特許の文書構成と分類情報を用いた類似特許検索方式の精度評価

NTCIR－6特許検索タスクにおける取り組み紹介

株式会社日立製作所　システム開発研究所
間瀬　久雄

PROFILE 平成2年（株）日立製作所入社、システム開発研究所に配属。以来、特許や新聞記事、Webページ等を対象とした分類自動付与、文書検索、文章要約、テキストマイニング等の日本語処理の研究に従事。

hisao.mase.qw@hitachi.com

1 はじめに

発明を無効化する過去の類似特許を検索する際には、発明内容を端的に表すキーワードから構成される論理式（AND/OR/NOT）を作成し、この論理式を満たす特許を検索するという方法が一般的である。適切な論理式を作成して所望の特許を高精度かつ効率良く検索するためには、発明内容およびその技術分野に関する知識だけでなく、特許の書き方・読み方に関する知識や、検索そのものに関するノウハウが必要不可欠である。しかし、これらの知識をすべて持ち合わせた利用者は少ないため、論理式を試行錯誤的に作成しながら所望の特許を検索することとなり、検索作業に多大な時間と労力がかかっているのが現状である。

適切な論理式の作成にかかる時間と労力を極力軽減する検索手法として、キーワード論理式の代わりに、発明内容を記述した文章をそのまま検索条件とする「自然言語文検索方式」が普及してきている。本方式は、文章中に出現する重み付きターム集合で発明内容を表現し、ターム集合間の類似度を定量化することによって、入力文章と各特許との間の類似性を判定するというものである。

このような技術動向を受け、国立情報学研究所（NII）が主催する情報検索システム評価用テストコレクション構築プロジェクトNTCIR [1] では、この自然言語文検索方式によって発明を無効化する過去の特許を検索する技術の精度向上を目指す「特許検索タスク」が行われている。筆者もこれまでに3回参加しており [3][4][5]、NTCIR-4特許検索タスク（2003-2004）では、Japioと共同で参加して好成績を収めている [3]。

本稿では、直近のNTCIR-6特許検索タスク（2006-2007）における筆者の取り組みとして、特許明細書文章と特許分類を入力として類似特許を検索する方式およびその精度評価について述べる [5]。すなわち、発明の内容を特徴付けるタームの抽出・重み付けにおいて、特許の文書構成の特性を利用する方式と、検索時に特許分類を用いて類似度スコアをチューニングする方式を提案し、NTCIR特許検索タスクで用意されたテストセットを用いてその有効性を評価する。

以下、2章ではNTCIR特許検索タスクの概要と筆者の取り組み方針について述べる。3章では提案する類似特許検索方式について詳しく述べる。4章では本方式の精度を評価した結果について述べる。

2 特許検索タスクへの取り組み方針

2.1 NTCIR特許検索タスクの概要

NTCIR特許検索タスクについて簡単に述べる。本タスクの詳細については、オーガナイザによる報告 [2] を参照されたい。

本タスクでは、特許明細書（公開特許公報）の代表請求項（主に請求項1）文章を入力として、そこに記載された発明内容を無効化する過去の特許（1993-2002年公開特許公報）を検索し、類似度の高い上位1000件を提出することを課題（必須ラン）としている。提出さ

れた結果から検索精度を算出し、参加チーム間で比較する。

また、本タスクでは必須ランとは別に、代表請求項文章以外の文章を入力として使用することが、オプションランとして認められている。2．2節で述べるように、本稿で述べる方式は特許明細書全文および明細書に付与された特許分類を入力としているため、このオプションランに相当する。

2.2　タスク取り組み方針

代表請求項文章は、発明内容を端的に記述しているため、本タスクの必須ランにおいて、請求項文章を類似特許検索の入力とすることは妥当である。

しかし、類似特許検索システムの利用者が実際に置かれる状況を踏まえると、代表請求項文章以外の情報も入力として利用できるケースが多い。例えば、出願直前に出願人が公知例調査を行う場合、請求項文章を含む特許明細書全文を入力情報として利用できる。また、特許庁において審査官が特許性を判定する場合、特許明細書全文に加え、これに付与されている特許分類や出願人などの書誌情報も利用することができる。

そこで筆者のチームは、入力として利用可能な様々な情報を最大限に活用することによって、現状の検索技術によって類似特許検索精度をどこまで向上できるかを見極めることを目的として、本タスクに取り組んだ。特に、特許明細書全文と特許分類を入力とした検索方式について検討することとした。

3　類似特許検索方式の提案

本稿では、「特許の文書構成を踏まえた検索方式（3．1節）」および「特許分類を用いた検索方式（3．2節）」を提案する。

一般に、文書中のタームを用いた類似文書検索において、その精度向上を検討する際に重要となる観点として、以下の三つが挙げられる。

（1）どのタームを検索に用いるか（ターム抽出）
（2）どのタームを重要視するか（ターム重み付け）
（3）文書間の類似度をどのように算出するか（類似度計算）

本稿で提案する「特許の文書構成を踏まえた検索方式」は、上記（1）（2）の観点に立ったものであり、「特許分類を用いた検索方式」は、上記（3）の観点に立ったものである。

3.1　特許の文書構成を踏まえた検索方式

（1）ターム抽出方式

特許明細書を入力文章として使用する場合、まず課題となるのが、検索に使用するタームを明細書のどの部分からどのような基準で抽出するか、である。

一般に、特許明細書の中では、要約や請求項が発明内容を最も端的に記述していると言われている。しかし、要約は発明者／出願人によってその記述精度には差があると言われている。また、請求項は発明内容を敢えてぼかして抽象的に記述することがある。これらの傾向を踏まえると、特許明細書のどの記載箇所に着目すべきかを固定することは得策ではない。とは言え、要約や請求項の中に有用なタームが多く含まれているというのも事実である。

そこで筆者は、特許明細書を構成する以下の三つの構成要素に着目して発明内容を特徴付けるタームを抽出し、検索タームとして使用することとした。

（a）請求項1の中のターム（請求項1ターム）
（b）要約の中のターム（要約ターム）
（c）明細書全文の中で重みの高い上位N個のターム（全文ターム）

ここでタームは、名詞、動詞、形容詞、未知語のみとしている。また、上記（c）で、全文タームの数を限定しているのは、明細書全文で使われているタームの異なり数が非常に多いので、ノイズタームによる精度低下を極力抑えるためである。

（2）ターム重み付け方式1：全文の頻度を利用

特許明細書を入力文章として使用する場合に、次に課

題となるのが、検索に使用するタームの重要度（重み）をどのように割り当てるか、である。

本方式では、重み付きタームからなるベクトルによって文書の特徴を表現するベクトル空間モデル[6]と、タームの出現頻度に基づいてタームの重み付けを行うTF-IDF法[6]を採用している。TF-IDF法では、タームが一つの文章中に出現する回数（Term Frequency, TF）と、そのタームが特許データベース中の文書に出現する回数（出現文書数）の逆数（Inverted Document Frequency, IDF）の積によって、タームの重みを算出する。

TF-IDF法に基づいてタームの重みを算出する場合、文章がある程度長くないと、TFの大きさとタームの重要度との間に相関がなくなる。しかし、請求項や要約文章を入力とする場合、文章が短くなることがしばしばある。

そこで本方式では、特許明細書のどの部分からタームを抽出するかにかかわらず、タームの重みを算出する際に使用するTFの値は、特許明細書全文を対象として算出することにより、重みの信頼性を高めている。

（3）ターム重み付け方式2：ターム共起の利用

3.1節（1）で述べたように、本方式では請求項1ターム、要約ターム、全文タームの3種類のターム集合を抽出する。ターム別の出現傾向を見ると、これら三つのどの集合にも含まれるタームもあれば、一つの集合にしか含まれないタームもある。

そこで本方式では、あるタームがどの記載箇所に出現しているかという観点に着目して、タームの重みを補正する。すなわち、3種類のターム集合のより多くに共起するタームを重要視する。例えば図1に示すように、ターム「検索」は、請求項1ターム、要約ターム、全文タームのどれにも含まれるタームであるので重みの値を据え置いているが、ターム「実行」は、請求項1タームにしか含まれてないので、その重みを強制的に低減させる。出現する箇所が少ない（図1で○の数が少ない）タームほど、重みを低く補正している。

ターム	ターム出現箇所			重み補正
	請求項1ターム	要約ターム	全文ターム	
検索	○	○	○	25 → 25
文書	○	○	×	38 → 30
ターム	○	×	○	22 → 18
入力	×	○	○	48 → 38
実行	○	×	×	32 → 19

図1　ターム出現共起による重み補正

3.2　特許分類を用いた検索方式

出願された特許には、特許分類（国際特許分類IPC、FI、テーマ、Fターム）が付与される。特許分類は技術体系であるため、互いに類似する特許には共通のあるいは近接する特許分類が付与される。本方式では、この特許分類の共通性に着目する。

分類を用いて検索範囲を絞り込む機能は珍しくなく、市販されている文書検索システムの多くで提供されている。すなわち、利用者によって指定された分類または入力文章に付与された分類と同じ分類が付与された特許のみを検索結果として抽出し、それ以外の特許は検索結果から除外する機能である。

しかし、入力文章の発明内容を無効化する特許が、入力文章と同じ技術分野であるとは限らない。筆者の調査によれば、過去に審査官によって引用された延べ特許件数の13%は、対応する出願特許と共通するテーマ分類が一つも付与されていない、技術分野の異なる特許である。単に特許分類が共通するか否かによって検索結果をフィルタリングしてしまうと、この13%の類似特許は検索結果から必ず漏れてしまうことになる。

そこで本方式では、特許分類の共通性に基づいて検索結果をフィルタリングするのではなく、類似度スコアをチューニングするというアプローチを採用する。すなわち、入力文章と共通の分類を持つ特許の類似度スコアを強制的に高くする。本方式では、特許分類のうち、テーマ分類（2600分類、1特許あたり平均約2個付与されている）を使用した。

4 提案方式の精度評価

4.1 実験環境

NTCIR特許検索タスクの実験環境で本方式の精度を評価した。検索エンジンとして、GETA*を用いている。

(1) 評価用特許データ

以下の2種類のデータセットを用いた。

・NTCIR-5データ

課題1189件、正解延べ件数2065件。正解は、審査官が拒絶する際に引用した特許（審査官引例）である。

・NTCIR-6データ

課題1685件、正解延べ件数9871件。正解は審査官引例である。ただし、1課題あたりの審査官引例件数が多い特許を意図的に集めている。

(2) 検索対象

1993年から2002年の公開特許公報（約350万件）を検索対象とした。ただし、入力特許の出願日以前に公開された特許に検索範囲を絞っている。

(3) 評価尺度

情報検索で使われる代表的な評価尺度である「平均精度[6]」と、再現率を表す評価尺度である「検索結果上位N位までに含まれる正解特許件数の割合」の2種類を採用する（Nを20および300とした）。

(4) 実験パターン

表1に示す7種類の実験パターンによる検索精度を比較した。実験Xはベースラインであり、入力文章として請求項1のみを使用し、請求項1内での出現頻度をもとにタームの重みを算出する。実験Aから実験Dでは、3章で提案した検索方式を一つずつ追加適用する。実験EおよびFでは、特許分類の活用方法の有効性を比較する。

表1　評価実験パターン

#	検索方式	実験パターン						
		X	A	B	C	D	E	F
1	請求項1タームを使用	○	○	○	○	○	○	○
2	請求項1での出現頻度を使用	○	−	−	−	−	−	−
	全文での出現頻度を使用	−	○	○	○	○	○	○
3	要約タームを追加	−	−	○	○	○	○	○
4	全文ターム（30語）を追加	−	−	−	○	○	○	○
5	ターム出現共起を利用した重み補正	−	−	−	−	○	○	○
6	テーマ分類によるフィルタリング	−	−	−	−	−	○	−
	テーマ分類によるスコアチューニング	−	−	−	−	−	−	○

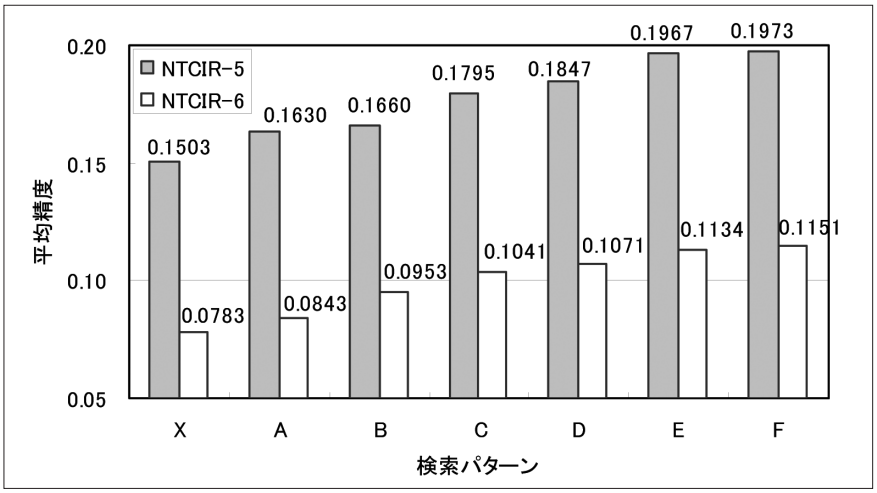

図2　提案方式の有効性評価（平均精度）

* GETA（汎用連想計算エンジン）は，独立行政法人情報処理推進機構（IPA）が実施した「独創的情報技術育成事業」の研究成果である．

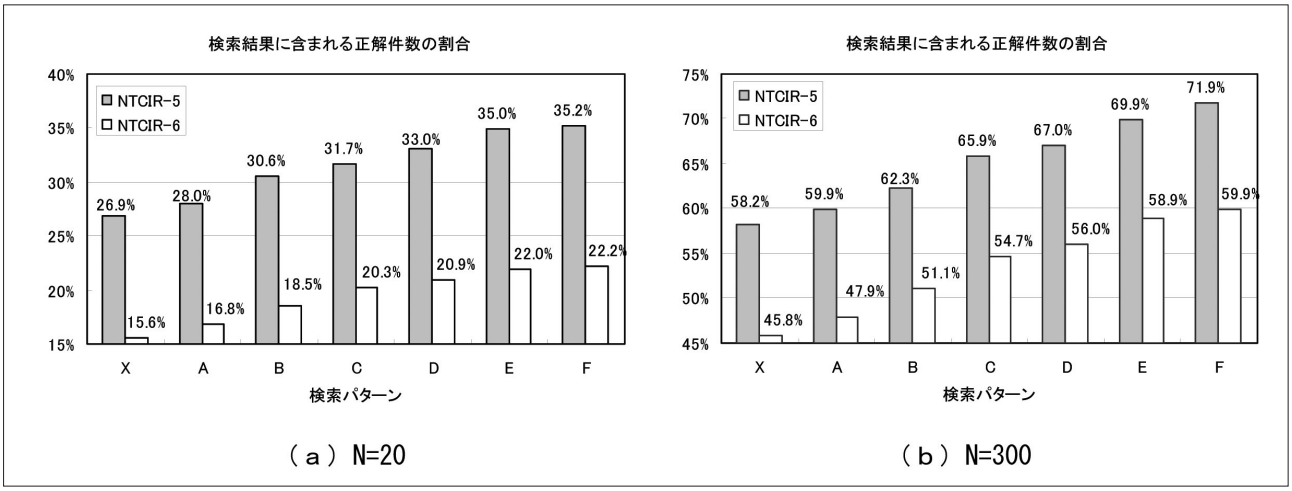

図3　提案方式の有効性評価（検索結果上位N位までに含まれる正解特許件数の割合）

4.2　実験結果

実験パターン別の平均精度を図2に示す。ベースライン（X）に比べ、提案方式を加えていく（A～F）につれて、平均精度の値が向上していくのがわかる。また、実験パターン別の「検索結果上位20位または300位までに含まれる正解特許件数の割合」を図3に示す。こちらも提案方式を加えていくにつれて値が向上しているのがわかる。

さらに、特許分類を用いた実験EとFは、特許分類を用いない実験Dに比べて平均精度も正解特許件数の割合も改善されている。また、正解特許件数の割合（N=300）について見ると、類似度スコアをチューニングする方式（実験F）の方が、フィルタリングする方式（実験E）よりも、NTCIR-5データで2ポイント（69.9%→71.9%）向上している。

4.3　結果の考察

実験の結果、請求項1文章のみを入力とするよりも、特許明細書全文と特許分類を入力とする方が、平均精度も再現率も大幅に改善されることがわかった。これは、入力として使用できる情報量が増えたことで、より多くの検索タームを抽出でき、その重要度を示す重みも精度良く付与できたことに起因していると考えられる。

図2に示したように、本方式により平均精度はNTCIR-5データでは31％（0.1503→0.1973）、NTCIR-6データでは32%（0.0783→0.1151）改善された。しかし、絶対数値で見ると0.2にも達していない。特にNTCIR-6データのように、1件あたりの審査官引例の数が多い案件の場合、発明内容の部分的な一致を認識する必要があるため、NTCIR-5データに比べて検索精度が低くなっている。

一方、図3に示したように、検索結果上位20位までに含まれる正解特許件数の割合は、NTCIR-5データで8.3ポイント（26.9%→35.2%）、NTCIR-6データで6.6ポイント（15.6%→22.2%）改善した。すなわち、正解の3～5件に1件は、本方式による検索結果の上位20件から見つかることを意味している。同様に、上位300位ではそれぞれ、13.7ポイント、14.1ポイントの向上が見られ、NTCIR-5データでは4件の正解特許のうち3件弱が見つかる。この検索精度の現状と、検索条件を指定しなくても特許明細書を指定するだけで類似特許を検索できるというメリットを踏まえると、本方式は技術者が出願前に公知例を検索する状況で特に有効と考える。

5 おわりに

特許明細書文章および特許分類を最大限に活用した類似特許検索方式について提案し、その有効性を検証した。入力の情報量を増やすことにより、請求項1文章のみを入力とする場合に比べて、検索精度を大幅に向上させることができた。

今後は、技術分野の特性を踏まえた検索方式について検討する必要がある。例えば、化学分野では、請求項1タームや要約タームに加え、全文タームが検索精度向上に有効であるが、物理分野では、全文タームが検索精度向上に比較的貢献していないという分析結果も出ている。その技術分野の特許の書き方や読み方を考慮して、ターム抽出・重み付けを最適化することにより、検索精度のさらなる向上が期待できる。また、今回は全自動による検索を想定したが、実際には利用者とのインタラクションによって、より効率良く検索を行えると考えている。どの時点でどのような情報を提示すると効果的な検索ができるのかについて検討していく予定である。

参考文献

[1] Kando, N.：Overview of the Sixth NTCIR Workshop, Proceedings of the Sixth NTCIR Workshop Meeting, pp.i-ix（2007）.

[2] Fujii, A., Iwayama, M. and Kando, N.：Overview of Patent Retrieval Task at NTCIR-6, Proceedings of the Sixth NTCIR Workshop Meeting, pp.359-365（2007）.

[3] 間瀬久雄，岩山真，松林忠孝，小川祐一，大塩只明：文章構造を利用した二段階特許検索方式の提案と評価，Japio創立20周年記念誌（2005）.

[4] Mase, H., Matsubayashi, T., Ogawa, Y., Yayoi, T., Sato, Y. and Iwayama, M.：NTCIR5 Patent Retrieval Experiments at Hitachi, Proceedings of the Fifth NTCIR Workshop Meeting, pp.318-323（2005）.

[5] Mase, H. and Iwayama, M.：NTCIR-6 Patent Retrieval Experiments at Hitachi, Proceedings of the Sixth NTCIR Workshop Meeting, pp.403-406（2007）.

[6] 徳永：情報検索と言語処理，東京大学出版会（1999）.