既存文書のレイアウト情報付き構造化手法

3S－5

石田 和生　市山俊治
NEC 関西 C&C 研究所

1 はじめに

近年、注目をあびているものに電子図書館システムがあるが、この電子図書館システムを実際に運用するためには、現在紙ベースで存在している既存文書を電子化して入力する必要がある。既存文書を電子化する方法として、文書をスキャナと OCR を用いてテキストデータに変換し、変換されたテキストデータから文書の持つ論理構造 (タイトルや、段落である、といった情報) を抽出して論理構造情報を含んだデータ (構造化テキスト) として格納するやり方がある [1,2]。しかし、文書には論理構造だけでなくレイアウトに関する情報も含まれているが、レイアウトの持つ意味は非常に大きいため、これを無視するわけにはいかない。本稿では、既存文書を電子化するときに、従来の論理構造情報記述手段の枠組の中で、文書のレイアウト情報も保存出来る情報構造化方式を提案し、試作システムについて述べる。

2 レイアウト情報付き構造化テキスト

2.1 レイアウト情報の重要性

構造化テキストのフォーマットとして SGML が標準となりつつある。SGML は通常、論理構造だけを記述してレイアウトに関する情報は持っていない。このため、既存文書を SGML 化したデータを表示アプリケーションで表示する場合、もとの文書のレイアウトの再現は保証されない。ところが文書の中には「右図参照」等のようにレイアウトが変わってしまうと正しく読み手に意味が通じなくなる記述が存在することがある。また、レイアウト情報は人間の記憶の中でも比較的大きな割合をしめており、以前見た文書を、「右上にグラフのあった論文」のようにレイアウトをもとにして覚えていることも多い。このような覚え方をしている場合、レイアウト情報の抜けた構造化テキストデータの中から目的の論文を検索することは非常に困難である。以上のようなことから、既存の文書を構造化テキストに変換する場合には、文書のレイアウト情報の持つ意味がたいへん重要であると言うことが出来る。

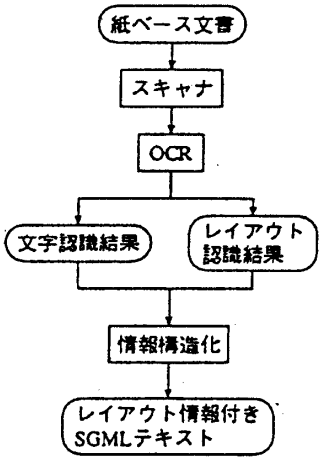

図 1: 処理の流れ

2.2 レイアウト情報つき SGML

既存文書を電子化する場合に、レイアウト情報を保存しておく方法には、1) イメージデータとして保存、2) 文書のおおまかなレイアウト (「右上が図で、左下が文章」といった程度のもの) をインデックスとして文書に付加する、といったものが考えられる。しかし、1 の方法だとデータサイズが大きくなり、さらに、検索を行うために別途インデックスを作成しなければならないという問題がある。2 の方法だと、データサイズをおさえながらレイアウト情報をもとにした検索がある程度実現できるが、特殊なインデックスを付加したデータになるため、従来の検索システムやデータ編集システムでは扱えないという問題点がある。

そこで既存文書を SGML フォーマットの構

Generating Structured Text with Layout from Printed Document
Kazuo ISHIDA, Shunji ICHIYAMA
Kansai C&C Research Laboratories, NEC Corp.
1-4-24 Shiromi, Chuo-Ku, Osaka 540, JAPAN

造化テキストに変換する際、レイアウト情報をSGML のタグとして保存しておく方式を考察した。この方式の処理の流れを図1 に示す。まず、スキャナで取り込んだページ画像を OCR によってテキストに変換するとき、文字コードと一緒にその文字がページ上で存在している位置情報も出力する。次に、出力されたテキストデータと位置情報をもとに文書の論理構造を抽出し、SGML 形式の構造化文書に変換する。このとき同時に、文字の位置情報を SGML のタグ (今回、新たに定義した <Layout> タグ) の属性に埋め込む。埋め込む情報は、1 文字毎の位置情報のままでもよいが、文書のおおまかなページレイアウト情報を保存しつつ、出来るだけデータサイズを押さえるために、今回は段落単位での位置情報を保存することにした。また、レイアウト情報の保存方法として、<Layout> タグを用いずに <Para> などの論理構造を表すタグの属性値に埋め込む方法も考えられるが、この方法だと例えば、ひとつの段落が複数ページにまたがる場合の対応などが困難であるため、今回は <Layout> タグによる方法を用いた。

以上のような処理を行うシステムを実際に試作し、図 2 の既存文書を SGML に変換した例を図 3 に示す。この図に含まれている <Layout> タグの属性にレイアウト情報が埋め込まれている。例えば、セクションタイトル「はじめに」は「2 ページ目の紙の左端から 0.3、上から 0.1 の場所に、横幅 0.2、高さ 0.05 の大きさ」で書かれていた、ことを表している。

このように SGML のタグとしてレイアウト情報を保存することによって

1. 表示するときに、もとの文書のレイアウトを再現することが出来る
2. レイアウト情報をもとにした検索が、容易に行える
3. <Layout> タグを無視すれば、論理構造のみを含んだ通常の SGML テキストとして扱える

といった利点がある。2 番目の「レイアウト情報をもとにした検索」というのは「右上にグラフのあった論文」のような検索のことで、<Layout> タグの属性 LEFT, TOP, WIDTH, HEIGHT の値を利用して実現することが出来る。

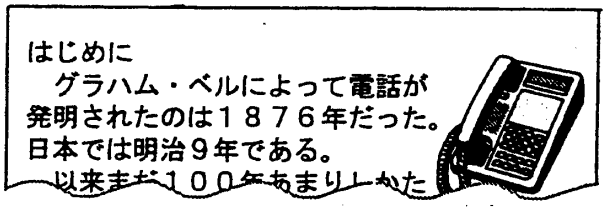

図 2: 紙ベース文書の例

```
<Section>
 <SecTitle>
  <Layout LEFT=0.02 TOP=0.1 WIDTH=0.2
          HEIGHT=0.07 PAGE=2>
    はじめに
  </Layout>
 </SecTitle>

 <Figure>
  <Layout LEFT=0.8 TOP=0.15 WIDTH=0.18
          HEIGHT=0.4 PAGE=2>
  </Layout>
 </Figure>

 <Para>
  <Layout LEFT=0.02 TOP=0.2 WIDTH=0.7
          HEIGHT=0.3 PAGE=2>
    グラハム・ベルによって電話が発明されたのは
    １８７６年だった。日本では明治９年である。
  </Layout>
 </Para>
```

図 3: レイアウト情報つき SGML データの例

3 おわりに

本稿では、電子図書館等のコンテンツ情報作成を目的として、SGML テキストにレイアウト情報を埋め込む手法の提案と、紙ベースの既存文書からレイアウト情報付き構造化テキストを生成するシステムについて述べた。現在、抽出しているレイアウト情報は文字列の存在している座標値だけであるが、文字のサイズやフォント情報等、保存するレイアウト情報の種類について、今後検討を行う予定である。また、本システムが生成するレイアウト情報付き構造化テキストは、検索だけでなく、例えば文書のダイジェスト表示のように、表示に関してもさまざまな利用用途が考えられるので、これらのデータ利用アプリケーションに関しても検討を行っていく。

参考文献

[1] M. Yamaoka, M. Sato, K. Iwane and O. Iwaki, A Document Understanding System for Converting Printed Documents to SGML Instances, Proc. ISDL '95, pp. 287–288, 1995.

[2] 成田 他: 科学技術論文プレーンテキストへの SGML タグ付けの自動化, 自然言語処理の応用に関するシンポジウム, pp. 49–56, 1995.