特許検索履歴を用いたシソーラスの自動構築

難波英嗣
広島市立大学大学院
情報科学研究科

竹澤寿幸
広島市立大学大学院
情報科学研究科

乾孝司
筑波大学大学院
システム情報工学研究科

岩山真
日立製作所
中央研究所

橋田浩一
産業技術総合研究所
社会知能技術研究ラボ

橋本泰一
東京工業大学
総合プロジェクト支援センター

藤井敦
東京工業大学
情報理工学研究科

1. はじめに

本稿では、特許データベースと特許検索履歴からシソーラスを自動的に構築する手法を提案する。シソーラスは、文献を検索したり特許や論文等の専門文書を執筆したりする上で、有用な情報源として活用されている。しかし、シソーラスを人手で構築し、更新することは非常にコストがかかるため、テキストデータベースからシソーラスを自動的に構築するという研究が近年活発に行われるようになってきた。テキストデータベースからシソーラスを構築する代表的な手法には、「A や B などの C」等の定型表現に着目して、用語の上位、下位概念を自動的に抽出するというものがある[Hearst 1992]。本研究では、このような既存の技術により抽出された知識を用いて特許検索履歴を解析することで、より網羅性の高い特許シソーラスを自動的に構築する。

知財の専門家による特許検索の履歴は、専門家の知識や検索経験の集約と考えることができ、上述の定型表現に着目した手法では得られない情報が含まれている可能性がある。他方、検索履歴には上位・下位関係や同義関係など、検索に用いた用語間の関係が明示されていないため、検索履歴だけを用いてシソーラスを構築することはできない。本研究では、特許テキストデータベース(以後、公開公報)から抽出した知識と検索履歴を用い、網羅的な特許シソーラスを構築する点が従来研究と異なる。

本論文の構成は以下のとおりである。次節では、検索履歴を用いたシソーラスの自動構築手法を提案する。3 節では、提案手法の有効性を調べるために行った実験について述べる。4 節では、関連研究について述べる。最後に 5 節で本論文をまとめる。

2. 特許検索履歴を用いたシソーラスの自動構築

2.1 シソーラス構築の手順

本研究で構築するシソーラスは、以下の 3 種類の関係を含んでいる。

(関係 1) 上位・下位関係
(関係 2) 効果表現リスト[1]
(関係 3) 同義関係

以下に、シソーラスの構築手順を示す。

(手順 1) 公開公報を解析し、関係 1 と 2 を抽出する。
(手順 2) 手順 1 の解析結果と検索履歴データを解析し、関係 3 および手順 1 で抽出されなかった新たな関係 1 と 2 を抽出する。

手順 1 および 2 について、2.2 節および 2.3 節で、それぞれ述べる。

2.2 上位・下位関係の抽出および効果表現リストの作成

(関係 1) 上位・下位関係の抽出

1 節で述べた Hearst ら[Hearst 1992]の定型表現法を用い Nanba は 10 年分の公開公報から以下に示す規模の上位・下位関係を抽出している[Nanba 2007]。

- 上位・下位関係(異なり数)：7,031,159 関係
- 全用語数(異なり数)：1,825,518 語

本研究では、このデータを用いてシソーラスの構築を行う。

(関係 2) 効果表現リストの抽出

国立情報学研究所主催の第 8 回評価ワークショップ NTCIR では、特許マイニングタスクが実施された[Nanba 2010:b]。このタスクでは、要素技術とその効果を示す表現を、特許から自動的に抽出することを目的のひとつとしている。例えば「PM 磁束制御用コイルを設けて閉ループフィードバック制御を施すため、電力損失を最小化できる。」という文が入力されると、図 1 に示すように、

[1] 効果表現リストとは、例えば「信頼性の向上」や「コストの低減」など、発明の効果に関する表現を収集して構築したリストのことである。

要素技術と効果を示す個所に、それぞれ“TECHNOLOGY”および“EFFECT”タグを自動的に付与する。ここで、“EFFECT”タグの中には、さらに“ATTRIBUTE”と“VALUE”という2種類のタグを自動的に付与するシステムを構築することが、このタスクで求められる。

PM 磁束制御用コイルを設けて<TECHNOLOGY>閉ループフィードバック制御</TECHNOLOGY>を施すため，<EFFECT><ATTRIBUTE>電力損失</ATTRIBUTE>を<VALUE>最小化</VALUE></EFFECT>できる.

図1　特許への要素技術と効果に関するタグ付与の例

我々は、上記のようなタグを自動付与するシステムをすでに構築している[Nanba 2010:a]。このシステムを用いて10年分の公開公報を解析し、自動的に抽出された属性(ATTRIBUTE)と属性値(VALUE)の対を「効果表現リスト」と呼ぶ。図2は、その一例である。各行、「頻度　属性　属性値」を示しており、2,599,368対(異なり)が抽出されている。このうち頻度3以上の96,435対を実験に利用する。

22393	信頼性	向上
21713	信頼性	高い
17362	構成	簡単
16870	生産性	向上
11283	作業性	向上
10522	操作性	向上
10376	コスト	低減
10175	製造コスト	低減

図2　効果表現リストの一例

2.3 同義関係の抽出

検索履歴中で、論理和で結合されている用語は同義関係にある可能性が高いが、そうでないケースも少なからず存在する。例えば、以下の例1-3の場合、例1はすべて同義語であるが、例2は上位語と下位語が、例3は属性と属性値が混在している。なお、本稿において「+」は論理和演算、「*」論理積演算を示す。

(例1) ジャガイモ+じゃがいも+バレイショ+馬鈴薯+ばれいしょ
(例2) 植物+茸+きのこ+キノコ
(例3) 解像度+デグレード+低下+劣化

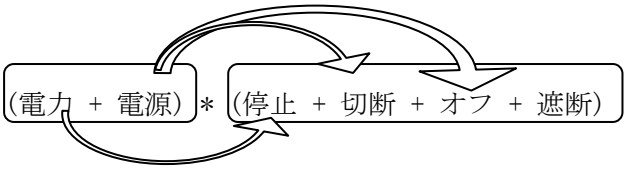
そこで、検索履歴の中で、論理和で結合されている一連の用語がすべて同義語である可能性の高い個所(上記の場合は例1)を見つけ、その個所から同義語を抽出する。このような個所の検出に、2節で述べた既存の上位・下位関係(関係1)と効果(関係3)を用いる。例えば、以下のような検索履歴を考える。

（電力 + 電源）＊（停止 + 切断 + オフ + 遮断）

また、効果表現リスト中に次の関係が含まれているとする。

「電源－オフ」「電源－切断」「電力－停止」

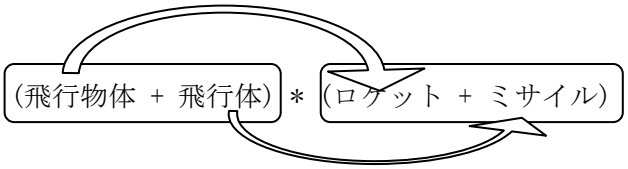
この時、「(電力 + 電源)」ブロックと「(停止 + 切断 + オフ + 遮断)」ブロックは属性と属性値の関係にあると推測される。



ここから、以下の知識が新たに得られる。

同義語：電力－電源
同義語：停止－切断－オフ－遮断
効果：電力－切断、電力－オフ、電力－遮断
効果：電源－停止、電源－遮断

上位・下位関係に関しても基本的には効果表現リストと同様の手順で新たな上位・下位関係を得ることができる。一方、同義関係については抽出しない。なぜならば、上位・下位関係の場合、同義関係でないものが誤って抽出される可能性があるからである。例えば、以下の例において「(飛行物体 +飛行体)」ブロックと「(ロケット+ミサイル)」ブロックが上位・下位関係にあることが分かっている場合、ここから新たに「飛行物体－ミサイル」と「飛行体－ミサイル」を新たな上位・下位関係として抽出できる。しかし、この場合、「ロケット」と「ミサイル」は同義関係にはない[2]。



なお、2つのブロック間に同一の関係Xが2回以上出現した場合は、ブロック間にXという関係があると判断する。

[2] 実際のデータを用いて試してみたところ、正しい同義表現が数多く抽出された一方で、「XML」と「SGML」と「HTML」なども同義表現として抽出された。

3. 実験

2 節で提案した手法のうち、効果表現リストと検索履歴を用いたシソーラス構築手法の有効性を調べるため、実験を行った。

3.1 実験方法

実験に用いるデータ

- 検索履歴データ：一般財団法人工業所有権協力センターにおいて 2001 年度から 2008 年度までの間に知財の専門家が作成した検索報告書に記載された検索式約 61 万件
- 特許全文データ：公開公報 1993-2002 年(3,496,252 文書、94.5GB)

評価方法

以下の 2 点において評価を行った。

(評価 1) 属性－属性値の抽出

提案手法により「属性－属性値関係にある」と判定された 2 つのブロックが、実際に属性－属性値の関係にあるものの割合

(評価 2) 同義関係の抽出

提案手法により「属性－属性値関係にある」と判断された 2 つのブロックそれぞれから抽出した任意の用語対が全て同義表現になっているものの割合

単純にブロック内の論理和で結合されている用語がすべて同義関係にあるか否かで評価するのではなく、評価 2 のような方法をとっている理由は、特に属性値と判断されたブロック内の用語を評価する場合、ブロック内の用語だけでは判断できない同義関係が数多く存在するからである。例えば、2.3 節に挙げた例において、「遮断」と「停止」が同義関係にあるかどうかは、両者が「電源」や「電力」といった語と属性－属性値の関係にある、という情報がなければ人間の被験者が判断しづらいと考えられるからである。

3.2 実験結果および考察

提案手法を用いて抽出された効果表現のうち、122 件について人手で評価を行った。評価結果を以下に示す。

(評価 1) 属性－属性値の抽出：89.3% (109/122)
(評価 2) 同義関係の抽出　　：81.9% (100/122)

図 3 は、提案手法を用いて、実際に正しく解析された例である。図 3 において、“ATTRIBUTE”と“VALUE”タグは提案手法により自動的に挿入されたものである。また、ATTRIBUTE と VALUE タグ間には対応関係を示す id が付与されている。例えば、id=1 で ATTRIBUTE タグ付与されている「電源」は VALUE タグが付与されている「切断」と対応関係にあることを示す。この関係は、2.2 節で述べた効果表現リストに含まれていた「電源－切断」というデータから導きだされたものである。

```
(<ATTRIBUTE id="1&2">電源
</ATTRIBUTE>+電力+パワー)*
(オフ+ＯＦＦ+<VALUE id="1">切断
</VALUE>+<VALUE id="2">遮断</VALUE>+
停止)
```

図 3　正しく解析できている例

図 3 の例において、この他に、「電源－遮断」も 2.2 節で述べた効果表現リストに含まれていたことから、提案手法は図の下線部と波線部のブロック間に属性－属性値の関係があると判断している。この結果、効果表現リストには、新たに「電源－オフ」「電力－オフ」「パワー－遮断」などのデータが追加されることになる。

次に、解析誤りについて考察する。評価 1 に関する誤りは以下の 2 種類に分けられる。

(原因 1) ブロック内に関係のない用語が混在している(8 件)
(原因 2) 元の効果表現リストに誤りがある(1 件)

以下に、失敗例をそれぞれ示す。

(原因 1)ブロック内に関係のない用語が混在している(8 件)

図 4 において、提案手法は下線部と波線部のブロック間に属性－属性値の関係があると判断した。しかしながら、VALUE ブロック内(波線部)には、「検知」と「検出」のような同義関係にある用語の他に「ログ」や「履歴」など、同義関係とは考えにくい用語も含まれていたため、下線部と波線部ブロック間には属性－属性値の関係がないと人間の被験者が判断した。

```
(<ATTRIBUTE id="0&1&2">エラー
</ATTRIBUTE>+<ATTRIBUTE id="3&4&5">
異常</ATTRIBUTE>+<ATTRIBUTE
id="6&7&8">障害</ATTRIBUTE>+イリーガル+
アラーム+アラート)*
(オペ+コメント+リトライ+再+<VALUE
id="0&6">回復</VALUE>+リカバ+ログ+履歴
+<VALUE id="3">検知</VALUE>+<VALUE
id="1&4&7">検出</VALUE>+<VALUE
id="2&5&8">発生</VALUE>+通知+報知+表示)
```

図 4　評価 1 の解析誤り例(その 1)

(原因 2) 元の効果表現リストに誤りがある (1 件)

図 5 において、元の効果表現リストに「ファイル

ー高速」という、本来ならば属性ー属性値の関係にないデータが含まれていたため、「ファイル」と「高速」にそれぞれ ATTRIBUTE と VALUE タグが誤って付与されている。

(<ATTRIBUTE id="1"> ファイル </ATTRIBUTE>+<ATTRIBUTE id="2"> 事故 </ATTRIBUTE>)*
(<VALUE id="1">高速</VALUE>+迅速+短縮 +<VALUE id="2">低減</VALUE>)

図 5　評価 1 の解析誤り例(その 2)

次に、評価 2 の解析誤りについて考察する。評価 2 に関する誤りは以下の 2 種類に分けられる。

(原因 1) 評価 1 における(原因 1)と同じ (8 件)
(原因 2) ブロック内の反義語や兄弟語 (8 件)

図 6 に、原因 2 の失敗例を示す。図 6 では、下線部と波線部内の任意の 2 語は属性ー属性値の関係にあると考えられるが、下線部内で「圧縮」と「解凍」は反義語であるため、同義関係にないと人間の被験者が判断した。

(効 果 +<ATTRIBUTE id="0&2"> 効 率 </ATTRIBUTE>+ 有 効 性 + 能 率 + 性 能 +<ATTRIBUTE id="1">圧縮率</ATTRIBUTE>+ 圧縮性能+圧縮効果)*
(<VALUE id="0&1">圧縮</VALUE>+凍結+解凍+コンプレッション+コンプレション+<VALUE id="2">コンパクト</VALUE>+コンパクタ+コンパクション+シュリンク)

図 6　評価 2 の解析誤り例

4. 関連研究

これまでに、Web の検索履歴から知識獲得を行う数多くの研究が行われている[小町 2008, Pasca 2007, 関口 2010]。その代表的な手順は、以下のとおりである。

(1) 「東京　大阪　名古屋」など、性質の似た用語集合をシードとしてシステムに与える。
(2) 各シードの用語と共起する語句を、2 語から構成される検索質問の履歴データから抽出する。
(3) 手順 2 で抽出された語集合と(検索履歴データ内で)共起頻度の高い語句を手順 1 で入力したシードの関連語として出力する。

ここで、手順 2 で抽出される共起語の多くは手順 1 で入力されるシードの属性(例えば、この例の場合「航空券」や「名所」など)であり、用語の属性を考慮する、という点においては、本研究と関連がある。しかしながら、これまで例に示したとおり、特許検索では数多くの用語を論理和と論理積で組み合わせて用いられることが一般的であり、Web の検索履歴を対象にした手法をそのまま適用することはできない。また Web の検索履歴を対象にした既存研究では人間がシードを与えることを前提にしているが、本研究では、シードの代わりに公開公報から自動的に抽出した知識を用いている点も既存研究と異なる。

5. おわりに

本稿では、公開公報から抽出した知識を用いて検索履歴を解析し、網羅的な特許シソーラスを構築する手法を提案した。実験の結果、81.9%の解析精度で特許シソーラスが構築できることが分かった。

謝辞

本研究を実施するにあたり、一般財団法人工業所有権協力センターから、特許検索履歴データを提供して頂きました。深く感謝致します。

参考文献

[Hearst 1992] Hearst, M.A. (1992) “Automatic Acquisition of Hyponyms from Large Text Corpora”. In Proceedings of the 14th International Conference on Computational Linguistics, pp.539-545.

[小町 2008] 小町守, 鈴木久美 (2008) “検索ログからの半教師あり意味知識獲得の改善”. 人工知能学会論文誌, Vol.23, No.3.

[Nanba 2010:a] Nanba, H., Kondo, T., and Takezawa, T. (2010) “Automatic Creation of a Technical Trend Map from Research Papers and Patents”. In Proceedings of the 3rd International CIKM Workshop on Patent Information Retrieval (PaIR'10), pp.11-15.

[Nanba 2010:b] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. (2010) “Overview of the Patent Mining Task at the NTCIR-8 Workshop”. In Proceedings of the 8th NTCIR Workshop, pp.293-302.

[Nanba 2007] Nanba, H. (2007) “Query Expansion using an Automatically Constructed Thesaurus ”. In Proceedings of the 6th NTCIR Workshop, pp.414-419.

[Pasca 2007] Pasca, M. (2007) “Organizing and Searching the World Wide Web of Facts Step Two: Harnessing the Wisdom of the Crowds”. In Proceedings of WWW 2007, pp.101-110.

[関口 2010] 関口裕一郎, 田中智博, 内山匡, 藤村滋, 望月崇由, 鈴木智也 (2010) “検索クエリログのセッション情報を利用した属性語句抽出”. DEIM Forum.