情 報 学 基 礎　66－19
デジタル・ドキュメント　32－19
（２００２．３．１５）

図表検索のための図表情報自動抽出の試み

市野 順子[*]，箕牧 数成[*]，山口 和泰[*]，垣 智[*]，東 郁雄[**]，古田 重信[**]

[*]TIS(株) 産業第２事業部 マルチメディアビジネス第１部 先端技術グループ
[**]関西電力(株) 総合技術研究所 情報通信研究室

e-mail : [*]{ichino, mimaki, kyamaguc, skaki}@karl.tis.co.jp, [**]{azuma, sfuruta }@rdd.kepco.co.jp

近年，様々な形式の電子文書が普及し，蓄積されている．しかし，それらの再利用や検索についてはあまり考慮されていない．本研究では，文書中の図表に着目し，様々な形式の電子文書から図表領域を特定し，図表に関連する様々な情報を網羅的に抽出することを目指す．本稿では図表領域及び，図表に関連するテキスト情報を抽出する手法について述べる．提案手法はルールベースを基本としている．11文書 90 図表を対象に抽出を行ったところ，図表領域の特定は，再現率 97%，適合率 80%，キャプション，本文図表説明文の抽出は，それぞれ 3 位適合率 85%，3 位適合率 90%の結果を得た．

キーワード：図表検索，電子文書，図表領域，キャプション，本文図表説明文，XML

Experiment in Automatic Extraction of Chart Information for Chart Retrieval

Junko ICHINO[*], Kazunari MIMAKI[*], Kazuhiro YAMAGUCHI[*], Satoshi KAKI[*],
Ikuo AZUMA[**], Shigenobu FURUTA[**]

[*]Multimedia Business Dept.1, Industrial Business Div.2, TIS Inc.
[**] Technical Research Center, The Kansai Electric Power Co., Inc.
[*]Toyotsu-cho, Suita-shi, Osaka, 564-0051, JAPAN
[**]Nakoji, Amagasaki-shi, Hyogo, 661-0974, JAPAN

e-mail : [*]{ichino, mimaki, kyamaguc, skaki}@karl.tis.co.jp, [**]{azuma, sfuruta }@rdd.kepco.co.jp

Although electronic documents have come into wide use and a mass of data is stored, reuse or retrieval of these documents has not been considered much. In this study, focusing on the charts in documents, we attempt to specify chart areas and extract information about charts from electronic documents in various forms. In this paper, the method of extraction of chart areas and text information is examined. The algorithm is based on rules. An experiment in the validity of chart retrieval was made on 11 documents and 90 charts, and the results show: recall 97% and precision 80% in specifying chart areas; and precision (3-CUTOFF) 85% and 90% in extracting captions and chart explanations respectively.

Keywords : chart retrieval, electronic documents, chart area, caption, chart explanation, XML

1　はじめに

近年のパソコンの急激な普及に伴い，膨大な量の電子文書が蓄積されつつある．これら電子文書の検索は，現在普及している情報検索技術を利用すれば，文書のタイトルやテキスト内容，人手によって付加された文書概要などをキーワードによって検索することができる．しかし，例えば以前見た文書のページ内にあった図表を再利用したいと思った場合は，その図表がありそうな文書をキーワードで絞り込んだ後，各文書のページを一つ一つ読んで探さなければ得ることができず，大きな労力を要する．

もし，文書中の図表や図表に関連する情報を自動的に抽出することができれば，図表検索を実現でき上記の問題は解消できると考えている．

そこで本研究では，種々の形式の電子文書を対象に，図表領域及び図表に関連する情報を抽出し，それらを用いて図表を効率よく検索するシステムの構築を目標とする．

本稿では，その途中経過として，文書から図表領域の特定及び，図表に関連するテキスト情報を抽出する手法を提案する．2 章で，図表検索に有用な情報を分析し本研究でのアプローチを示す．3 章で図表領域を特定する手法，4 章で図表に関連するテキスト情報を抽出する手法をそれぞれ説明し，5 章でそれらの評価実験を行う．最後に，6 章で成果及び今後の課題をまとめる．

2 本研究のアプローチ

2.1 図表検索に有用な情報

まず，文書中の図表を探す場合に手掛かりとなる情報にはどのようなものがあるかを明らかにする．図表を含む文書全体の情報は，図表の上位概念を表す場合が多く有用な手掛かりとなる．また検索対象文書を絞り込む情報としても活用できる．これには，ファイル名・ファイル作成者・ファイル作成日時といったファイル属性情報のほかに，文書の表題・著者・日付や文書量，文書レイアウト情報，文書全体の要約などがある．一方，文書中の個々の図表に関連する情報には，図表の領域（位置，範囲）や，グラフ・写真・テクニカルイラスト・表といった図表の種類のほかに，図表の形状・色・構図といった視覚的な特徴も有用な情報である．また，図表に関連したテキスト情報には，図表に隣接するものとして，図表のタイトルであるキャプションや図表内に含まれる文字情報がある．さらに本文中において図表を直接的・間接的に説明した箇所も有用であると考えられる．以上より，図表検索に有用な情報を表 1 に整理する．

文書から図表領域を特定・抽出する従来の研究は，文書を画像データとして扱ったものが中心であった[1][2][3][4]．これはスキャナで取り込んだ紙文書を対象とした場合に有用である．一方，Word，PowerPoint などで作成された電子文書では，内部的に，四角，線分，テキストといった個々の描画データを保持している．これらのデータをそのまま活用すれば，より正確な図表の情報を把握でき，従来の手法より的確に特定を行うことが可能になると考えられる．

表 1：図表検索に有用な情報

図表を含む 文書全体の情報	ファイル属性情報
	文書の表題,著者,日付
	文書量(ページ数)
	文書のレイアウト(段組 等)
	文書全体の要約
個々の図表に 関連する情報	図表領域(位置,範囲)
	図表種類
	視覚的特徴
	キャプション
	図表内文字情報
	本文図表説明箇所

また，図表に関連する情報の抽出を試みた研究としては，対象図表の位置情報を検索の手掛かりとする研究がある[5][6]．また，図表の視覚的情報を抽出する技術としては，画像検索の分野で広く研究されており，例えば画像認識により色や形状の特徴付けを行うものがある [7]．図表に関連するテキスト情報の抽出を行った研究としては [8][9][10]がある．これらはいずれも図表に関連する情報を部分的に抽出しているにとどまっており，不十分である．

以上より，本研究では，画像データとしての文書ではなく，文書作成ツールを使って作成された電子文書を対象とした場合の図表情報の抽出を行う．また，図表に関連する情報を網羅的に抽出することを目指し，本稿では，表 1 のうち，網掛けで示した情報の抽出を行う．

2.2 対象とする文書

本研究が対象とする文書は，ある特定の文書ではなく，できるだけ広範囲なものにしたい．

対象とする電子文書は，一般に普及している文書作成ツールで作成されたもののうち，文書を構成する個々のデータにアクセス可能なものである．このようなツールには，Word, PowerPoint, Excel, 一太郎, Lotus123, PDF などがあるが，なるべくこれらのフォーマットに依存せずに処理したい．図表情報抽出を目的とした場合，文書を構成する各データの種類，座標，サイズ，テキスト内容などがわかればよい．よって，ファイルから構成要素データを取得する処理はフォーマットごとに必要となるが，構成要素データから図表情報を抽出する処理は，フォーマットに依存しない処理を目指す．

上記であげたフォーマットは，いずれも PDF への変換機能をもつ．しかし，PowerPoint は PDF に変換すると余分な図表情報をもつため，PDF に変換されると図

表領域特定が困難であることがわかった．Word，Excel については，PDF に変換せずに直接データを取得可能だが，複数の座標体系が存在し図表情報を抽出するのは困難と思われる．以上より，今回は，現在対応可能なものとして PDF，PowerPoint 及びそれらに変換可能なフォーマットに絞り込んだ．

一方，文書をファイル形式ではなく，図表という観点から文書の書式を見た場合，学術論文・マニュアルなど文字情報が圧倒的に多く図表はその補助的な役割を担っているもの（以降，「一般文書」と呼ぶ）と，プレゼンテーション資料やカタログなど図表がその中心的役割を担っているもの（以降，「プレゼンテーション文書」と呼ぶ）の二つに分けることができる．本研究ではいずれの文書書式も対象とする．

以上より，本研究の対象範囲は，ファイル形式という観点からは PDF，PowerPoint 及びそれらに変換可能なもの，文書書式という観点からは一般文書及びプレゼンテーション文書とした．

2.3 図表検索システムの概要

本研究の図表検索システムの概念図を図 1 に示す．

まず, PDF 及び PowerPoint ファイルから, API や SDK を利用して文書の構成要素データを取得し，共通データ形式に変換する．これから，最初に，図表領域及びテキスト情報の抽出を行う．次に，一般文書またはプレゼンテーション文書に分類後，図表検索に有用な情報をそれぞれ抽出する．抽出された情報は図表情報ファイルとして書き出され，図表検索の際に検索対象となる．尚，図表情報ファイルは XML 形式で記述する．

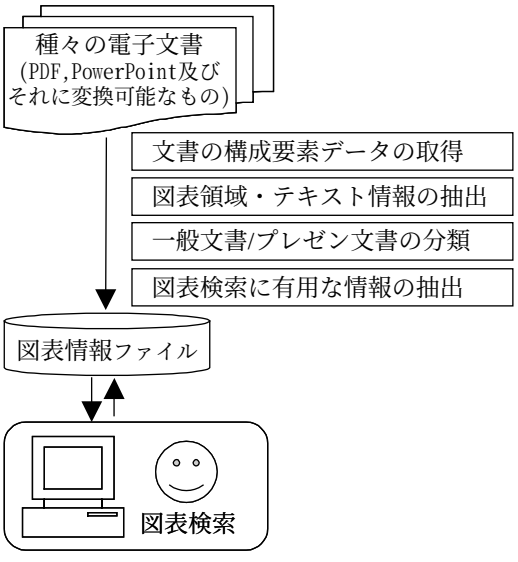

図 1：図表検索システムの概念図

3 図表領域の抽出

図表領域とは，ここでは一つの図表とみなすことのできる範囲（座標やサイズ）を指す．

電子文書における描画データは，直接図表領域を表す情報をもっておらず，実際には図表を構成する最小単位である基本図形の情報しかもっていない．各基本図形は，種類（四角，線分，矢印，テキスト，イメージなど），座標，その基本図形の矩形領域のサイズといった情報をもっている．この基本図形が複数集まったものを一つの図表として我々が見ているだけであり，描画データの中には，図表の範囲を直接示すデータは存在しない．このため，独立して存在する基本図形の集合から，何らかの方法で一つのまとまりをもった図表として識別し直す必要がある．

そこで，我々が普段図表を作成するプロセスを考察する．図表を描く場合，基本図形を順に作成しながら，それらを重ねて配置したり，接して配置したりする．また，接していないが近くに配置することもある．このことより，以下の手法で一つの図表を特定した．

表 2：図表の特定

1. 基本図形の矩形領域に対して，その領域面積の 0〜15%程度のマージン幅を周囲に付加する．
2. マージン幅が付加された基本図形の矩形領域同士が重なる場合，それらを 1 つのグループにまとめる．
3. 2 でできたグループの矩形領域同士が重なる場合，それらをさらに大きな 1 つのグループにまとめる．
4. グループの矩形領域同士が一つも重ならなくなるまで 3 を繰り返し，最後にまとめられたグループを図表とする．

この手法によって，複数の基本図形が一つの図表と特定される例を図 2 に示す．実線が基本図形，グレーで囲まれたものがマージン幅を付加された矩形領域，破線がグループの矩形領域を表す．

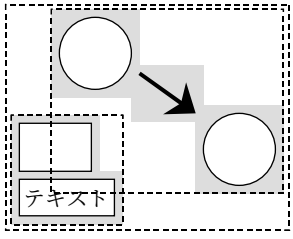

図 2：5 つの基本図形が 1 つの図表に特定される例

ところが，上記手法によって特定されたグループの中には，単なるテキスト，囲み線や網かけといった文字飾り，ページ全体を囲む枠線や章の区切りとしての境界線，といった明らかに図表ではないものも含まれる．それらは，以下のようなルールを設定することで図表の対象から外した．

- グループにテキストの基本図形しか存在しない（単なるテキスト）
- グループにテキストとそれ以外の基本図形が存在し，その両方の矩形領域がほぼ同じ座標，同じサイズである（文字飾り）
- グループには基本図形が一つしかなく，それがイメージ以外である（枠線，境界線など）

以上より，最終的に図表と特定されたグループの矩形領域が図表領域となる．図表領域特定後，データは図表領域とそれ以外のテキスト情報に分けられる．

4　図表テキスト情報の抽出

ここでは，図表に関連するテキスト情報の抽出を行う．抽出する情報は，キャプション，図表内に含まれる文字情報，本文中の図表説明箇所（以降，「図表説明文」と呼ぶ），図表種類の 4 種類である．以下に，キャプション及び図表説明文の抽出方法について述べる．

4.1　キャプションの抽出

一般的にキャプションは図表に隣接した「図１ ○○の構成」といった文字列を指すが，本研究ではキャプションを広義にとらえ「図表に隣接し，簡潔に図表内容を表現しているテキスト」と定義する．まず，キャプションがどのような特徴をもっているかを分析し，それを踏まえ抽出方法の検討を行う．

4.1.1　キャプションの表層的特徴

まず，キャプションの特徴を把握するために，著者の異なる複数の文書を調査したところ，一般文書中の図表に記述されるキャプションは，「第 1 図」，「表 2」といった図表番号を含むもの（以降，「定型キャプション」と呼ぶ）と，含まないもの（以降，「非定型キャプション」と呼ぶ）に大きく分類でき，いずれも図表の上下に位置する傾向が強いことがわかった．一方，プレゼンテーション文書の場合は，図表番号は存在せず，フォントサイズや位置にその特徴があった．

ここで，キャプションを，一般文書とプレゼンテーション文書という観点ではなく，文字情報とレイアウト情報という観点から見ると，以下の共通した特徴をもつことがわかった．

(1) 文字情報に関する特徴

図表を明示的に指し示す語を含む場合がこれに当たるが，以下の 2 つに分類できる．

- 図表番号が行頭に出現するテキスト
- 「～する図」,「～の表」,「～の例」といった語が行末に出現するテキスト

(2) レイアウト情報に関する特徴

位置や文字サイズには以下の特徴がある．

- 図表の上端もしくは下端に位置する場合が多い．
- 図表の周囲に複数のテキスト情報がある場合，フォントサイズが他より大きいものが該当する場合が多い．

4.1.2　キャプション抽出ルール

前節で述べた，キャプションの文字情報とレイアウト情報を利用してキャプションの抽出を行う．抽出の手順は，文字情報に関するルールと，レイアウト情報に関するルールそれぞれをベースとした抽出を行い，双方の結果を総合的に判断しキャプションを決定する．抽出の対象とするテキストは，図表領域及びその周辺の本文テキストである．

文字情報による抽出は，キャプションの定型的な表現を正規表現で表し，パターンマッチングによって候補を特定する．表 3 に正規表現の一例を示す．

レイアウト情報による抽出は，前節でまとめた特徴をルールとし，ルールベースによる抽出を行う．

表 3：パターンマッチングによるキャプションの抽出

正規表現	マッチング
第?.?[0-9|０-９]+¥s?(図|表){1}	行頭
(図|表|写真){1}.*?[0-9|０-９]+	行頭
(図|表|例|写真){1}	行頭・行末
(Fig| Figure| Tab| Table){1}*?[0-9]+	行頭

4.2　図表説明文の抽出

本研究では，本文中で図表を直接的に説明している部分や，図表に深く関連する内容が述べられている部分を図表説明文と定義する．まず，図表説明文がどのような特徴をもっているかを分析し，それを踏まえて抽出アルゴリズムの検討を行う．

4.2.1　図表説明文の表層的特徴

一般文書における図表説明文の表層的特徴を把握するために，10 文書 65 図表を調査したところ次のようなことがわかった．

まず，図表の物理的な分布傾向を考察する．図表説

明文は図表に近接しているとは限らないが，図表の前後 1 ページの中に位置することが多く，同一ページであることが最も多かった．その出現位置は，図表より先または後といった傾向は見られなかった．図表説明文としての文書量は，図表を直接的に指し示している文に前後数文を含めた量になることが多く，図表毎にかなりばらつきはあるが，平均すると 200 文字程度であった．次に，文書の論理構造的な観点から考察すると，図表説明文と図表は同じ論理単位（章，節等）の中で出現することがわかった．また，図表説明文中に出現する単語と，図表領域内に出現する単語を比較したところ，両者で同一の単語が使われる場合が多いことがわかった．これは，文書中で図表を説明する場合，図表のタイトルとも言えるキャプション中の図表番号や単語を用いて明示的に引用することが多いためであると思われる．

一方，プレゼンテーション文書の場合，ページ間の連続性が一般文書に比べて低く，ある図表に対して別のページで説明を行うようなケースは希であった．また，その利用目的から図表そのものが文書の中心となり，一般文書と比較して文字情報は極めて少ない．図表の説明をプレゼンテーションの発表者が口頭で行う割合が高いため，文字として記述される情報は発表のポイントや補足のみで，文章形式でないものが多かった．しかし，それらの文字情報は，図表と同一ページ内にある場合，ほとんどが図表と密接に関連した内容であることがわかった．また，図表とそれら文字情報との位置関係に，関係の強さといった特定の傾向を見ることも難しいことがわかった．

以上より，図表説明文の表層的特徴を表 4 にまとめることができる．

表 4：図表説明文の特徴

●一般文書
①図表の前後 1 ページ中に出現する場合が多い
②図表を明示的に参照している文の前後数文からなる場合が多く，その長さは 200 文字程度である
③図表と同じ章または節の中に出現する
④図表領域内の図表番号や単語が頻繁に使われる
●プレゼンテーション文書
⑤図表の説明が別のページで行われることは少ない
⑥ページ内のほとんどの文字情報が図表に密接に関連した内容である
⑦図表との位置的な関連性は低い

4.2.2 図表説明文を抽出するアルゴリズム

前節で行った分析をもとに，ここでは図表説明文を抽出するアルゴリズムについて考察する．

まず，一般文書における抽出について述べる．表 4 ④より図表領域中の図表番号や単語が多く出現する部分がその図表の図表説明文である可能性が高いという特徴にもとづき，テキストマッチング技術を用いて図表説明文抽出を行う．キーワードと本文中におけるその偏出度を用いて説明文を特定する研究として[11]があるが，本研究は図表番号の存在する文を重視し他よりも高い重みを与えた．前節の分析を踏まえ，以下のように図表説明文を抽出する．

あるまとまった文章を図表説明文とするため，本文テキストを章や節といった論理単位(表 4③)からそのタイトル部分を除いた文章部分を抽出の対象範囲とする．3 章で抽出されたテキスト情報から取得できる行単位のテキスト情報を文単位に区切り直す．このとき，フォントサイズの大きいテキスト情報を章や節のタイトルと想定し，小さいテキスト情報をヘッダー，フッターと想定し対象から除く．次に，図表説明文の抽出にキャプション情報を用いる(表 4④)．キャプションからパターンマッチングにより図表番号を抽出後，図表番号以外の文字列を日本語形態素解析システム「茶筌」[12]を用いて形態素解析し「名詞」及び「未知語」を抽出する．先に抽出した本文文章部分の各文に対し，図表と同一ページに存在する場合(表 4①)や，キャプションの図表番号や単語がマッチする場合に重み付けした得点を与え，図表説明文の候補を抽出する．一定以上の得点をもつ候補文を中心に，隣り合う前後の文を交互に 200 文字程度になるまで連結し（表 4②），図表説明文を生成する．

一般文書における図表説明文の抽出アルゴリズムを以下にまとめる．

(1) 本文文章部分を抽出する
(2) キャプションから図表番号,単語を抽出する
(3) 本文文章部分の各文に対して(2)の抽出文字列の出現状態に対応した重み付けを行う
(4) (3)で抽出した候補文に前後数文を連結し図表説明文を生成する

次に，プレゼンテーション文書における図表説明文について述べる．表 4 ⑤より，同一ページのテキスト情報のみを対象とする．また⑥,⑦より，テキスト同士のつながりやテキストの重要度の差を判別しにくいため，各テキスト情報に対して一般文書での抽出で行ったような重み付け行うことは難しい．以上より，プレゼンテーション文書における図表説明文を，図表と同一ページの，図表領域以外のすべてのテキストとする．

4.1 節及び 4.2 節ではキャプション及び図表説明文の

抽出方法について述べた．その他の図表テキスト情報として，図表内に含まれる文字情報は，3 章で抽出された図表領域に含まれるテキスト情報をすべて抽出し，それを図表内文字情報とする．また，図表種類は，文書中のテキスト情報から特定可能なレベルを考え，図／表／写真／イメージ／その他 に分類する．種類の特定は，4.1 節で抽出したキャプションを利用しパターンマッチングにより行う．

5 評価実験

提案手法の有効性を確認するために，図表領域及び図表テキスト情報を抽出するプロトタイプシステムを開発し，11 文書 90 図表を対象にし，一般文書とプレゼンテーション文書に分けて評価実験を行った．

5.1 図表領域の抽出結果

各サンプルについて 3 章で述べた図表の定義にもとづき，人手で正解図表を設定した．抽出結果に対して，正解図表と抽出図表が完全に一致するもの，正解図表の一部分が抽出されたもの，正解図表を包含して抽出されたもののうち，図表として意味があるかどうかを人手で判断し，意味のあるものを正解とした．対応付けの結果を，正解図表に対する再現率，適合率を用いて評価した．

表 5：図表領域抽出結果

文書の種類	再現率	適合率
一般文書	96%	76%
プレゼンテーション文書	100%	100%
計	97%	80%

再現率 ：Rf = |Ef| / |Tw|
適合率 ：Pf = |Ef| / |Tf|
Tw：全正解図表数
Tf：抽出図表数
Ef：抽出図表のうち正解図表数

結果より，図表抽出のためのルールはほぼ有効に機能していることがわかる．しかし，一般文書の適合率は 76%と低い．これは，一つの正解図表が複数の意味の無い図表に分割して抽出されたり，本文中の線やデザインとして挿入された矩形領域などの意味の無い図表が抽出されたりしたことが原因と考えられる．これらに対応可能なルールを再検討する必要がある．

5.2 キャプションの抽出結果

非定型キャプションやプレゼンテーション文書におけるキャプションは，正解キャプションを唯一に特定しにくいため，前節の図表抽出で用いた再現率を出すことは難しい．よってキャプションの抽出結果は 1,3 位適合率として評価する．抽出結果に対して，4 章で述べたキャプションの定義にもとづいたテキストを正解とし，単体では図表を特定できないテキストや，文章形式のものは不正解とした．一般文書は，定型キャプションをもつ図表からなる文書と，非定型キャプションをもつ図表からなる文書に分けて評価した．

表 6：キャプション抽出結果

文書の種類		1 位 適合率	3 位 適合率
一 般 文書	定型キャプション	97%	97%
	非定型キャプション	80%	80%
プレゼンテーション文書		77%	77%
計		85%	85%

適合率：Pc = |Ec| / |Tc|
Tc：Ef のうちキャプションを抽出できた図表数
Ec：Tc のうち抽出スコアの上位 1,3 位で正解キャプションを抽出できた図表数

結果より，文字情報とレイアウト情報を利用した抽出が概ね有効であることがわかる．また，1 位適合率と 3 位適合率が同じであることから，抽出のための重み付けが有効に機能し上位での抽出に成功していることがわかる．

しかし，抽出に失敗した例として，段組設定された文書において右段の図表に対して左段のテキストが抽出されており，これは抽出対象領域の特定が不十分であることが考えられる．このため，段組情報や正確な論理構造の情報を把握する必要がある．また，キャプションが全く抽出されないものに関しては，パターンマッチングでは候補を抽出できず，レイアウト情報のみを利用したが抽出に失敗していた．表 3 で示したパターンの追加や，レイアウト情報による抽出ルールを再検討する必要がある．

5.3 図表説明文の抽出結果

前節のキャプションと同様，図表説明文も正解を唯一に特定できないため，抽出結果を 1,3 位適合率として評価する．抽出結果に対して，4 章で述べた図表説明文の定義にもとづき，図表を直接的に説明している部分を含むものや，図表に深く関連する内容が述べられているものを正解とし，図表の理解につながらないものや他の図表を説明しているものは不正解とした．なお，プレゼンテーション文書については，4 章より容易に抽出可能なため評価対象から外した．

表 7：図表説明文抽出結果

文書の種類		1 位 適合率	3 位 適合率
一 般 文書	定型キャプション	86%	100%
	非定型キャプション	70%	75%
プレゼンテーション文書			
計		80%	90%

適合率：Pe = |Ee| / |Te|

Te：Ec のうち図表説明文を抽出できた図表数

Ee：Te のうち抽出スコアの上位 1,3 位で正解図表説明文を抽出できた図表数

結果より，テキストマッチングによる図表説明文の抽出が概ね有効に機能していると言える．特に図表番号をもつ図表からなる文書である定型キャプションの場合の抽出結果が良いことから，図表番号に着目したマッチングが有効であることを確認できた．

一方，正確に抽出されなかった例として，図表説明文の中にキャプション自体が含まれる等，不完全な文章のものがあった．本文文章部分の抽出段階でより正確な論理構造を抽出する必要がある．正確に抽出できない原因の一つに，PDF 文書から抽出したテキスト情報の出現順序が表示順序と一致していないことがあげられる．また，キャプション中のどの単語も本文に出現しない場合や，キャプションの文字列長が短い場合に抽出結果が悪かった．このため，キャプション情報だけでなく，図表内の文字情報やシソーラスの利用も検討する必要がある．

以上，図表領域，キャプション，図表説明文の具体的な抽出結果例を図 2 に示す．これは，定型キャプションをもつ図表を含む一般文書「関西電力 R&D News Kansai 2001.11」に対する抽出結果である．この例では，ページ右側の正解図表に対して，3 つの破線で囲まれた部分，上からそれぞれ図表説明文，図表領域，キャプションを抽出できた．

さらに，上記 5.1 節～5.3 節の評価実験とは別に，全文検索との比較をするために簡単な実験を行った．28 文書に対して，全文検索ツールを用いてあるキーワードで検索したところ 18 文書がヒットした．一方，開発システムを用いて同じキーワードで検索したところ，1 つの図表が抽出され，実際にそのキーワードに関連する図表は 28 文書中その図表 1 つだった．この結果からも，開発システムが有効に機能していることを確認できた．

図 2：図表情報の抽出結果例

6 おわりに

本稿では，各種電子文書の図表に焦点を当て，図表検索システムに必要な図表領域の抽出及び図表テキスト情報の抽出を行った．実験より，ファイル形式に依存しない汎用的なルールを用いた図表領域の特定，文字情報とレイアウト情報を利用したキャプション抽出，図表領域中の図表番号や単語を用いたテキストマッチングによる図表説明文の抽出，それぞれの有効性を確認できた．

特に各種ファイル形式からの変換が容易な PDF 文書への対応を実現できたことにより，本研究で取り組む図表検索システムの汎用性が高まった．これらの成果を利用すると，単に図表を検索するためのツールにとどまらず，図表内容に対する情報を利用したシステムとして，例えば論文検索システム，テキストマイニングツール，電子図書館との融合といった可能性も検討できる．

今後の課題として次のことがあげられる．

- 正確な論理構造の抽出
  キャプション及び図表説明文の抽出実験の考察から，正確な論理構造の抽出により各処理の抽出精

度が向上すると思われる．

- 視覚的特徴の抽出手法の検討
  個々の描画データの種類情報を利用し，視覚的特徴を抽出することで，図表に関連する情報を網羅的に抽出でき，図表検索の実用性が高まる．
- 図表検索システムに有効なユーザインタフェースの設計・実現
  抽出した種々の図表情報を，検索及び検索結果表示において有効に利用し，効率良くユーザに提示するための工夫が必要となる．

参考文献

[1] Yanping Zhou, Chew Lim Tan："Chart analysis and recognition in document images", Proc. Sixth International Conference on Document Analysis and Recognition, pp.1055-1058, 2001

[2] Saitoh, Yamaai et al.：”Document Image Segmentation and Layout Analysis (Special Issue on Document Analysis and Recognition)”, IEICE transactions on information and systems, Vol.E77-D, No.7, pp.778-784, 1994

[3] 平山：「複雑なカラム構造をもつ文書イメージの領域分割法」, 電子情報通信学会論文誌, Vol. J79-D-2, No.11, pp.1790-1799 (1996.11)

[4] 岩崎, 黄：「文書中の図領域検索方式の提案」, 情報処理学会全国大会講演論文集, Vol. 第 55 回平成 9 年後期, No. 3, pp.196-197, 1997

[5] 高橋, 島他：「位置情報を手がかりとする画像検索法」, 情報処理学会論文誌, Vol.31, No.11, pp.1636-1643, 1990

[6] Chang, S.K., Yan, C.W., Dimitroff, D.C., Arndt, T.："An intelligent image database system", IEEE Transactions on Software Engineering, Vol.14, No.5, pp.681-688, 1988

[7] 串間, 赤間他：「オブジェクトに基づく高速画像検索システム : ExSight」, 情報処理学会論文誌, Vol.40, No. 2, pp.732-741, 1999

[8] 岩崎, 黄：「文書中の図領域検索方式の提案」, 情報処理学会全国大会講演論文集, Vol.第 55 回平成 9 年後期, No.3, pp.196-197, 1997

[9] 小平, 久保田：「図表や写真に含まれる文字列の抽出方法」, 電子情報通信学会ソサイエティ大会講演論文集, Vol.1998 年.情報・システム, pp.243, 1998

[10] Google イメージ検索
URL：http://www.google.com/imghp?hl=ja

[11] 水野, 黄瀬他：「単語の出現密度分布と偏出度を用いた図表と説明テキストの対応付け」, 情報処理学会論文誌, Vol.40, No.12, pp.4400-4403, 1999

[12] 日本語形態素解析システム「茶筌」
URL: http://chasen.aist-nara.ac.jp/index.html.ja