社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2003－FI－ 72　(9)
2003－NL－157　(9)
2003／9／29

特許文献を用いた因果関係に基づく知識構造化の試み

石川大介†　石塚英弘‡　宇陀則彦‡　藤原譲⋆

図書館情報大学大学院情報メディア研究科†
図書館情報大学図書館情報学部‡
独立行政法人工業所有権総合情報館⋆

特許文献には、発明の手段とその結果もたらされる効果が記述されている。我々は、その手段－効果の関係を因果関係と見なし、繊維工学の分野の特許文献から文型パターンと用語リストを用いて因果関係を抽出した。今回の実験では化合物とその性質の因果関係を抽出した。また、これらの因果関係から抽出した知識の構造化を試みた。今回は、特に接着性がある化合物に関する事例を報告する。

An attempt to structure knowledge extracted from causal relationship described in patent document

Daisuke Ishikawa†　Hidehiro Ishizuka‡　Norihiko Uda‡　Yuzuru Fujiwara⋆

Graduate School of Information and Media Studies, University of Liblary and Infomation Science†
Faculty of Liblary and Infomation Science, University of Liblaryand Infomation Science‡
National Center for Industrial Property Information⋆

A method in an innovation and effects caused by it are described in patent document. We regard the method-caused effect as causal relationship, therefore we extracted some causal relationships from patent documents in fiber engineering using phrase-pattern and term list. Some causal relationships between compound and its property are obtained. We also attempted to structure knowledge extracted from the causal relationships. The case of adhesive compound will be reported.

1　はじめに

近年、電子化された文書が氾濫するようになったため、これらの大量なテキストから有用な情報を抽出する目的で、テキストマイニングの研究が行なわれている [1]。特に、ライフサイエンス分野に対象を絞ったテキストマイニングが盛んに研究され [2]、主に学術文献を対象にして遺伝子や病気に関する因果関係の抽出が行なわれている。また、テキストマイニング技術を応用した特許検索・分析支援システム：ACCENT がある [3]。これは企業別の特許出願傾向やキーワード間の関連性を可視化し、特許の概要把握と比較分析を支援するツールである。

特許とは、自然法則を利用した産業利用可能な発明を保護する制度であり、特許文献にはその具体的な手段と、その発明によりどのような効果がもたらされるのかが記載されている [4]。本研究ではこの手段に関する記述と効果に関する記述の関係を因果関係と考え、特許で扱われている科学技

術分野を対象として因果関係の抽出を行なった。これらの因果関係から、思考に必要な汎用的な知識を得られるものと考えている。

我々は前の研究 [5] において、「ことにより」という表現に注目し、自動車工学の分野を対象に因果関係の抽出を行なった。また、抽出した用語を解析し、その結果を利用することによって、「ことにより」の表現が使われない文献も解析し、効果の記述も抽出した [6]。

本研究では、繊維工学の分野を対象に、同じく「ことにより」の表現を用いて、予め抽出対象とする用語を辞書に登録し、その辞書を利用して因果関係を抽出した。また、得られた因果関係の統合方法についても議論し、知識の構造化を試みた。

2 対象テキストの生成

2.1 元データ

実験で使用するデータは、NTCIR-3 [7] の「JAPIO 出願抄録データ 98(日本語)」である。これは、JAPIO((財) 日本特許情報機構) により作成された特許抄録コーパスであり、公開公報には、もともと出願人により要約が付与されているが、JAPIO 出願抄録は、これらを専門家が (必要であれば) 修正したものである [8]。

このデータは一行に一件の特許出願抄録が収録され、図 1 のようにタグ付けされている。

一件の抄録データにつき、出願番号、出願人、国際特許分類などの基本的な書誌情報と、タイトル、キーワード、概要、詳細のテキスト情報がタグ付けされて収録されている。テキスト部分のタグは以下の通りである。

```
<SDOAB LA='J'>
<P>概要　£ キーワード</P>
<P>詳細な内容</P>
</SDOAB>
```

各抄録データの概要の部分を実験の対象テキストとした。

2.2 対象領域の選定

特許抄録コーパスの中から、国際特許分類 IPC を利用して分野を選定した。今回の研究では繊維工学の領域を対象とした。具体的には各特許抄録に付与されている IPC コードに「D06M」を含むものを対象とした。IPC コード「D06M」とそのメインクラスの説明を、特許庁 [4] で配布されている IPC 第 7 版から抜粋し、簡単に表 2 にまとめた。

IPC コード	内容
D06	繊維または類似のものの処理
D06M	繊維の他に分類されない処理
D06M 10	物理的処理 (例：超音波、コロナ放電) 化学的な処理と組み合わせた物理的処理
D06M 11	無期物質またはその錯体による処理
D06M 13	非高分子有機化合物による処理
D06M 14	炭素-炭素不飽和結合を含有する 単量体のグラフト重合
D06M 15	高分子化合物による処理
D06M 16	生化学的処理 (例：酵素)
D06M 17	多層織物の製造
D06M 19	羽毛の処理
D06M 23	プロセスに特徴のある処理 (例：泡やエアロゾルの処理剤、片面処理)

図 2: IPC コード「D06M」

この IPC コードを持つものを特許抄録コーパス 34 万件から検索し、見つかった文献 863 件を対象文献とした。

2.3 語分割テキストの生成

対象文献の対象テキスト部分について、「ことにより」を含む文献 541 件を対象に、形態素解析システムの茶筌 [9] を利用し、形態素解析を行なった。

ここで、「ことにより」が複数使われていた文献は 2 件あるが、それは対象外とした。また、句点が対象テキストの文末部分以外にも使用されている文献は 3 件あるが、それは文頭から最初の句点が見つかるまでを処理対象とした。

形態素解析した結果について、「名詞、未知語、記号、接頭辞」をひとまとまりにして用語とみなし、［と］でマーク付けした。以下、このマーク付けされた部分を本稿では用語と呼び、それ以外の形態素の集合を語と呼ぶことにする。

「ことにより」と「ことにより、」は「->」に変換し、「、」は「,」に変換し、文末の句点は取り除いた。マーク付けと変換の処理の様子を図 3 に示す。

この処理したテキストを、語分割テキストと呼ぶこととする。この語分割テキストは、対象テキストの記述を「ことにより」という表現を手掛か

<PATDOC>
<B210>1996244060</B210>
<B220>19960827</B220>
<B110>1998072774</B110>
<B140>19980317</B140>
<B711>東レ　ダウコーニング　シリコーン　（株）</B711>
<B721>石川　裕規</B721>
<B721>長縄　努</B721>
<B721>小名　功</B721>
<B511>D06M 13/513　　</B511>
<B511>D06M 15/647　　</B511>
<B541>繊維処理剤用添加剤および繊維処理剤</B541>
<SDOAB LA='J'>
<P>（Ｊ）特定の有機ケイ素化合物を含有させることにより、繊維に対して湿潤性、浸透性を著しく向上させる繊維処理剤用添加剤を得る。￡風合、平滑性、ポリプロピレン、ポリエステル、ナイロン</P>
<P>式†１〔Ｒは（置換）一価炭化水素；Ｒ↑１は二価炭化水素；Ｒ↑２はＨ，Ｃ↓１↓～↓５一価炭化水素、Ｒ↑３ＣＯ－（Ｒ↑３はＣ↓１↓～↓５一価炭化水素）；ｎは４～１５〕で表わされる、式†２のオルガノトリシロキサン等の化合物を含有させる。尚、この繊維処理剤用添加剤を０．００５～１．０重量含有させて繊維処理剤を調製するのが好ましく、また、この繊維処理剤は、エマルジヨンまたは水溶液の形態で使用される。</P>
</SDOAB>
</PATDOC>

図 1: データの中身

（Ｊ）特定の有機ケイ素化合物を含有させることにより、繊維に対して湿潤性、浸透性を著しく向上させる繊維処理剤用添加剤を得る。

⇓

[（Ｊ）特定] の [有機ケイ素化合物] を [含有] させる -> [繊維] に対して [湿潤性] , [浸透性] を著しく [向上] させる [繊維処理剤用添加剤] を得る

図 3: マーク付けと変換の処理による語分割テキストの生成

りに、手段の記述の部分と効果の記述部分を以下のような形式にしたと言える。

「手段の記述部分 -> 効果の記述部分」

以下、この語分割テキストを用いて実験を行なった。

2.4　語の使用頻度

生成された語分割テキストについて、手段の記述部分に使用される用語と語、そして効果の記述部分に使用される用語と語の、それぞれの使用頻度 (上位 10 件まで) と合計数、異なり数を調べた。図 4 にその結果を示す。

3　化合物とその性質の因果関係抽出

図 2 に示されているように、今回実験で対象とする分野では、化合物を使用した処理がある。そして、概要の具体例を見ると、その化合物を利用したことによって、得られる効果が「～性」という表現を用いて記述されている。この化合物とその効果に関する関係を因果関係と考え、その抽出を試みた。

3.1　化合物名の抽出

語分割テキストの手段の記述部分から、化合物を表す用語を抽出するために、「[～化合物]」という用語を抽出し、これを化合物名辞書に登録した。ただし、[化合物] と [特定化合物] という用語は、化合物名を表していない用語であるため、除外した。辞書に登録された用語は 68 件である。これを用いて以下の実験を行なった。

なお、化合物名のうち頻度順に並べた上位 10 件を図 5 に示す。

使用頻度	手段の語
581	を
556	の
379	する
372	,
303	に
170	と
160	で
115	させる
101	した
97	及び
合計	3968
異なり	298

使用頻度	手段の用語
119	[（J）特定]
102	[特定]
85	[含有]
48	[付与]
45	[形成]
43	[処理]
41	[繊維]
39	[後]
29	[表面]
27	[布]
合計	3677
異なり	2013

使用頻度	効果の語
736	,
374	を得る
340	の
315	を
218	な
209	に
144	する
131	に優れた
94	や
73	で
合計	4236
異なり	454

使用頻度	効果の用語
75	[こと]
63	[有用]
60	[向上]
58	[付与]
56	[耐久性]
44	[防止]
35	[良好]
35	[好適]
33	[風合]
31	[風合い]
合計	3686
異なり	1861

図 4: 手段と効果の語と用語の頻度

使用頻度	用語
10	[化合物]
4	[無機系化合物]
4	[特定化合物]
4	[ヒドラジド化合物]
4	[シリコーン系化合物]
3	[ポリエーテル化合物]
2	[有機化合物]
2	[環状尿素化合物]
2	[ポリエポキシド化合物]
2	[ポリウレタン化合物]
合計	99
異なり	70

図 5: 化合物名の高頻度出現語 (一部)

使用頻度	用語
56	[耐久性]
28	[柔軟性]
22	[洗濯耐久性]
18	[性]
17	[接着性]
17	[抗菌性]
13	[耐熱性]
12	[防縮性]
12	[消臭性]
12	[吸水性]
合計	671
異なり	259

図 6: 性質を表す用語の高頻度出現語 (一部)

3.2 性質を表す用語の抽出

語分割テキストの効果の記述部分から、性質を表す用語 (以下、性質表現用語) を抽出するために、「[〜性]」という用語を抽出した。これを頻度順に並べた一部を図 6 に示す。

これらの用語 259 件のうち結果の [性] のみの用語は、語分割テキスト生成の処理時に、正しく用語としてマーク付けされなかった用語であるため、除外し、258 件の用語を辞書に登録した。これを性質表現用語辞書として実験に使用した。

3.3 化合物名と性質名の抽出

化合物名辞書と性質表現用語辞書を利用し、語分割テキストから用語の抽出を行なった。手段の記述部分からは化合物名辞書に登録されている用語を抽出し、効果の記述部分からは性質名辞書に登録されている用語を抽出した。そして、手段と効果の両方から少なくても一つ以上の用語が抽出されているものを最終的に抽出した。541 件の語分割テキストの中から、52 件の文献について化合物名とその性質の記述を抽出した。この抽出結果を、化合物-性質関係と呼ぶこととする。

4 汎用語の除去による因果関係抽出

図 3 の手段の記述では、手段として使用された材料は「[有機ケイ素化合物]」であり、他の手段の用語「[（J）特定]」と「[含有]」は材料を表していない。これらの材料を表していない用語は図 4 が示すように、高頻度で使用される汎用語である。そこで、これらの材料を表さない用語を除去することで、残った用語を材料と考え、その材料とその性質の関係の抽出を試みた。

4.1 汎用語除去辞書の生成

手段の記述において、汎用的に利用される用語は図 4 の通りである。ここから、高頻度で使用される上位 4 位までの用語「[（Ｊ）特定]、[特定]、[含有]、[付与]」を汎用語除去辞書に登録し、これを以下の実験で使用した。

4.2 汎用語除去辞書を利用した抽出

汎用語除去辞書と性質名辞書を利用し、語分割テキストから用語の抽出を行なった。手段の記述部分からは汎用語除去辞書に登録されている用語を消去して残った用語を抽出した。効果の記述部分からは性質名辞書に登録されている用語を抽出した。手段の用語の数は一つだけ抽出されているものであって、かつ、効果の用語が少なくても一つ以上抽出されているものを抽出した。541 件の語分割テキスト中、10 件の抽出結果を得た。この抽出結果を未特定材料-性質関係と呼ぶこととする。

5 抽出結果

5.1 化合物-性質関係の調査

化合物名辞書と性質名辞書を利用して抽出した化合物-性質関係が、元の文献の記述から表現の上で正しいかどうかを目視により調査した。手段の記述部分のみを調査対象に、手段に使われる化合物や材料を示す用語が全て抽出されたかを調べた。その結果、52 件の関係中、16 件の関係が、元の文献の表現からその関係が正しいと判断された。正しいと判断された文献の例を以下に示す。なお、文献番号とは、処理の対象文献 863 件の中での通し番号である。

文献番号 205：東レ ダウコーニング シリコーン（株），特願 1996-244060
（Ｊ）特定の 有機ケイ素化合物 を含有させることにより、繊維に対して湿潤性、浸透性を著しく向上させる繊維処理剤用添加剤を得る。

この場合の化合物-性質関係は次の形式で示すことができる。

有機ケイ素化合物 → 湿潤性 , 浸透性

正しいと判断できない結果もあった。それは元の文献の表現上からは抽出した化合物のみが、抽出したその性質に起因するとは断定できないものである。以下、その例を示す。

文献番号 144：花王（株），特願 1996-211050
（Ｊ）特定の アンモニウム塩 と アスパラギン酸化合物 を配合することにより、抗菌活性が高く、刺激性が少ない、繊維、床、壁等の抗菌剤として有用な組成物を得る。

下線部分が抽出対象の材料を表す用語である。抽出結果では「アスパラギン酸化合物」のみが抽出されているが、本文では「アンモニウム塩」も得られた性質に関係していることが分かる。そのため、「アスパラギン酸化合物」だけが抗菌活性と刺激性に関係するとは断定できない。

5.2 未特定材料-性質関係の調査

汎用語除去辞書と性質名辞書を利用して抽出した未特定材料-性質関係が、元の文献の記述から表現の上で正しいかどうかを目視により調査した。その結果、10 件の関係中、8 件の関係が、元の文献の表現からその関係が正しいと判断された。

正しいと判断された化合物-性質関係と、未特定材料-性質関係の抽出結果の一部を、抽出した原文と出願人、出願番号と合わせて図 7 に示す。なお、抽出結果の用語の「[]」や「(Ｊ)」は取り除いてある。

残り 2 件は未特定材料が「親水性処理剤」と「エマルジョン」である。これらは、特定の材料名を示していない特許特有の表現方法であるが、前者は明らかに何らかの物質を暗示している。また、後者の「エマルジョン」は国語辞典：広辞苑によれば「液体状の微粒子がこれと混合しない他の液体中に分散して乳状をなすもの。」と書かれてあり、この「微粒子」が物質を暗示している。従って、この 2 件の結果も特に専門知識の無い人間が正しいと判断できる範囲にある。

なお、未特定材料が「[〜化合物]」の場合も 3 件抽出された。これらの抽出結果は前述の化合物-性質関係で得た抽出結果と重複する。

6 考察

今回実験で抽出した結果の統合について議論する。まず、抽出した効果の関係の中から、「接着性」という用語が抽出されている関係を全て列挙した。そして、これらの関係の中で、化合物名が

化合物、未特定材料　→　性質	文献番号
カルボキシル変性ポリシロキサン　→　精練性	155
三洋化成工業　（株）, 特願 1996-219495 （J）特定のカルボキシル変性ポリシロキサンを用いることにより、繊維同士の膠着が少なく精練性や染色ムラの発生を抑制しポリウレタン弾性繊維の紡糸から後加工まで安定な操業ができる弾性繊維用油剤を得る。	
膨潤性粘土鉱物　→　柔軟性	199
一方社油脂工業　（株）, 特願 1996-244260 （J）膨潤性粘土鉱物を含有させることにより、繊維類や染料等に対する影響もなく、スレ・アタリ・シワの発生を防止し、柔軟性、風合、品質に優れた繊維類の得られる、標記加工処理剤を得る。	
有機ケイ素化合物　→　湿潤性 , 浸透性	205
東レ　ダウコーニング　シリコーン　（株）, 特願 1996-244060 （J）特定の有機ケイ素化合物を含有させることにより、繊維に対して湿潤性、浸透性を著しく向上させる繊維処理剤用添加剤を得る。	
3級アミン化合物　→　柔軟性 , 環境安全性	551
花王　（株）, 特願 1997-008199 （J）特定の3級アミン化合物等の所定量を含有させることにより、優れた柔軟性及び弾力性（ふつくら感）を付与でき、環境安全性の高い衣料用柔軟仕上剤組成物を得る。	
ピレスロイド系化合物　→　害虫忌避性 , 洗濯耐久性	568
帝人　（株）, 特願 1997-021499 （J）ポリエステル繊維を、ピレスロイド系化合物を含有する液中で、特定の条件で加熱処理することにより、風合、害虫忌避性及びその洗濯耐久性が良好で、蚊、ブヨ、ノミ、ダニ、シラミ等の害虫による被害を防止しうるポリエステル繊維構造物乃至繊維製品を得る。	

図 7: 抽出した化合物-性質関係と未特定材料-性質関係 (一部)

同一か、部分一致するもの同士を矢印で結んだ。その結果を図 8 に示す。

同一の化合物「ポリエポキシド化合物」は双方向の矢印で結び、化合物名の図の下線部が部分一致する「ポリエポキシド化合物」と「芳香族ポリエポキシド化合物」、及び「ブロツクドイソシアネート化合物」と「ブロツクドポリイソシアネート化合物」とを片方向矢印で結んだ。そして、矢印で結ばれた用語のまとまりを枠で囲った。

こうして統合した結果、以下の点が言える。

1. ポリエポキシド化合物と芳香族ポリエポキシド化合物に接着性があるため、共通するポリエポキシド化合物の部分に接着性があると推定される。
2. ブロックドイソシアネート化合物とブロックドポリイソシアネート化合物に接着性があるため、共通するイソシアネート化合物の部分に接着性があると推定される。

このようにして、得られた化合物を接着性という因果関係に基づいて統合した。ここから、個別の因果関係を知識として構造化することが可能と考えられる。

7　おわりに

本研究では、特許コーパスの概要に記述されている、手段と効果に関する記述部分において、その間の関係を因果関係と考えた。そして、「ことにより」という表現を手掛かりに、手段と効果からそれぞれ抽出対象とする用語辞書を作成し、それを利用して因果関係の抽出を行なった。また、抽出対象としない不要な用語を予め辞書に登録し、それを使った抽出も行なった。得られた因果関係から、化合物名の完全一致、部分一致によってそれぞれを結ぶことによって統合を試み、その結果について議論した。これらによって、特許文献から抽出された科学技術に関する因果関係を知識として構造化することを試みた。

今回の実験では文献の内部から得られる用語を利用して辞書を生成した。この辞書の生成は、例えば JICST シソーラス [10] など、外部から得られる用語を利用することも考えられる。シソーラスには用語間の上位下位関係の構造が記述されているため、この構造を辞書に取り入れることも可能である。また、特許文献で扱われる科学技術分野間の統合も考えられ、これは現在研究中である。

化合物	→	性質	文献番号
芳香族ポリエポキシド化合物 , ブロツクドポリイソシアネート化合物	→	接着性	139
ブロツクドイソシアネート化合物	→	接着性	320
エチレン尿素化合物	→	接着性	411
ポリエポキシド化合物	→	接着性	707
イミダゾール化合物 , シラン化合物	→	接着性	731
ポリエポキシド化合物 , ポリアリルアミン化合物	→	接着性	736

⇓

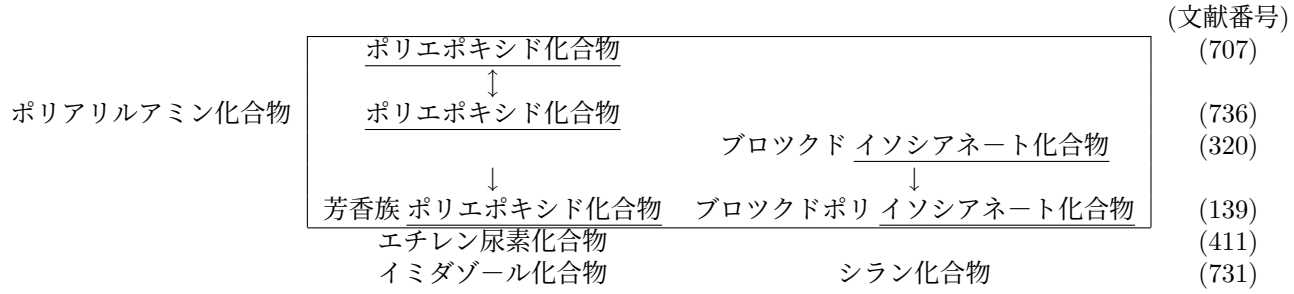

図 8: 接着性のある因果関係と、それらの統合

これらの研究は、最終的に学習・思考機能の開発 [11] につながるものと考えている。

謝辞　本研究において、国立情報学研究所で作成された NII-NACSIS コレクションの NTCIR-3 を使用しました。深く感謝いたします。

参考文献

[1] 市村由美、長谷川隆明、渡部勇、佐藤光弘：テキストマイニング-事例紹介、人工知能学会誌、Vol.16, No.2, pp.192-200, 2001.

[2] 浦本直彦、松澤裕史、猪口明博、武田浩一：ライフサイエンス分野におけるテキストマイニング技術適用の動向、情報処理学会研究報告 2003-FI-71(4), pp.25-32, 2003.

[3] 渡部勇：富士通研究所による特許検索・分析支援システム「ACCENT」、INFOSTA シンポジウム 2002 予稿集, pp.7-12, 2002.

[4] 特許庁：http://www.jpo.go.jp/indexj.htm

[5] 石川大介、石塚英弘、宇陀則彦、藤原譲:特許文献における因果関係の抽出、第 11 回研究報告会講演論文集、情報知識学会、pp.31-38、2003

[6] 石川大介、石塚英弘、宇陀則彦、藤原譲:特許文献における因果関係の抽出と用語の解析、情報科学技術フォーラム (FIT2003), D-003, 2003 (発表予定)

[7] NTCIR：http://research.nii.ac.jp/ntcir/index-ja.html

[8] 岩山真, 藤井敦, 高野明彦, 神門典子：特許コーパスを用いた検索タスクの提案、情報処理学会研究報告 2001-FI-63, pp.49-56, 2001.

[9] 形態素解析システム 茶筌：http://chasen.aist-nara.ac.jp/index.html.ja

[10] 日本科学技術情報センター編：JICST 科学技術用語シソーラス, 日本科学技術情報センター, 1993.

[11] 藤原譲：情報学基礎論の現状と展望-学習・思考機構と超脳計算機への応用-、情報知識学会誌、vol.9、no.1、pp.13-29、1999