特許情報から抽出した複合名詞を利用した文書類似度の検証

Verification of document similarity making use of compound noun that is extracted from patent information

柳堀恭子 [1*]　津田和彦 [1]

Kyoko Yanagihori[1], Kazuhiko Tsuda[1,]

[1] 筑波大学大学院

[1]Graduate School of Business Sciences, University of Tsukuba

Abstract: when patent examiner indicates a refusal to patent application, "Notice of Reason for Refusal" is issued. We have created a synonym dictionary of compound nouns using the information extracted from the Notice of Reason for Refusal, and extracted the dependency relationship of compound nouns and verbs in the claims full text written in patent applications. Between this patent application and the patent publication cited, we measured a similarity by using created formula. Then, when using the synonym dictionary, we investigated whether this synonym dictionary is effective.

１　はじめに

特許における先行研究調査は、出願前に自社の発明が他社のもつ特許に抵触していないかどうかを調査する目的で行われる場合が多い。特許侵害においては、製品の差止請求ならびに損害賠償請求をされる可能性も含み、損害請求額は億にのぼる場合もあり、訴えられた側の損失は非常に大きい。[1] 特に開発や製品化が進んでしまったあと、他社から「貴社の製品は弊社の特許を侵害している」との通知を受けてしまうと、取り返しのつかない損失になりかねない。このように、開発の前段階などに先行研究調査を行うことは重要なことであり、数多ある文献について完全なる調査を行うことは難しいが、なるべく漏れのないように調査を進めるべきである。

このような背景から、本研究で先行研究調査を効率よくすすめるための提案を行う。

調査漏れの原因の１つに特許文書から意図する発明をくみ取ることが難しいということがあげられる。特許文書、とりわけ発明の範囲を決定する請求項の部分では、独特な記載方法がとられている。句点が使われず、長い文章であっても１文で書かれていること。不明確性をなくす意味で、指示語はなるべく使われていないこと。発明の新規性を強調するために、出願者によって辞書にないような語として、名詞をつなげて新たな複合名詞を作りだして発明を説明している場合があること。これらが請求項を難読化させてしまっている要因となっている。

本研究では、漏れのない調査を行うためにまず、文書検索に複合名詞を利用することを考えた。難読な請求項を解釈しやすくするために、自然言語処理におけるテキストマイニングの手法を利用して請求項中に出現する複合名詞を係り元として係り受け解析を行い、整理をした。そして、類似文書検索に応用するために、類似する複合名詞を特許文書中から抽出し、複合名詞の辞書の作成とその辞書を利用し、係り受け解析した請求項文同士を比較し、文書間類似度を検証した。

２　検証

複合名詞の類似辞書を作成し、抽出データに作成した辞書を適用した場合と、しなかった場合での類似度を測り、辞書の有効性を検討する。

*連絡先：筑波大学大学院ビジネス科学研究科
〒112-0012　東京都文京区大塚 3-29-1
E-mail kyoko@gssm.otsuka.tsukuba.ac.jp

下記図１の流れで検証を行った。

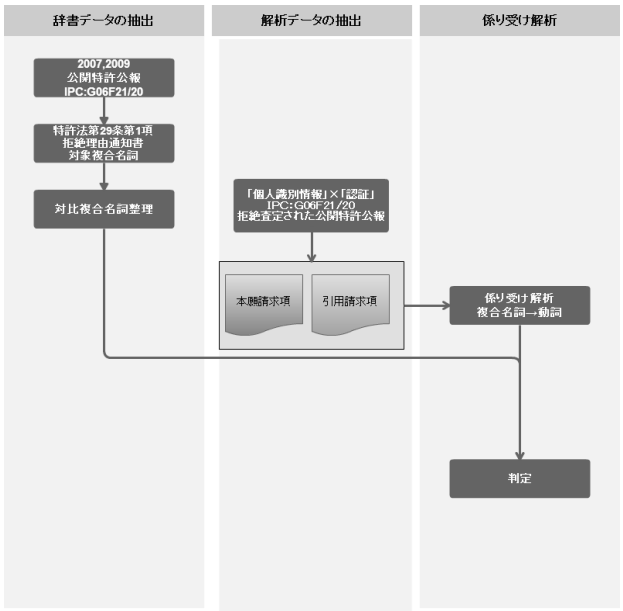

図１：解析の手順

２．１　辞書データの作成

辞書データの抽出は、特許電子図書館 IPDL（http://www.ipdl.inpit.go.jp/homepg.ipdl）のデータベース[2]を利用した。

抽出対象は、国際分類番号（IPC）G06F21/20 とした。物理学器械セッションの電子的デジタルデータ処理不正行為から計算機を保護するためのセキュリティ装置の中のコンピュータシステム コンピュータネットワークのノードへのアクセスの制限によるものを対象とした分野である。この分野で 2007 年 1 月 1 日から 2007 年 12 月 31 日まで、および 2009 年 1 月 1 日から 2009 年 12 月 31 日までに出願公開された公開特許公報のうち、特許審査官より拒絶査定を受け、その際に示される拒絶理由通知書に類似の対比箇所の記載がある部分を抽出することで作成したものである。

２．２　解析データの抽出

次に、解析対象データを抽出した。
キーワード「個人識別情報」∩「認証」を請求項にもつ公開特許公報を 246 本抽出した。このうち、審査請求された特許出願（以下、本願とする）に対し、特許審査官が、拒絶査定をし、公報その理由として拒絶理由通知書で類似する引用文献（以下、引用とする）を示し、辞書抽出と同じ分野となる国際分類番号G06F21/20 を本願にもつ18組の本願と引用のペアを比較した。

２．３　類似判定の方法

類似度の判定には式１の「係り受け類似度」を使用した。従来の文書類似比較の手法では、複合名詞を最小単位の名詞に分割する形態素解析を利用するため、本来の複合名詞がもつ意味ではなく細切れの単語の出現により類似度が変わってしまう。[3]
例えば、「個人情報認証システム」という複合名詞が「個人」「情報」「認証」「システム」に分割されてしまうため、特許文書で多く使われる「システム」が書かれている文書だと、どのようなシステムであれ文書が類似してしまうことになる。そこで、新たに複合名詞による類似度を測る式を定めた。

$$F(sim)= \frac{1}{N}\sum_{i=1}^{N} k_i\alpha\beta \qquad \text{（式１）}$$

N は、本願のリンク数（係り受け抽出数）
M は、引用のリンク数（係り受け抽出数）
式２で表される Ki は、X を本願の係り先数、Y を引用の係り先数としたときの係数とする。
係り元が類似であっても、係り先が同じ出ない場合は係数 Ki で重み付けする。係り先が同じ場合は Ki を１とする。

$$K_i= 1 - \frac{|X \cap Y|}{|X \cup Y|} \qquad \text{（式２）}$$

図２の α と β は、各ノードのリンクへの寄与率を示す。

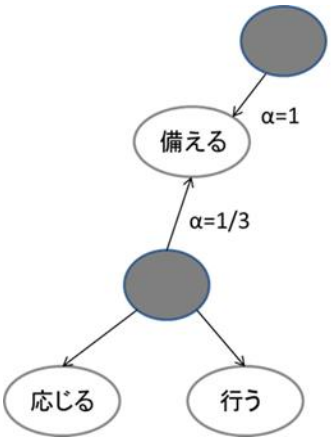

図２：リンク寄与率

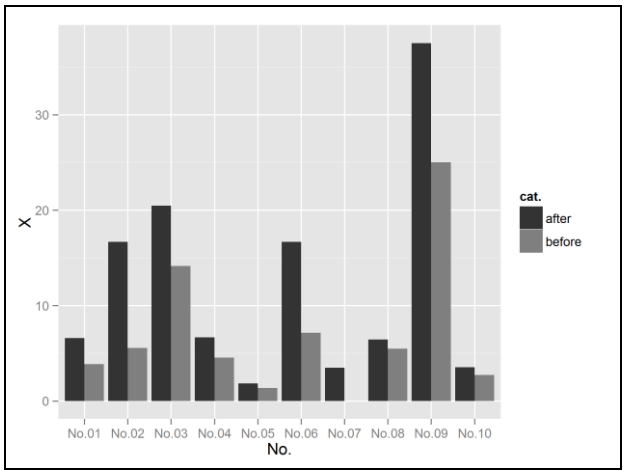

図３　辞書適用前後の結果

また、本願に対する類似度をみるが、係り受け解析によって、引用の係り受け関係の抽出が本願からの抽出数に比べて著しく少ない場合は、N の数を $N=\mu H$　と、式３に示すように、N と M の調和平均に置き換えて計算することで、調整をした。

$$\mu H = \frac{2NM}{N+M} \quad \text{（式３）}$$

２．４　係り受け解析

係り受け解析について、テキストマイニングスタジオ[4]を利用し、係り元を名詞が２語以上連なる複合名詞とし、係り先を自動詞に設定し、本願と引用の各請求項全文より抽出した。

２．５　判定

対象となった、本願と引用 18 本のうち、抽出数 0 となった本願が 1 本あったため、残りの 17 本で係り受け解析した結果を判定し、表 1 に示す。

表 1：類似判定の結果

類似度が上昇した	10
類似度に変化なし	4
比較不能	3

図３より、類似度が上昇したものについて、１つ１つの類似度が違うので、平均値ではなく、辞書適用前と後での類似度の平均でみると、辞書適用前平均と辞書適用後平均で 5%上昇したことがわかった。辞書前との変化率（上昇率）をみると、0.696 となった。

比較不能となった 3 本について、その原因として、請求項全文はある程度の文字数が揃うはずだが、請求項数が 1 つあるいは、字数が 100 字程度と少ない場合は、係り受けが 1 つあるいは 2 つ程度しか抽出されず、引用との比較ができないということがあげられる。

３　考察

係り受け類似度式をもとに辞書を利用した場合としなかった場合の類似度を比較した。

辞書の範囲である IPC G06F21/20 をもつ本願 18 組と、それに対応する審査官が類似判断した引用をみると、引用は同じ分類番号 G06F21/20 もっていないことから、特許検索で行われるキーワード×分類番号検索でヒットしないことになる。ヒットさせるためには、キーワードの拡張を行わなければならない。例えば、公開特許公報番号 2008-117179 である本願内「携帯電話」と公開特許公報番号 2002-223253 の引用内「端末装置」の係り先が同じ「含む」であることから、係り元が類似である可能性がある。作成した辞書を適用することにより、本願と引用の「携帯電話」と「端末装置」がこれら語の上位概念となる「通信端末」で置き換わることになり、その結果文書の類似度はあがることになる。

つまり、拒絶理由通知書から抽出した複合名詞により作成された辞書を使うことによって、係り受けが同義あるいは類似となることにより、係り受け解

析からの類似度が高くなる。これにより、辞書の有効性が示されることになる。そして、このようにキーワードを拡張していくことで検索への応用ができると考えられる。

４　おわりに

通常の文書検索の手法では形態素解析手法がベースとなっているため、「装置」「機器」などのどの文書にも書かれているような単名詞が存在すると、内容的にまったく違う装置であっても類似文書と見なされ、検索の再現率は高いが、適合率は低くなってしまう。この手法だと、類似候補の文献はたくさんあるが、本当に適合する文書にあたる確率は低いということになり、効率的な検索を行うことはできない。適合率と再現率はトレードオフの関係にあるため、類似文書検索にあたり出現文書の再現率が下がってしまうことは望ましくないが、漏れのないだけではなく、漏れがなく正しい結果を求める検索を行うためには、検索の際のヒントでもある類似複合名詞の辞書の規模を拡大し、多くのキーワードで適合する検索対象文書を狭めていく必要がある。

さらに複合名詞を拒絶理由通知書から抽出した辞書は有効に活用できるものと考え、さらに大きなデータを対象として解析していきたい。

参考文献

[１] 東京地裁平成 20 年（ワ）第 36814 号特許権侵害差止等請求事件

[２] IPC, International Patent Classification. http://www.wipo.int/classifications/ipc/en/

[３] 長尾真: 自然言語処理, 岩波講座ソフトウェア科学 15, 岩波書店, (1996)

[４] テキストマイニングスタジオ http://www.msi.co.jp/tmstudio/