3D-5

概念グラフを用いた特許明細書群からの関連語抽出

下司 義寛 † 和多 太樹 † 廣川 佐千男 ††
† 九州大学大学院システム情報科学府 †† 九州大学情報基盤センター

1 はじめに

一般の Web ページや文書検索では, キーワードで検索する. これに対し特許明細書についての専門家の検索は, FI や F タームを効率よく利用する. これらの分類コードは, 特許庁の特許審査官が付与するもので, 多少の揺らぎはあるとしても, 先行事例の網羅的検索や絞り込みのための限定には非常に有効である. しかし, それらの分類コードは, 多岐にわたり特許の専門家であっても全てを把握しているわけではない. いずれにせよこのような分類コードを活用できるか否かが専門家と非専門家の違いと言われている. 専門家であっても, 予備的調査においてはキーワード検索を行い, 関連する FI を獲得することが多い.

本稿では, 著者等が提案する概念グラフの手法を特許明細書文書群に適用し, キーワードに対する FI 群, あるいは 1 つの FI にかかわるキーワード群を自動的に抽出するシステムを構築した.

2 概念グラフ

筆者等は, シソーラスを自動構築する手法, 概念グラフシステムを提案してきた [1], [3], [4]. 一般のシソーラス構築技術では与えられた文書全体での単語の関係を抽出する. 概念グラフでは, ユーザの与えるクエリ q によって決まる文書集合 $D(q)$ に着目して特徴語の抽出を行い, $D(q)$ の文書の中での特徴語の上位下位関係を抽出する.

2.1 特徴語抽出

ある単語 w の $D(q)$ での特徴量 $s(w, D(q))$ を式 (1) で定義する.

$$s(w, D(q)) = \frac{df(w, D(q))}{df(w, U)} \tag{1}$$

この式は単語 w を含む文書集合 D(w) と D(q) の交わりの文書集合 $D(w)$ での割合を表している。また,D(w) と D(q) の交わりの $D(q)$ での割合を w のもう一つの特徴量と考える。

$$s_2(w, D(q)) = \frac{df(w, D(q))}{|D(q)|} \tag{2}$$

$s(w, D(q))$ かつ $df(w, U) \leq |D(q)|$ となる単語 w を $D(q)$ の下位の特徴語, $s_2(w, D(q))$ かつ $df(w, U) > |D(q)|$ となる単語 w を $D(q)$ の上位の特徴語と呼ぶ.

2.2 上位下位関係抽出

特徴語間の関連抽出には二つの特徴語 u, v の $D(q)$ での共起頻度 $df(u * v, D(q))$ を用いて, v からみた u の関連度 $r(v, u)$ として定義する。

$$r(v, u) = \frac{df(u * v, D(q))}{df(v, D(q))} \tag{3}$$

$df(u, D(q)) > df(v, D(q)) \wedge r(u, v) > 0.5$ であるような, すなわち v よりも一般的でかつ関連度が 0.5 を超えような u を v の上位語と呼び,u は v の上位であると言う. 特徴語と同様に特徴語間の上位下位関係も着目する文書集合 $D(q)$ によって変化する。全ての特徴語対の間の上位下位の関連を計算し, 下位の語から上位の語へパスを引いた有向グラフを概念グラフと定義する.

3 特許データ

本稿では, IPDL 広報テキスト検索データベースから, (容器∪瓶∪壜)∩(リサイクル∪リサイクリング∪再利用 ∪ 再資源) の検索式で検索された 2302 件の特許データを元にシステムを構築した. 特許データ群の発明者および出願人の名前, その識別番号 (ID), FI(File Index), F ターム (File Forming Term) , 特許技術の説明の 5 項目から成るデータベースを作成した.

ユーザは検索用データベースと出力用のデータベースを選択する. 例えば, 発明者の名前で検索し, FI と F タームで出力することで, その発明者がどの分野で技術を発明しているかを調べることが可能である.

4 実験と考察

本節では前節の特許明細書群から特許技術の説明部分のキーワードとそれに対応する FI および F ターム群を自動抽出し概念グラフとして出力する. ただし, グラフ中では枝で結ばれた左側の節が上位, 右側の節が下位の関連語となる.

「ダイオキシン」というクエリに対する, FI, F タームは図 1 のようになった. グラフ中の A62D_0300 は

Related Words Extraction from Patents using Concept Graph

† Yoshihiro SHIMOJI(y-shimo@i.kyushu-u.ac.jp)
† Taiki WADA
†† Sachio HIROKAWA
Graduate School of Information Science and Electrical Engineering, Kyushu University (†)
Computing and Communications Center, Kyushu University (††)

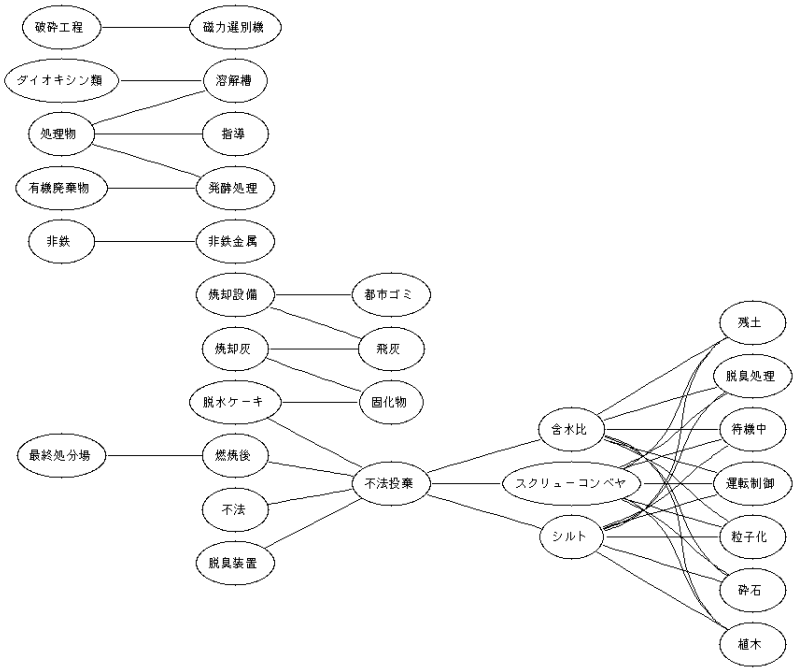

図 2: FI=「B09B」に対する概念グラフ

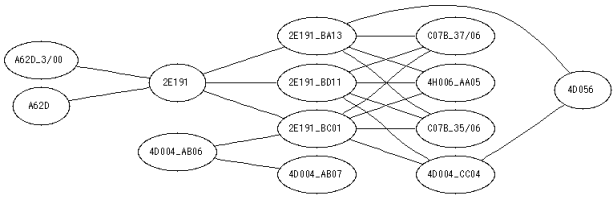

図 1: キーワード=「ダイオキシン」に対する概念グラフ

「物質の科学変化の影響により有害な化学剤を無害にするまたはより有害でなくする方法」に関する FI である. 2E191 はこの FI に対応する F タームのテーマコードである. 4D004 は「固体廃棄物の処理」に, 4D056 は「抽出, 液体の置換」に関する F タームのテーマコードである. この概念グラフからダイオキシンに関係のある特許にどのような分類コードが付与されているかがわかる.

次に, 「B09B」という FI についてその関連語群を特許明細書群から抽出しグラフにしたものが図 2 である. B09B という FI は固体廃棄物の処理についての特許に対応しており, 概念グラフにも廃棄物処理についての単語がこの FI の関連語として出現していることがわかる.

5 結論と今後の課題

本稿では, 特許明細書群について概念グラフを適用することで, ユーザの指定したキーワードに対応する分類コード群を抽出, あるいは分類コードに対応するキーワード群を自動的に抽出するシステムを提案した. このシステムを利用することで特許の非専門家と専門家の特許データベースの検索能力の差を縮小するか, または専門家であっても検索に必要な労力を削減することが可能である. また, 各キーワードに対する関連する分類コードおよび各分類コードに対する関連するキーワードを適切性や再現率の観点で評価することが課題である.

参考文献

[1] 廣川佐千男, 下司義寛, 三輪眞木子, シラバスデータを使った分野ごとの概念マップの生成, 第 68 回情報処理学会全国大会講演論文集 3, pp.9-10, 2006

[2] 広報テキスト検索, http://www7.ipdl.ncipi.go.jp/Tokujitu/tjkta.ipdl?N0000=108

[3] 下司義寛, 廣川佐千男, 学会講演データにおける著者やキーワードの関連分析システム, 人工知能学会 第 63 回 人工知能基本問題研究会, 2006

[4] 下司義寛, 和多太樹, 廣川佐千男, 英和辞典からの知識抽出, 第 68 回情報処理学会全国大会講演論文集 3, pp.19-20, 2006