2E-5

文書属性に基づく可視化手法を用いた特許情報検索システムの試作

岡野 祐一　　平野 敬　　亀代 泰三　　岡田 康裕
三菱電機(株) 情報技術総合研究所

1. はじめに

近年、大量に蓄積された電子文書から所望の文書を効率良く検索したいという要求が高まっている。これに対し、我々はこれまで多種多様な文書を対象に精度良く全文検索を行い、文書から抽出した属性情報を基に検索結果を視覚的にわかり易く表示する可視化手法について検討を行ってきた[1][2]。

一方、企業等では競争力強化のために知的財産の有効活用が重要となっており、各種特許検索・調査システムが利用されている。しかし、文書属性に基づいて様々な視点から特許検索結果を分類し、この情報を基に絞込み検索を行うシステムはあまりない。

そこで、特許調査結果の有効利用を目的として、特許情報に対して全文検索、及び文書可視化手法を適用した特許情報検索システムを試作した。

2. 文書属性に基づく可視化手法の概要

文書内に記述された文字列を解析することにより、文書の内容把握や分類に有効な情報（文書属性）を抽出する。具体的には以下の方法により自動的に属性抽出を行う。

(1) EXCEL や HTML,XML 文書の特定領域から文字列を抽出し、これを属性として抽出する。

(2)言語情報、レイアウト情報を用いて文字列を抽出し属性を抽出する。

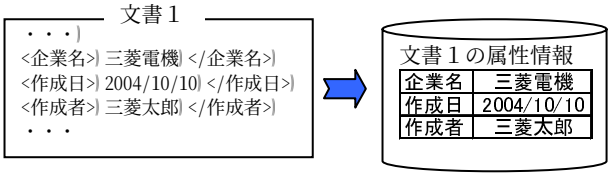

図 1　特定領域からの属性抽出

(1)では図 1 に示すように、例えば特定のタグ付けされた領域をその文書に関する属性情報として登録する。(2)では図 2 に示すように、文書中の文字列に対して形態素解析処理、人名・組織名リスト照合等により属性値となり得る部分文字列（属性値プリミティブ）を抽出する。次に文字列のレイアウト情報や文字種の並びの検定処理等に基づく属性抽出ルールにより属性値を抽出する。例えば、”品詞が人名名詞で、かつ文書上の「報告者」に近くに位置する”という条件で、文字列「小林」を属性”報告者”の属性値として抽出する。

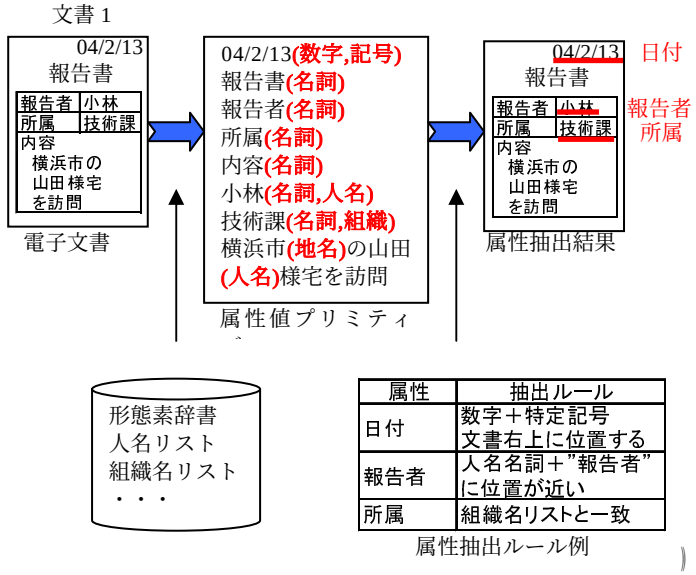

図 2　言語情報、レイアウト情報を用いた属性抽出

このようにして抽出した文書属性情報を用い、

・属性値に応じたランキング表示

・複数属性に関するマトリックス表示

・イメージ上への文書オブジェクトのマッピング

等の可視化を行う。

これにより、文書の傾向を直感的に把握したり、大量の検索結果一覧の中から、所望の文書を容易に探し出すことが可能となる。

A Patent Retrieval System based on Visualizing Information of Document Attributes.
Yuichi Okano,Takashi Hirano,Taizo Kameshiro,Yasuhiro Okada
Information Technology R&D Center, Mitsubishi Electric Corp.
5-1-1 Ofuna, Kamakura, Kanagawa, 247, Japan

3. 試作システム概要

3.1 全体システム構成

全文検索、及び文書属性に基づく可視化手法を使った特許情報検索システムを試作した。今回試作したシステムでは、特許調査の結果をまとめたエクセルファイルの情報を基に特許情報と、対応する属性値を登録するシステムとした。具体的には特許調査の結果から、特許の公報番号、発明の名称、出願人、出願日、概要、特許内容分類情報等をまとめたエクセルファイルを基に、特定セル内の文字列をその特許情報の属性値として登録を行う。システム全体の概要を図 3 に示す。

特許調査結果をまとめたファイルから各特許情報を全文検索するための情報と、各特許文書に対応した属性値を抽出し、サーバに登録する。クライアント側はイントラネットを介して Web ブラウザからサーバに登録した特許情報の検索、及び検索結果一覧の可視化により、所望の特許情報を得るシステムとなっている。

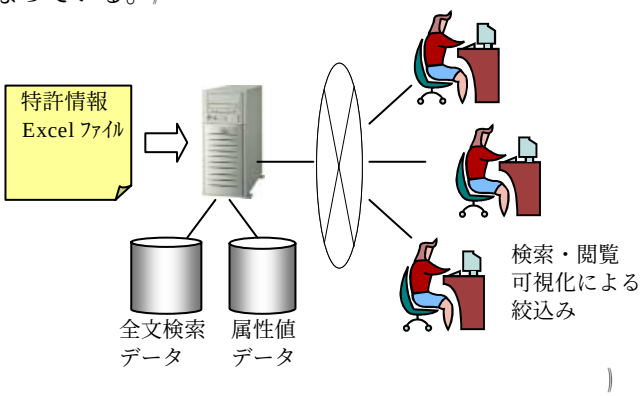

図 3 全体システム構成

3.2 システム画面例

図 4 に試作システムの画面例を示す。キーワードによる検索を行った検索結果一覧画面に加えて、属性値によるランキング表示や表形式による可視化画面を表示する。

属性値による可視化画面では、可視化条件をあらかじめ定義しておき（例えば出願年と出願人の表形式等）、この定義ファイルを指定するだけで、現在得られている検索結果一覧に対する指定条件での可視化が可能となっている。

また、任意の属性を選択してのランキング表示や表形式での表示が可能となっている。

図 4 試作システム画面例

3.3 検索・可視化連携による絞込み

可視化画面では、ランキング表示や表形式での表示に関して、属性値ごとに文書オブジェクトが割り当てられており、必要に応じて属性値単位の絞込みが可能である。また、検索結果一覧画面、可視化画面はすべて連携して動作しており、例えば出願年別ランキング表示の可視化画面で特定年に絞込みを行うと、別の可視化画面（表形式の可視化画面）や検索結果一覧画面においても特定年による絞込み結果が反映される。このように検索結果一覧の情報が大量にある場合には、様々な視点からの属性値による絞込みを行うことにより、容易に目的の文書の検索が可能となる。

4. おわりに

全文検索、文書属性による可視化手法を用いた特許情報検索システムを試作した。今後は大量のデータを使った本システムの有効性の評価を行う予定である。また、今回は特許情報の概要をまとめたデータファイルから文書属性抽出を行っているが、2.で述べた、文書中からの自動属性抽出を用いて特許本文から各種属性を抽出する機能の追加、及び特許本文の内容に応じて文書を分類する機能[3]等の検討を行うことにより、さらに効率の良い検索・可視化絞込みシステムを検討する予定である。

[参考文献]

[1]平野他 “PDL データの解析による…“ PRMU04-145-13
[2]亀代他 “自由書式文書からの…” PRMU03-256
[3]牧田他 “特許検索における分類…” NL2002 No.151-014