人工知能学会 第21回知識流通ネットワーク研究会
SIG-KSN-021-06

公開特許公報を対象としたシソーラス作成法の研究

林 祐介[1]，山本 修一郎[1]

[1] 名古屋大学大学院　情報科学研究科

A thesaurus development method from patent information

Yusuke Hayashi[1]，Shuichiroh Yamamoto[1]

[1] Graduate school of Informatics, Nagoya University

概要

特許情報をシソーラスに変換することで既存特許の可読性を高める手法を提案する。具体的には特許情報プラットホームで公開されている公開特許公報を対象とし、その中で出現する名詞、動詞の属性、さらには位置関係に着目し提案規則を適用することでシソーラスを作成する方法を提案する。また公開特許公報からシソーラスへの情報反映率、精度、第三者がシソーラスを読んだときの理解度を明らかにする。

Abstract

We propose a method to increase understandability of patent by using thesaurus developed from patent descriptions. Specifically, we propose a method to analyze patent publications disclosed on the patent information platform. The proposed method creates a thesaurus by using rules on nouns, verb attributes and their relationships of positions. We also clarify the information reflection rate from patent publications, the accuracy, and the understandability of the thesaurus developed by the method.

1　はじめに

米国では、特許件数に対する IT 関連特許件数は 80 %を超えている [1]. したがって、知的財産活動における IT 特許の存在は大きくなってきている。したがって、IT 特許を創造するための手法が必要である。

従来から、既存特許に基づいて新たな特許を体系的に創造する手法として TRIZ[2] が知られている。ところが、TRIZ は物理現象を対象としているため、IT 特許に対して、そのままでは適用できないという問題があった。このため、IT 特許に基づいて新たな IT 特許を創造する手法が求められている。

本稿では特許情報プラットホームから取得できる公開特許公報に基づきシソーラスを作成することで早く正確に特許情報を理解する方法を提案する. 公開特許公報の項目の一つである「解決手段」項にある名詞と動詞に着目し、動詞の分類および名詞、動詞間の位置関係から機械的にシソーラスを作成する手法を提案する.

本稿の構成は次の通りである. 第 2 章では我々の研究分野である特許工学について、類似研究および対象となる公開特許公報について説明する. 第 3 章では提案手法とその適用例を述べる. 第 4 章では評価実験の内容とその結果ついて述べる。第 5 章では実験仮説、作成されたシソーラスの有効性と限界について考察する. 第 6 章では、まとめと今後の課題について述べる.

2　関連研究

本研究の研究領域は特許工学と呼ばれる比較的新しい研究領域である. 特許工学とは特許出願戦略の立案から特許権の消滅までの約 20 年にわたる特許ライフサイクルの各種活動に対して, 情報工学的アプローチにより支援する学問である.[3] 本研究は発明構築フェーズの発想支援ツール, 発明支援ツールの作成にフォーカスを当てている.

類似研究として Kobayashi らの研究 [4] がある.Kobayashi らの研究では複数の特許情報から

名詞間のシソーラスを自動的に作成する方法を提案し, パテントマップの作成に活用しようとしている. しかし特許一件一件を表現するシソーラス作成法の研究はなされていない. そこで我々は 1 つの特許情報から 1 つのシソーラスを作成することで特許情報を整理し, 素早く理解するための方法を提案する. 他の類似研究では, フィーチャモデルを用いて特許の構成要素を整理することで新たな IT 特許を作成する研究もなされている [5].

本研究で対象としているのは特許情報プラットホームで公開されている公開特許公報である. この公開特許公報は PDF 形式で公開されているためデータが構造化されておらず, 分析するのには不向きである. また公開特許公報は「課題」,「解決手段」,「特許請求の範囲」,「発明の効果」などさまざまな項目が記載されているが, これらは公開特許公報の筆者によって項目の有無に差異がある. このことも公開特許公報の利用を妨げる原因になっていると考えられる. よって本研究では項目の有無に左右されない手法を提案する.

3 提案手法

本手法の目的は自然言語文書で公開されている特許情報からシソーラスを作成することで特許情報を整理し, 既存特許の理解を容易にすることである.

3.1 提案規則

本手法では公開特許公報の「解決手段」項に現れる名詞と動詞に着目しそれらの前後関係から機械的にシソーラスを作成する. 本手法では動詞を以下の 4 つに分類する.

- 入力動詞
  特許システムの入力要素の述語となる動詞.
  例:「使う」
- 出力動詞
  特許システムの出力要素の述語となる動詞.
  特許システムの処理要素となる.
  例:「査定する」
- 修飾動詞
  文中の名詞を修飾する動詞.
  例:「関連する」
- 未分類動詞
  上記 3 つのどれにも分類できない動詞.

本手法では以下の手順で公開特許公報の解決手段項からシソーラスを作成する.

- 手順 1:解決手段項の文を動詞で区切る.
- 手順 2:文中の動詞は入力動詞、出力動詞、修飾動詞に分類し、どれとも確定できないものは未分類動詞とする.
- 手順 3:入力動詞の文に含まれる名詞を入力名詞, 出力動詞の文に含まれる名詞を出力名詞, 修飾動詞の文に含まれる名詞を修飾名詞として関係づける.
- 手順 4:出力動詞は前に存在する入力動詞と関係づける.
- 手順 5:修飾動詞は後ろにある名詞に関係づける.

以下では各手順に基づきシソーラスの作成例を示す. 例で用いる公開特許公報の解決手段項の内容の一部は「広告トリガに応答して、候補ビデオ広告を識別し、前記候補ビデオ広告のビデオ関連の特徴に関連する 1 組の少なくとも 1 つのキーを識別し…」である.

3.1.1 手順 1

動詞で文を区切る
「広告トリガに応答して、」「候補ビデオ広告を識別し、」「前記候補ビデオ広告のビデオ関連の特徴に関連する」「1 組の少なくとも 1 つのキーを識別し」…

3.1.2 手順 2

動詞を分類する
入力動詞は「応答する」、出力動詞は「識別する」、修飾動詞は「関連する」、未分類動詞はない.

3.1.3 手順 3

動詞と名詞を関連付ける
「広告トリガ」は入力名詞として「応答する」と関係づける.「候補ビデオ広告」、「一組」、「キー」は出力名詞として「識別する」と関連付ける.「前記候補ビデオ広告のビデオ関連の特徴」は修飾名詞として「関連する」と関連付ける.

3.1.4 手順 4

入力動詞と出力動詞を関連付ける
入力動詞「応答する」と 2 つの出力動詞「識別する」を関連付ける.

3.1.5　手順 5

修飾動詞を名詞と関連付ける
修飾動詞「関連する」は直後に現れる出力名詞「一組」、「キー」に関連付ける.

手順 1〜5 を行うことで作成されたシソーラスは以下の図 1 である。

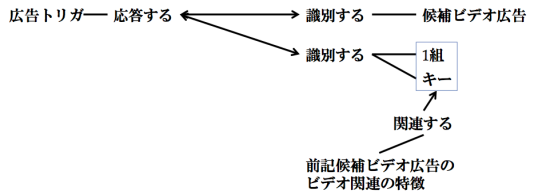

図 1: 作成されたシソーラス例

提案手法のシソーラスで各ノードをつなぐリンクを図 2, リンクとノードの関係を図 3 で示す.

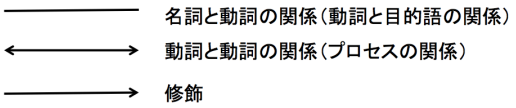

図 2: シソーラスのリンク

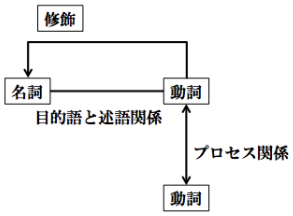

図 3: リンクとノードの関係

3.2　提案規則適用

3.2.1　適用対象

適用対象となる公開特許公報は特許情報プラットホームから「IT」and「広告」で検索し取得した 5 件である.

3.3　適用結果

作成されたシソーラスの一部を以下の図 4,5 で示す.

また、以下の評価項目で評価した結果を表 1 に示す

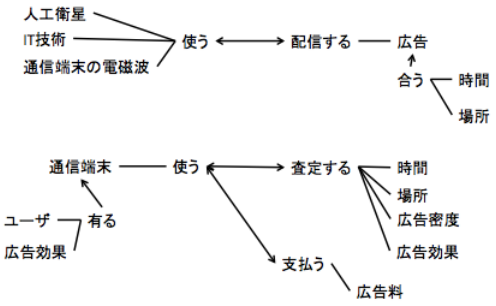

図 4: 作成されたシソーラス 1

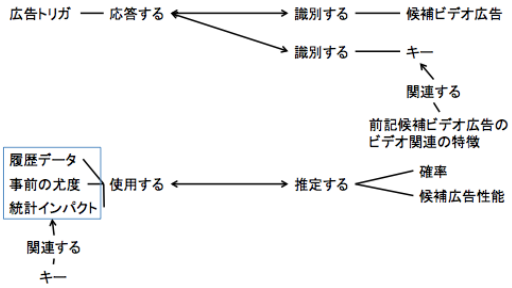

図 5: 作成されたシソーラス 2

1. 修正率
   機械的に作成した後, 目視でどれだけ修正したか. これが低いほど人間が行う作業が少なくて済む.
2. 利用率
   文書内に存在する名詞, 動詞がどれだけシソーラスに利用されているか. これが高いほど文書の情報が反映されていると言える.

修正率、利用率はそれぞれ以下の式で計算される.
修正率　＝　修正箇所/利用された名詞、動詞数＋生成された関連付け数
利用率　＝　利用された名詞、動詞数/文書中の名詞、動詞の総数

表 1: 評価結果

対象	修正率	利用率
特許文書 A	5.0%	60.7%
特許文書 B	7.1%	69.2%
特許文書 C	19.0%	67.7%
特許文書 D	12.0%	61.0%
特許文書 E	10.5%	42.3%

4　評価実験

4.1　実験仮説

評価実験を行うにあたって立てた仮説は以下の 4 つである.

- 実験仮説 1:特許の構成要素が公開特許公報に比べてシソーラスのほうが早く抽出できる.
- 実験仮説 2:特許システムの構成要素が公開特許公報に比べてシソーラスのほうが多く抽出できる.
- 実験仮説 3:特許システムの構成要素が公開特許公報に比べてシソータスのほうが正確に抽出できる.
- 実験仮説 4:入力構成要素と出力構成要素の関係が正確に取れる.

4.2　実験概要

本実験を行うにあたって以下の 4 つの資料を作成した. 便宜上 A1,A2,B1,B2 と名付ける.

- A1
  形式:自然言語資料
  対象特許:IT 広告をインターネットをしない人に届ける方法
- A2
  形式:自然言語資料
  対象特許::ビデオ広告クリエイティブに伴う改善された広告
- B1
  形式:シソーラス（図 4）
  対象特許:IT 広告をインターネットをしない人に届ける方法
- B2
  形式:シソーラス（図 5）
  対象特許::ビデオ広告クリエイティブに伴う改善された広告

被験者は公開特許公報を一度も読んだことのない学生 8 人である. 被験者 8 人をグループ α（被験者 1〜4)、グループ β（被験者 5〜8) の二つのグループに分ける. グループ α には A1 → B2、グループ β には B1 → A2 の順番にそれぞれ文書あるいはシソーラスを見ながら表 2 の設問に答えてもらう. また資料を読んでからすべての設問を答えるのにかかった時間を計測する. ただし Q4 だけは対象特許によって設問が異なる.

表 2: 設問内容

ID	設問内容
Q1	本特許システムの入力要素をすべて書き抜け
Q2	本特許システムの処理要素をすべて書き抜け
Q3	本特許システムの出力要素をすべて書き抜け
Q4-1	資料に記載されている「確率」は何を用いて推定されるか
Q4-2	資料に記載されている「広告効果」は何によって査定されるか

4.3　実験結果

ますはじめに設問への解答時間の分布を図 6 に示す.

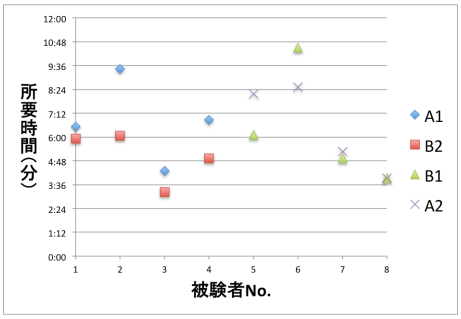

図 6: 比較結果:所要時間

次に平均正答数を表 3 で示す.

同じ特許を対象とした自然言語文書とシソーラスでの解答時間と正答率の分布図を図 7,8 で示す.

図 9 は平均所要時間,10 は平均正答率の比較である.

5　考察

5.1　実験仮説の検証

図 6 から、8 人中 7 人はシソーラスを用いたときのほうが早く設問を解いているのが分かる. さらに図 7,8 から解答時間に対する正答率がシソーラスを用いたときの方が高い位置に分布していることが分かる. このことから実験仮説 1 は成立したと言える.

表 3: 設問と平均正答率

A1Q1	B1Q1	A2Q1	B2Q1
0.25	2	1.5	2.5
A1Q2	B1Q2	A2Q2	B2Q2
1.75	2.75	1.5	2
A1Q3	B1Q3	A2Q3	B2Q3
1.25	0.75	1.25	3.25
A1Q4-1	B1Q4-1	A2Q4-2	B2Q4-2
3	1	3	3

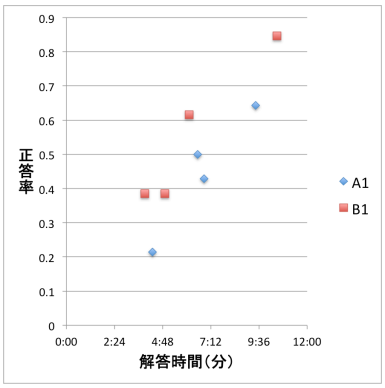

図 7: IT 広告をインターネットをしない人に届ける方法の時間と正答率

次に表 3 より B1Q3、B1Q4-1 を除くすべての項目でシソーラスを用いたときのほうが平均正答率が高い. このことからシソーラスを用いた方がより多くの構成要素を抽出できている. このことから実験仮説 2 は成立したと言える.

同様に表 3 より B1Q3、B1Q4-1 を除くすべての項目でシソーラスを用いたときのほうが平均正答率が高い。このことからシソーラスを用いたほうがより正確に構成要素を抽出できている. よって実験仮説 3 は成立したと言える.

実験仮説 4 を実証するために準備した設問は Q4-1,Q4-2 である. 表 3 より A1Q4-1 より B1Q4-1 のほうが正答率が悪いことがわかる. このことや実験後のアンケートから、指定された単語に関する情報はシソーラスより文章で読んだほうが文脈の関係から正確に読み取れるということがわかった. よって Q4-1,Q4-2 では実験仮説 4 は成立しなかった.

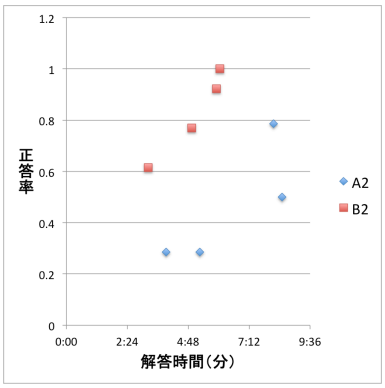

図 8: :ビデオ広告クリエイティブに伴う改善された広告の時間と正答率

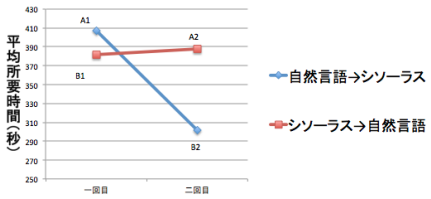

図 9: 平均所要時間の比較

5.2 提案手法の有効性

評価実験の結果から自然言語文書より提案手法により作成されたシソーラスを用いたほうが構成要素を早く、多くそして正確に抽出できることが分かった.

表 1 より機械的にシソーラスを作成したところ, 修正率が約 1 割となることが分かった. そのため, プログラムによる自動化が容易であると考えられる. よってシソーラス作成の工数が削減できるため短い時間でより多くの特許を理解できると考えられる.

また図 7,8 から所要時間に対しての正答率がシソーラスを用いたときのほうが良いのがわかる. このことから本手法で作成されたシソーラスを用いることで早く正確に既存特許を理解することができると考えられる.

さらには、図 9,10 からわかるように実験順序に関係なくどちらのグループもシソーラスを用いたほうが自然言語文書よりも時間効率、正答率について成績が良いことからも本手法のシソーラスが有効であると言える.

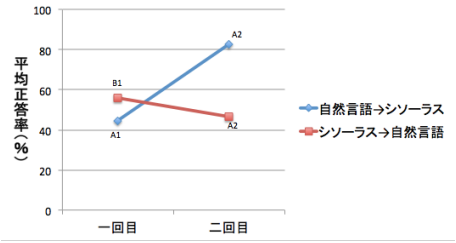

図 10: 平均正答率の比較

5.3 提案手法の限界

未分類動詞の割合が高いと利用率が低くなる。そのため表 1 の特許文書 E の利用率が低くなってしまった。利用率が低いと本来特許を理解するために必要な情報が抜け落ちている可能性がある.

出力動詞と考えられる動詞が非常に多いため, 今後自動化するにあたって事前に作成する動詞リストが膨大なものになってしまう可能性が挙げられる. また, 受動の関係により, 入力と出力が反転する場合がある. そのため, 現在の提案手法では修正率が全て 0%になることは無いと考えられる.

6 おわりに

6.1 まとめ

本稿では公開特許公報からシソーラスを作成する手法を提案し、その適用例を示した. 本手法を利用することで、自然言語文書より早く正確に特許を理解できることを示した.

6.2 今後の課題

本研究の適用実験は全て手作業で行った. この作業を削減するためにシソーラス自動作成ツールの作成を行わなくてはならない. 自動化ツールを作成するにあたっての課題は動詞の分類をどれだけ自動化できるかである. 現在はすべて手作業で分類を行っているため、分類規則を明確にすることで半自動化する予定である.

本研究では, 適用の対象広告分野のみと限定的であったため, 情報技術の他分野への適用及び評価が必要となる.

一部の被験者からシソーラスで入力要素、処理要素、出力要素がどこに記述されているか現在の記法ではわかりにくいという意見があった. そこで改善案として図 11 を考案した. 他にもカラーリングやアイコンの活用などを行うことで直感的に理解できるシソーラスにしていく予定である.

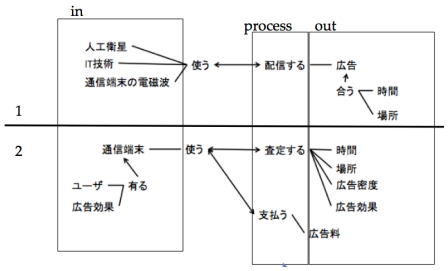

図 11: シソーラス改良案

また利用率を上げる試みとして、入力動詞、出力動詞、修飾動詞以外の動詞の分類や助詞の利用などを検討している. 自然言語文書と比較すると文脈の不足に不安を感じるという意見があったので文脈を補う工夫も課題である.

他の課題としては, 本研究で作成したシソーラスを起点とした新しい特許創造技法を作成することで, 特許情報の再利用性を高めることが挙げられる.

参考文献

[1] 和田恭, 米国の IT 企業における知的財産戦略の動向, http://www.ipa.go.jp/files/000006073.pdf, 2011

[2] 高木芳徳, トリーズ (TRIZ) の発明原理 40 あらゆる問題解決に使える [科学的] 思考支援ツール, ディスカヴァー・トゥエンティワン, 2014

[3] 谷川英和, 森本悟道 特許ライフサイクルに情報学を適用した新しい研究領域 2008 情報処理学会 VOL.49 No.4 p458 465

[4] Akio Kobayashi,Hirofumi Nonaka,Shigeru Masuyama,Hiroyuki Sakai An Automatic Thesaurus Construction Method for Technological Terms in Patent Maps 2010 Computers and Industrial Engineering (CIE), 40th International Conference

[5] 林祐介, 山本修一郎 フィーチャモデルを用いた IT 特許アイデア作成法の研究 2016 SIG-KSN-018