検索精度向上への取り組み

類似文献検索の特許検索への適用に係わる検討３

一般財団法人工業所有権協力センター　研究所調査員　**藤間　孝雄**

PROFILE

平成 11 年４月より特許検索、分類付与関係に従事、平成 20 年４月より現職

一般財団法人工業所有権協力センター　研究所研究部長　**垣本　和則**

PROFILE

平成15年４月より現職

1 はじめに

一般財団法人工業所有権協力センター（ IPCC：Industrial Property Cooperation Center、以下「財団」と表す。）研究所は、財団の主たる事業である先行技術文献調査事業、特許分類付与事業のさらなる効率化を目指し、 IPCCシソーラス等の独自データ資産を整備するとともに、それらの効果的な活用方法に取り組んでいる。その一環として３年前から、類似文献検索と組みあわせて利用するプロトタイプを構築し、評価を行っている [1]。また２年前からは、 IPCCシソーラスの代表語を特許専門用語として利用し、形態素解析の改良に利用する試みなども評価してきた [2]。

そして、昨年度は、類似文献検索技術を財団の検索業務における利用形態に、より近づけた応用方法を検討した。

2 過去の取り組みで分かった課題

類似文献検索技術を検索業務に利用する場合は、ＦＩ、Ｆターム、検索語による論理式検索と比較して、検索式を考えずにすみ、また検索語（及びその類義語）の漏れに強いという長所があるが、本願との類似性はシャープには特定できず、Ｘ文献、Ｙ文献の判定能力を持たないという短所がある。さらに利用上の大きな課題として以下の点がある。

・課題１

図１の「分野別正解文献 100 位再現率」（詳細は [1] 参照）に見られるように正解文献の100位までの再現率が適用するテーマや本願の特性で大きく異なり、結果が安定しないことである。正解文献のランキングが10位以内に入る場合もあれば、 1000位以内にも入らない場合もある。この特性は平均順位では表れない応用上の障害となり、安定した再現率を得ることが必要となる。

・課題２

従来のFI、Ｆターム、検索語による論理式検索手法等と融合した検索手法として確立することが必要である。

3 今回の取り組み

過去の取り組みで浮かび上がった課題に対し、実用性の立場から、検索実務に用いている手法を取り入れ「領域を限定した類似文献検索」、「類義語を同一視した類似文献検索」、「論理式検索と組み合わせた類似文献検索」を新たな試みとして実施し、解決策の可能性を検証・評価した。

３．１　領域を限定した類似文献検索

文献全体が必ずしも類似ではないが、特定の段落が

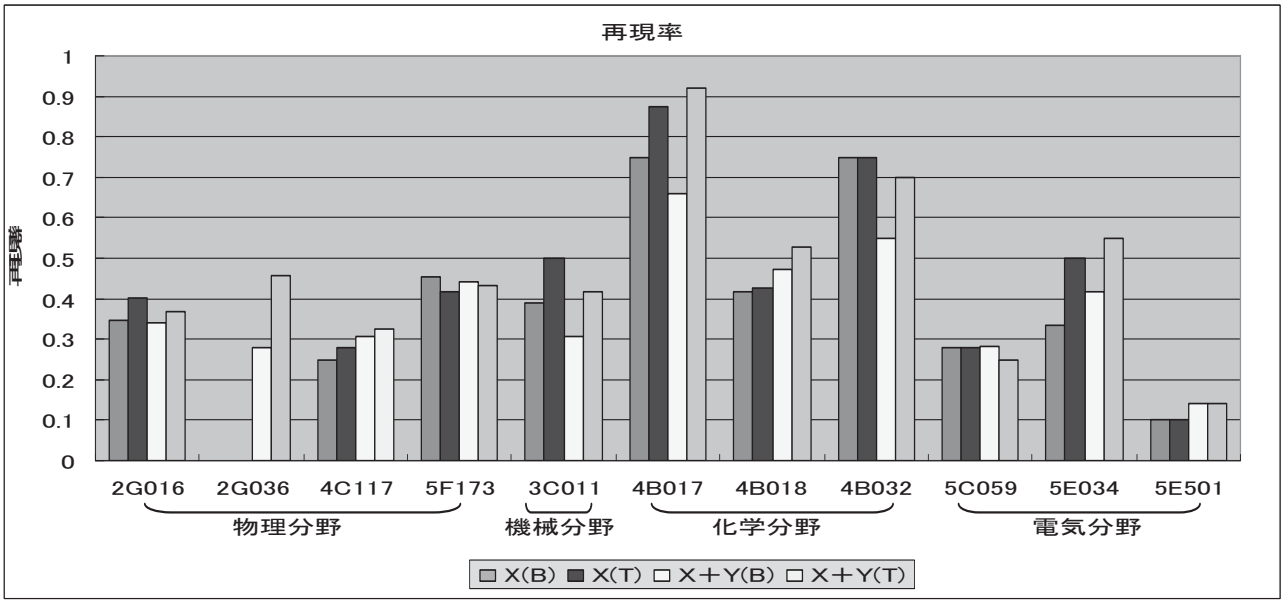

図１　分野別正解文献 100 位再現率

類似であることを判定できるように類似文献検索の機能を拡張する。以下、本稿では文献の領域として「発明を構成する要件」に類似する段落を想定する。

発明を構成する要件

ここで言う発明の構成要件は広く捉えて、発明の構成要素、発明の目的等とする。具体的に構成要件は、

① 第１クレームを複数の構成要素に分け、それぞれの分説した発明の構成要素を当業者なら分かる程度に平明な短文で記載したもの

② 発明のポイントを記載した平明な短文と発明の目的等である。

複数要件を具備する文献の類似文献検索

ここで複数要件を具備する類似文献検索として次のような類似文献検索を行う。

① 前述の各構成要件毎に作成した短文を類似文献検索の入力クエリーとして、それぞれ独立に対象文献集合に対して類似文献検索を行い、対象文献集合の総ての文献の総ての段落のテキストと構成要件毎のクエリー文との類似度スコアを求め、文献を構成する段落の最高スコアをその文献の各構成要件に対応した類似度スコアとする。

② 前記類似度スコアを一定の閾値で足切りをし、閾値以下の文献はその構成要件は具備していないと判断する。閾値以上の文献はその構成要件を具備すると考え、以下構成要件候補集合と呼ぶことにする。総ての構成要件毎に、それぞれの構成要件毎の足切り閾値を計算して足切りを行い、構成要件候補集合を求める。

③ 正解文献（ここではＸ文献を想定する。）は総ての構成要件を具備すると考え、図２に示すように総ての構成要件候補集合の積集合を求める。この積集合をここでは正解文献候補集合と呼ぶ。

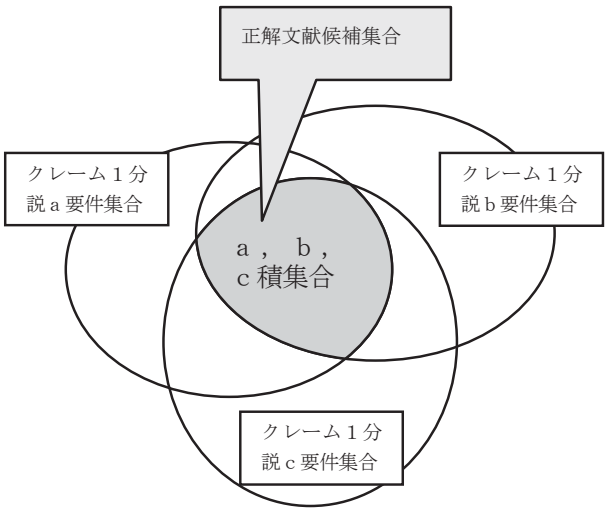

図２　正解文献候補集合の概念

閾値の決め方

ここで閾値を決めることは重要な意味を持ち、閾値が低過ぎれば対象文献集合から正解文献候補集合を求めても絞込みが効かず、構成要件毎の類似文献検索を行なった意味がなくなる。

また閾値が高過ぎると正解文献が正解文献候補集合から漏れてしまう。ここで類似度スコア（Okapi　BM25アルゴリズムで類似度の計算をする。）はクエリー文から抽出した特徴語とその類義語に対して求めた個別特徴語毎のスコアの総和であるから、ここでは、

閾値＝K＊（総特徴語数）　・・・　式1

K：係数

とする。

適正な閾値が構成要素毎に総特徴語数関数として決めることができれば、正解文献を漏らすことなく対象文献集合から正解文献候補集合を絞り込むことができることになる。

正解文献候補集合の縮小率と正解文献再現率

正解文献を漏らすことなく対象文献集合から正解文献候補集合を絞り込むことができるか否かを検証するために、正解文献候補集合の縮小率と正解文献再現率の関係についてテスト用本願と正解引例を選択し、前述の複数要件を具備する文献の類似文献検索方法による実験を行った。ここで、

正解文献候補集合の縮小率＝｛（対象文献集合―正解文献候補集合）／対象文献集合｝＊100

正解文献再現率＝｛（正解文献候補集合に含まれる正解文献数）／正解文献数｝＊100

である。

実験方法

・テーマ特性を考慮してテーマ2G016（遮断器と発電機・電動機と電池等の試験）、4B017（非アルコール性飲料）、5E501（デジタル計算機のユーザインターフェイス）、5G503（電池等の充放電回路）、5H180（交通制御システム）を選択し各５件のテスト用本願を選択した。正解文献は検索者が報告している拒絶引用文献とした。

・発明の構成要件としては第１クレームを分説し、各構成要素毎に当該分野の専門家が実施例を読み、その具体的内容を簡潔な短文にまとめ、類似文献検索のクエリー文とした。

・比較の目的で第１クレームを特に分説せずに１つの短文に纏め類似検索のクエリー文とする実験も行うこととした。

・式１のＫを変化させることにより構成要件毎の総特徴語数に応じて閾値が変わるので、その結果、正解文献候補集合の縮小率が変わる。その時の正解文献候補集合に含まれる正解文献の数を調べることにより再現率を求め、縮小率と再現率のグラフを求めた。

・類似文献検索は財団のプロトタイプ[1],[2]を一部修正して用いた。具体的には

類似文献検索ツール：情報処理振興事業協会（IPA）の研究成果である汎用検索エンジン（GETA）を用い形態素解析器としてChaSenを、形態素解析用辞書としてIPADICを利用するとともに、類義語拡張として財団のシソーラス辞書を利用

WAMの構成文献：テスト用本願毎にテスト用本願の検索テーマに含まれる全文献に対して、それを構成する全段落を類似比較対象テキストの単位とみなして特徴語抽出をし、テーマ毎に纏めWAMを構成

類似度スコアの算出方法：Okapi　BM25

実験結果

正解文献候補集合の縮小率と再現率について実験結果を図３に示す。ここに示されるように類似文献検索でこのような絞り込みを行うと縮小率が80を超えるあたりから再現率は急速に下がる。

結果の考察

実験結果から、今回のような類似文献検索を行った場合には、テーマ全体の文献に対して縮小率を８０以下に抑えれば再現率を高くできることが分かる。このような方法で正解文献候集合を絞る場合には最初の論理式検索で絞り込んだ集合サイズの１／４、１／５以上に類似文

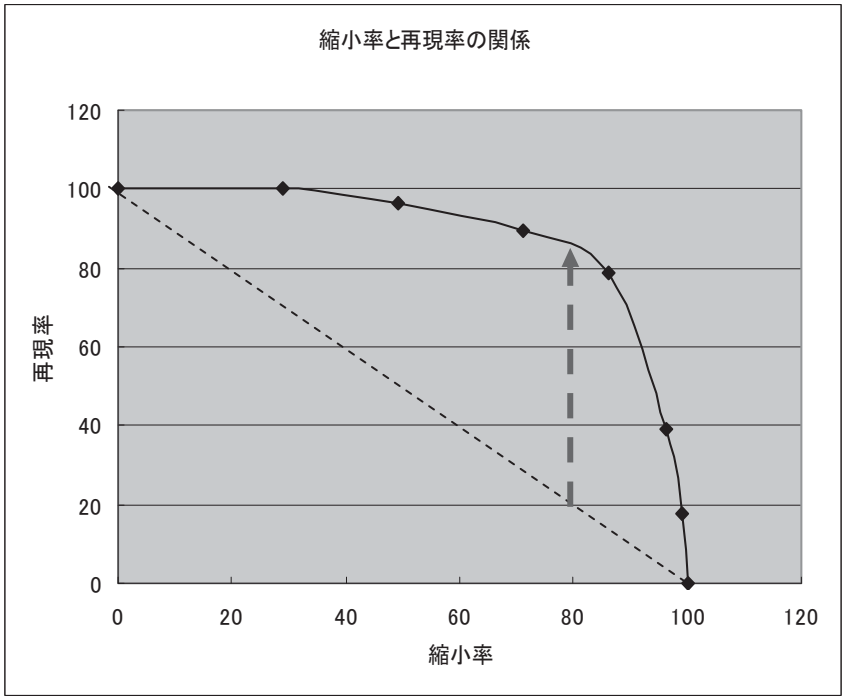

図３　縮小率と再現率の関係

献検索で絞り込むと再現率は急速に悪くなり、それ以内の絞り込みに抑えれば高い再現率が得られると言うことを意味している。

このことは、対象文献集合を類似文献検索以外の方法（FI、Ｆターム、検索語による論理式検索等）で絞り込んでおいても、テーマ内の文献であれば同様な傾向が期待できると考えられ、課題１に対する１つの対策となりうることを示唆している。

また、各構成要素毎に評価する類似文献検索では、ある構成要素のスコアが高く、別の構成要素のスコアが低いという文献は総合スコアが良くても正解文献候補集合には上がってこないフィルタ効果が期待できる。

３．２　論理式検索と組み合わせた類似文献検索

前述の考察どおり、論理式検索で正解文献候補集合を図３の高い再現率が得られる絞り込み領域に絞り込んだ後に類似文献検索を行うことで安定した再現率を得ることができるか検証を行った。

実験方法

①従来の特許検索システムを用いてFI、Ｆターム、検索語による論理式検索を行い、1000件以下の類似文献検索の対象集合をテスト用本願毎に抽出した。この実験では、図１の実験で成績が特に悪いテーマ5E501（デジタル計算機のユーザインターフェイス）に限定して実験した。

②３．１で前述した「発明を構成する重要な要件毎の類似文献検索」の要領に従って、発明の構成要素毎にクエリー文を作成し領域を限定した類似文献検索を行い、検索対象集合を構成する全文献に対して類似度スコアを求め、閾値との比較により足切りをし、構成要素毎に要件を満たす文献集合を求めた。この場合、類似文献検索の対象集合は①の論理式で求めた集合である。

③構成要件毎に要件を満たす文献集合の積集合を求め正解文献候補集合とした。ここで再度、構成要素のクエリー文を合体させたクエリー文を作成し、正解文献候補集合の総ての文献に対して類似度スコアを計算し、ランキング表示をした。

表１　論理式検索と組み合わせた場合の類似文献検索

テスト本願	1次絞り込み	正解引例 (種別)	1次絞り込み後の構成要素毎の正解文献順位（A）	全テーマ対象検索正解文献順位（B）
テスト本願A	テーマ5E501 論理式絞り込み 871 件	2002-014752 (X)	19	1000位外
		2001-013945(Y)	60	1000位外
テスト本願B	テーマ5E501 論理式絞り込み 730 件	2003-00588(X)	22	1
テスト本願C	テーマ5E501 論理式絞り込み 754 件	2002-047813(X)	5	43
テスト本願D	テーマ5E501 論理式絞り込み 992 件	2002-90529(X)	55	560
		2002-1968855(Y)	4	86
テスト本願E	テーマ5E501 論理式絞り込み 993 件	2001-216069(X)	7	1000位外
テスト本願F	テーマ5E501 論理式絞り込み 765 件	2003-022154(X)	65	1000位外
		2000-305746(Y)	152	1000位外

実験結果

表１にテーマ5E501の６件のテスト本願に対する正解引例、論理式検索で１次絞り込みした文献集合の文献数、論理式による１次絞り込みを実施した後の構成要素毎の類似文献検索方法による正解文献の順位、全テーマを対象とした無テーマ類似文献検索をした場合の正解順位を示す。

結果の考察

全テーマを対象にした類似文献検索結果（B）欄とテーマ内を論理式で絞り込んだ後に発明の構成要素を考慮した「発明を構成する重要な要件毎の類似文献検索」の順位結果（A）欄と比較すると、（B）欄の正解文献の順位は１位から1000位以下と大きくばらついている。一方、（A）欄の正解文献順位は概ね100位以内に押さえ込まれている。順位が高くなることは論理式検索によって類似検索の対象集合が1000件以内に絞ってあるから当然であるが、注目すべきことは正解文献候補集合を1000件以内に論理式検索で絞っておけば正解文献が高い再現率で100位以内に入ってくることである。

実務的な立場からするとテーマ内のFIやＦタームに精通した人であれば1000件以内に絞ることは比較的容易であり、その中から正解文献を効率的に探し出すことが求められる。今回の実験では図３グラフ中の上矢印線で示すように縮小率に比例して再現率が下がるのではなく、縮小率を適当に抑えておけば再現率の低下が比較的に少ないという効果が出ている。従って、論理式検索で一旦絞り込んだ後に、今回の様な類似文献検索技術で絞り込む方式は検索業務効率の向上につながると考えられる。

4 終わりに

今回の実験は、検証した件数が少ないため、検索業務に類似文献検索が上手く利用できると結論づけることはできないが、一定の方向性を確認することができた。また類似文献検索の所定順位内に正解文献の再現率を高くできても、集合を絞れば漏れも生じてくるので、類似度絞り込みをした場合には、実際にスクリーニングした文献を明確にし、従来のFI、Fターム、検索語による論理式検索と融合して検索漏れ防止を保証できることが求められる。

今後は、これらの知見をもとに類似文献検索技術の先行技術文献調査業務への適用領域を探る実用化研究を進めて行きたいと考える。

参考文献

[1] 居島一仁，検索精度向上への取り組み（類似文献検索の特許検索への適用に係わる検討），Japio 2009 YEAR BOOK，pp 168-171，2009

[2] 土居仁士，検索精度向上への取り組み（類似文献検索の特許検索への適用に係わる検討２），Japio 2010 YEAR BOOK，pp 180-181，2010