** appeared before, or after a citation of a paper, e.g., “例えば (for example)”, “文献 (article)”, “として (as)”, “ 記載 (written)”, “開示 (disclosed)”, “述べられている (described)”, “参照 (referred)”, “記述 (described)”, “発表 (published)”, “ 紹介 (introduced)”, “提案 (proposed)”, “知られている (is known)”, “示されている (is presented)”, “論文 (article)” “一例 (one such example)”, and “に記載 (described in)”.

– **Internal cues** appeared in the citation of a paper, e.g., “論文誌 (journal)”, “大会 (meeting)”, “学会 (association)”, “proceedings of”, “journal”, “workshop”, “conference”, “university”, “international”, “symposium”, “transaction”, “letters”, “pp.”, “p.”, “vol.”, “ 巻 (vol.)”, “号 (no.)”, “no.”, “ページ (pages)”, “(19|20)** (four digits of numbers)”, “年 (year)”, “ “ ”, “「”, “」”, “[”, “]”, and “)”

Negative cues did not appeared in sentences that cited research papers,
e.g., “新聞 (newspaper article)” and “特許 (patent)”. In other words, sentences that contained negative cues should not be detected.

References

1. Baré, R.: Results of a Statistical Study of the References Cited in the Search Reports Established by the EPO. World Patent Information, **Vol. 3, No. 2**, 56–60 (1981)
2. Borkar, V., Deshmukh, K., Sarawagi, S.: Automatic Segmentation of Text into Structured Records. In Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, 175–186 (2001)
3. Fujii, A., Iwayama, M., Kando, N.: Overview of Patent Retrieval Task at NTCIR-4. In Working Notes of the 4th NTCIR Workshop, 225–232 (2004)
4. Fujii, A., Iwayama, M., Kando, N.: Overview of Patent Retrieval Task at NTCIR-5. In Proceedings of the 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, 269–277 (2005)
5. Fujii, A., Iwayama, M., Kando, N.: Overview of the Patent Retrieval Task at NTCIR-6 Workshop. In Proceedings of the 6th NTCIR Workshop Meeting, 359–365 (2007)
6. Galhardas, D., Florescu, D., Shasha, D.: AJAX: An Extensible Data Cleaning Tool. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 590 (2000)
7. Giles, C.L., Bollacker, K., Lawrence, S.: An Automatic Citation Indexing System. In Proceedings of the 3rd ACM International Conference on Digital Libraries, 89–98 (1998)
8. Hitchcock, S., Carr, L., Harris, S. Hey, J.M.N., Hall, W.: Citation Linking: Improving Access to Online Journals. In Proceedings of the 2nd ACM International Conference on Digital Libraries, 115–122 (1997)
9. Ikeda, D., Fujiki, T., Okumura, M.: Automatically Linking News Articles to Blog Entries. In Proceedings of AAAI Spring Symposium Series Computational Approaches to Analyzing Weblogs, 78–82 (2006)
10. Itoh, H., Mano, H., Ogawa, Y.: Term Distillation for Cross-db Retrieval. In Proceedings of Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task, 11–14 (2002)
11. Iwayama, M., Fujii, A., Kando, N., Takano, A.: Overview of Patent Retrieval Task at NTCIR-3. In Proceedings of Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task, 1–10 (2002)
12. Mase, H., Iwayama, H.: NTCIR-6 Patent Retrieval Experiments at Hitachi. In Proceedings of the 6th NTCIR Workshop, 403–406 (2007)
13. Mayer, M.: Does Science Push Technology? Patents Citing Scientific Literature. Research Policy, **Vol. 29**, 409–434 (2000)
14. Nanba, H., Okumura, M.: Towards Multi-paper Summarization Using Reference Information. In Proceedings of the 16th IJCAI, 926–931 (1999)
15. Nanba, H., Kando, N., Okumura, M.: Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation. In Proceedings of the American Society for Information Science / the 11th SIG Classification Research Workshop, Classification for User Support and Learning, 117–134 (2000)
16. Nanba, H., Abekawa, T., Okumura, M., Saito, S.: Bilingual PRESRI: Integration of Multiple Research Paper Databases. In Proceedings of RIAO 2004, 195–211 (2004)
17. Nanno, T., Saito, S., Okumura, M.: Zero-Click: a System to Support Web Browsing. The 11th International World Wide Web Conference (2002)
18. Narin, F., Olivastro, D., Stevens, K.A., Bibliometrics/Theory, Practice and Problems, Evaluation Review, **Vol. 18, No. 1**, 65–76 (1994)
19. Needleman, S.B., Wunsch, C.D.: A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. Journal of Molecular Biology, **Vol. 48**, 443–453, (1970)
20. Schmoch, U., Kirsch, N., Lay, W., Plescher, E., Jung, K.O.: Analysis of Technical Spin-off Effects of Space-related R&D by Means of Patent Indicators. Acta Astronautica, **Vol. 24**, 353–362 (1991)
21. Schmoch, U.: Tracing the Knowledge Transfer from Science to Technology as Reflected in Patent Indicators. Scientometrics, **Vol.26, No. 1**, 193–211 (1993)
22. Takasu, A.: Bibliographic Attribute Extraction from Erroneous Reference based on a Statistical Model. In Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries 2003, 49–60 (2003)
23. Teufel, S., Moens, M.: Summarizing Scientific Articles–Experiments with Relevance and Rhetorical Status. Computational Linguistics, **Vol. 28, No. 4**, 409–445 (2002)