文書中の図領域検索方式の提案

3Q－6

岩崎 雅二郎　　黄 英傑
（株）リコー 情報通信研究所

1 はじめに

昨今、テキストを検索する手法として高速な全文検索が様々なアプリケーションで利用されるようになってきた。こういったアプリケーションは文書やページといった単位での検索は可能だが、文書にはテキストだけではなく様々な図、表、写真といったオブジェクトが存在する。そこで文書を構成するオブジェクト単位に検索したいという要求が新たに現れ始めた。例えば、「○○のシステム構成図」や「日本の人口推移のグラフ」が欲しいといったような要求がその一例である。

図や写真を検索する手段として一つには図や写真から画像特徴を抽出してそれを基に検索する手段が考えられる。しかし、現在の技術では図や写真から十分な画像特徴が得られず、利用者の検索要求に十分に応えることができない。

しかし、文書中の図、表、写真の内容は、通常キャプションや本文で記述されており、逆に説明がないような図は挿絵といった検索対象とはなりにくいものであると言える。したがって、図の内容を説明するキャプションや本文のテキストを利用することにより図、表、写真を効率よく検索することが可能である。

そこで本稿では文書中の図、表、写真といったオブジェクトについて効率よく検索する方式を提案する。

2 概要

本研究では実際に文書中の図を検索するシステムを構築した。検索対象となる文書を登録する処理を図1に示す。本システムの入力は紙文書をスキャナで取り込んだ画像データであるが、WP 文書に対して本方式を適応することも可能である。まず、文書中の図 (写真、表を含む)、キャプション、本文の領域に分割 [1] し、図は画像 DB に登録する。キャプションと本文のテキスト領域は OCR 処理を行いテキストデータを生成する。その後、図の内容を示すテキスト（キーテキスト）を抽出し、抽出したテキストを図とリンク付けして全文検索 DB に登録する。

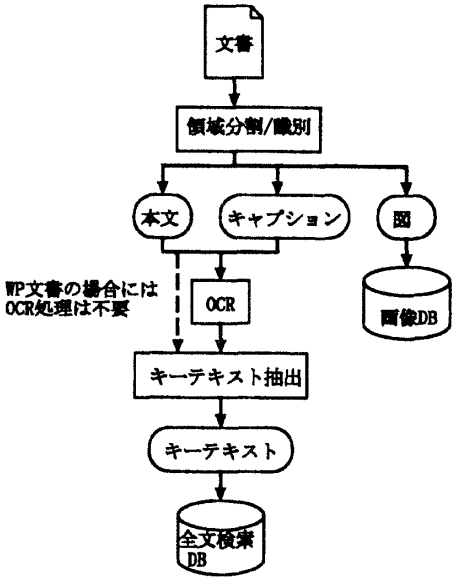

図 1: 登録処理の流れ

検索時には、ユーザが入力した検索語により全文検索 DB を検索し、ヒットしたキーテキストにリンク付けされている図をユーザに提示する。こうすることでユーザが思いついた検索語で文書中の図を検索することが可能となる。

3 キーテキスト抽出

人手により図のキーワードを抽出するテストを行ったところ、1) 図のキャプション、2) 図を直接参照している文（キーセンテンス）、3) 図を直接参照している文を含む段落（キーパラグラフ）の順で抽出したキーワードの出現頻度が高かった。したがって、ユーザが入力した検索語が検索時にどの個所でヒットしたかによって、入力した検索語とヒットした画像の関連度が判断できる。本研究では、さらに漏れをなくすためにキーパラグラフを含むページ（キーページ）を加え、図2のように4種類のテキスト（キーテキスト）を抽出す

A Method for Retrieving Graphic Areas in Documents
Masajirou IWASAKI, Ying-Jieh HUANG
Information and Communication R&D Center
RICOH Co.,Ltd.

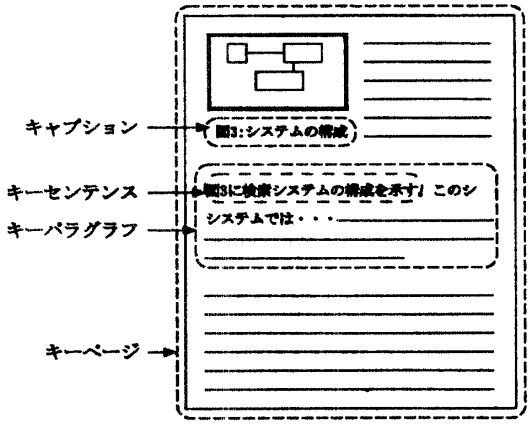

図 2: 文書中のキーテキスト

る。各キーテキストを抽出する方法を以下に示す。

キャプション：図に付与されているキャプション
領域分割/識別処理により図に近接する数行からなるテキスト領域をキャプションと判断する。

キーセンテンス：図を直接参照している一文
キャプションがあり、かつ、キャプションに図番がある場合には本文のテキストをサーチし図番による参照がある文を探し出し、それをキーセンテンスとする。図番やキャプションがない場合でも、本文中に図を指し示す言葉である、「上図」や「右の図」といった指示語があれば、その指示方向に存在する図を探し、探し出した図のキーセンテンスとする。ただし、図番付きのキャプションも指示語も存在しない場合には、キーセンテンスは抽出できない。

キーパラグラフ：キーセンテンスを含む一段落
キーセンテンスがある場合にはキーセンテンスを包含する段落をキーパラグラフとする。図番付きのキャプションが無い場合には説明文との対応を取るために、通常、図のそばで説明されている。そこで、キーセンテンスがない場合には、図に近接するテキスト領域をキーパラグラフとする。なお、包含するキーセンテンスのテキストは除く。

キーページ：キーパラグラフを含む一ページ
キーパラグラフを包含するページをキーページとする。なお、包含する上記キーテキストは除く。

これらの４種類のキーテキストを図３で示すように図とリンク付けしデータベースに登録する。

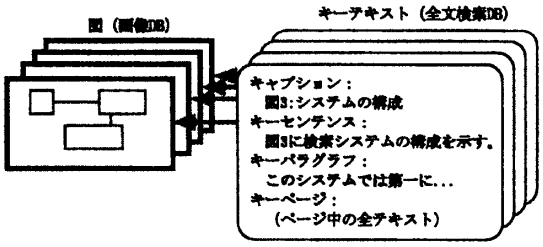

図 3: 図とキーテキストのリンク

4　検索結果のランキング

ユーザが検索語を入力すると、システムはその検索語をキーテキストから全文検索により検索する。そしてヒットしたキーテキストにリンク付けされた画像を検索結果としてユーザに提示する。画像を提示する時にヒットした場所がキャプションからキーページの順にユーザへ提示することで関連度の高い順にランキング付けし表示することができる。

70 文書（報告書、論文、雑誌、新聞記事など）を登録し、実際に人手によって図のキーワードを抽出した場合と比較し、適合率及び再現率 (表 1) を求めた。表からわかるように適合率と再現率はトレードオフの関係にあり、ユーザが網羅性を重視するか、または、正確さを重視するかを自由に選択が可能である。

表 1: 適合率及び再現率

ランク	キーテキスト	適合率 (%)	再現率 (%)
1	C	93.2	12.7
2	C+S	90.2	14.6
3	C+S+Pr	70.9	30.7
4	C+S+Pr+Pg	50.3	73.1

C:キャプション、S:キーセンテンス、Pr:キーパラグラフ、Pg:キーページ

5　おわりに

文書中のテキストを利用して図を検索する方式を提案した。本方式では図との関連度の異なるキーテキストを抽出することにより、検索時に４段階の関連度でユーザに検索結果をランキングして提示することができ、効率よく検索することを可能とした。

参考文献

[1] Saitoh,T., Yamaai,T.and Tachikawa,M., **Document Image Segmentation and Layout Analysis**, IEICE Trans. Inf. & Syst, Vol. E77-D, No.7, July, 1994