文字認識誤りを含むテキストからの全文検索方式の一検討

1V－3

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

1. はじめに

OCRを用いて印刷文書をコード化し、無修正で文字認識結果を蓄積する場合、文字認識性能が十分でないためにデータ中に誤認識文字が発生する。このため文字認識結果から単純に全文検索を行うと、誤認識に起因する検索もれ、検索ノイズが発生する。この問題を解決するために遊佐ら[1]は文字認識を行わず画像特徴で文字を表現し、仙田ら[2]は文字切り出し／文字認識候補をラティスで表現することで対処した。しかしこれらの方法は文字切り出し誤りに対応していない、データ保持のための容量が大きい等の問題点があった。そこで我々は、データ蓄積時に認識候補文字と各文字の画像特徴の双方を作成し、これらを併用した検索を行うことによって誤認識および文字切り出しエラーに耐性を持つ全文検索方式の検討を行った。

2. 文書データ蓄積処理

文書画像に対し、文書データ蓄積方法の流れを以下に示す。

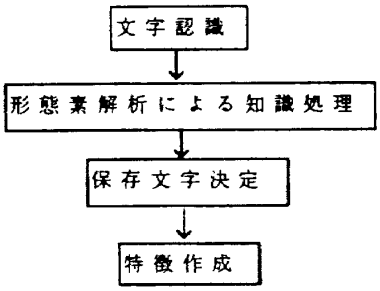

図1　文書データ蓄積処理

以下、各処理について述べる。

A Study on a Retrieval Method Using Recognized Texts.
Taizo Kameshiro ,Takashi Hirano ,Yasuhiro Okada , Fumio Yoda
Imformation Technology R&D Center,
Mitsubishi Electric Co.

2.1 文字認識処理

文字認識は、多値パターンから作成した特徴1024次元を正準判別分析により256次元に圧縮する。各文字画像に対し、標準パターンと比較してそれぞれ10個の認識候補文字を決定する。

2.2 知識処理

第10位までの認識候補文字から形態素解析を行い、日本語として適した文章を出力する。形態素辞書に認識結果が存在しなかったり、接続条件を満たさない文字は第1位認識候補文字を出力する。

2.3 保存文字の決定

文字認識と知識処理の結果を用いて第1位認識候補文字が正解であるか否かを推定する。推定には、正解グループと不正解グループとのマハラノビス距離を用いる。いま、文字認識処理で得た第1位候補文字の類似度x1、第1位候補文字と第2位候補文字の類似度差 x2、知識処理で得た形態素長 x3、および文字種類（英数字記号かそれ以外）x4からなる変数 $\mathbf{x}=(x1, x2, x3, x4)$ と学習データより予め求めた正解グループの平均 $\mathbf{y}=(y1, y2, y3, y4)$ に対し

$$D^2=(\mathbf{x}-\mathbf{y})'\,\mathbf{R}^{-1}(\mathbf{x}-\mathbf{y}) \quad \text{.....(1)}$$

により正解グループとの距離を計算する。Rは正解グループの分散共分散行列である。同様に非正解グループとの距離を求め、距離を比較することにより文字認識第1位候補が正解であるか否かを判定する。正解と判定した文字は、第1位候補文字のみを保存する。正解でないと判定した文字は、複数の候補文字と文字の画像特徴の双方を保存する。画像特徴の作成方法を以下に示す。

2.4 特徴作成

2.3 で特徴を保存すると判定した文字に対し文字認識後の文字矩形を領域分割し、各領域内における

文字画像の輪郭方向特徴を作成する。特徴量は8[バイト／文字]である。

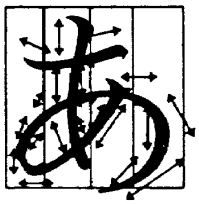

図2 方向成分特徴の作成

3 検索処理

検索処理の概要を以下に述べる。

3.1 文字コードによる照合

文書データ内の文字コードとキーワードとの照合を行い、キーワードと一定割合以上の文字コードが一致する部分を選択候補領域とする。

3.2 特徴による照合

選択候補領域内において文字コードが一致しない部分に対し特徴による照合を行う。照合はDPマッチングを用いる。キーワードと一致しない部分は標準パターンの特徴を用い、文字パターンはそのまま用いる。

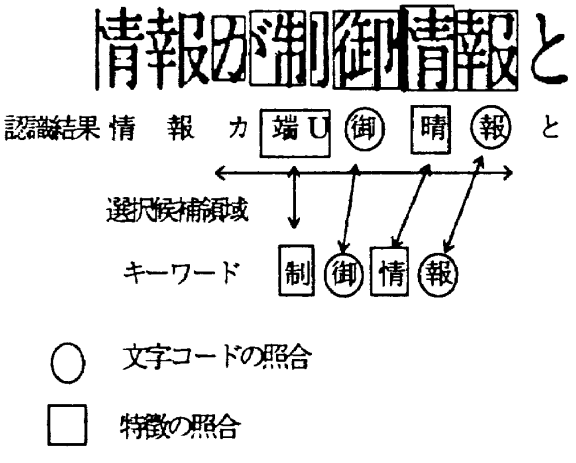

図3 検索処理の例

3.3 候補領域の決定

一致する文字数が一定割合以上で、DPの評価値が一定値以下の場合の領域を検索結果の候補とする。

4 実験

本方法の性能を検証するために以下の評価実験を行った。評価データは、技術論文の要約文(約400文字)40ページを10ポイント文字で印刷し、スキャナの濃度を適正、非常に薄い、非常に濃いの3種類で計120枚作成した。平均認識率は94.1%、10位累積分類率は97.1%である。これに対し本方法で文書データを作成し、34種類のキーワード（ひらがな、カタカナ、英数字、漢字からなる平均単語長3.4文字）を用いて検索実験を行った。作成した文書データ容量はテキストの1.7倍であった。比較のために文字認識の第1位候補のみを保持した場合(従来例1)と従来例1で1文字の不一致文字を許すワイルドカード検索（従来例2）と本方式について再現率、適合率の評価を行った。

表1 各方式の再現率、適合率

	再現率	適合率
従来例1	83.8%	99.5%
従来例2	94.5%	24.6%
本方式	97.5%	92.1%

この結果本手法は、従来例1に比べ再現率が非常に高くなっており、従来例2に比べ再現率、適合率の何れも高く、特に適合率は非常に高くなっている。適合率が従来例1に比べ低いが、特徴間の距離でソーティングすることにより、キーワードに近い形状の候補が上位に並び正しい候補領域を見つけやすい。

5 まとめ

文字コードと画像特徴を用いることで、誤認識に耐性のある検索が可能となることを示した。実験で作成した文書データはテキストの1.7倍程度であり、少ない保存容量で実現できることも示した。今後は、英文への対応および検索速度の向上を図る。

参考文献

[1]遊佐他 “トランスメディアシステムの日本語への拡張”,1994,49回情処全大

[2]仙田他 “全文検索可能な文書画像データベースシステムの開発”,1996,第8回ディジタル図書館ワークショップ（図書館情報大）