131 印刷漢字認識における
パターンマッチング方式

鈴木達郎 (東大) 田中英彦 (東大) 元岡達 (東大)

§1 はじめに

昨年発表した印刷漢字認識システム[1]特にパターンマッチング方式に関する評価結果がまとまったので、報告する。

本システムは漢字の幾何学的構造に注目して、グラフ構造としてとらえている。その時に巨視的な情報(大きさ、長さなど)をいかにグラフ構造作成に反映させるかを述べ、その効果の大きいことを示した。

今回述べるパターンマッチング方式(LCSマッチング)は、このグラフ構造どうしの類似度を定量化するものである。

現在、グラフ構造の類似度を求めるのに確立した手法はないが、本方式は計算時間などにおいても、充分実用に耐えうるものである。

§2 グラフ構造のマッチングアルゴリズム

漢字から特徴抽出して作成したグラフ構造をplex構造として表現する。

これには、各ノードの記号、座標、線幅が含まれるが、図1のように、ノードに記号をつけたグラフと見ることができる。

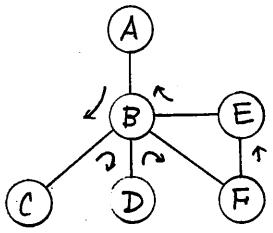

図1 グラフ構造

これをまず一次元化する。例えば、pre-order方式で取り出すことができる。即ち、上から下へ、左から右へ順にたどって行く。(ただしloopを考慮する必要がある。) 図1の例だと、

Ⓐ—Ⓑ—Ⓒ—Ⓓ—Ⓕ—Ⓔ—Ⓑ となる。

この一次元化の操作は、一意に定まる。また各記号ごとに、そのノードから出る枝の数が一定なので、一意に復元できる。

次にこの一次元化されたコード列どうしを比較すればよい。

§3 LCSマッチング

LCS (Largest Common Subsequence)とは、2つの記号列に対して、どちらにも順序をくずさずに含まれる部分列のうちで、最大の長さを持つもののことである。

(例えば ⒶⒷⒸⒹⒻⒺⒷ と ⒶⒶⒸⒺⒹⒻⒼⒷ とのLCSは、ACDFB である。)

ここで2つの記号列M, N (長さがそれぞれm, n)のLCSの長さがlだとすると、M, Nの距離は、

$$m+n-2l$$

で表わすことができる。

これはediting距離[2]と呼ばれるもので、記号列Mから記号列Nへ、最も少ない操作で変換しても、(m−l)回のdeleteと、(n−l)回のinsertの合計(m+n−2l)回の操作が必要なことを示している。

グラフ構造どうしの類似度として、このediting距離を、そのまま採用してもよいが、これを0から1の間に正規化して

$$\frac{l}{\max(m,n)}$$

を類似度として、採用した。

§4 LCSの長さを求めるアルゴリズム

LCS及びその長さを求めるアルゴリズムは、いくつか発表されているが[2]~[4]、長さを求めるだけでよいので、m×n

の時間と，これの記憶場所で実現できる。

§5　LCSマッチング方式の特徴

- グラフ構造の類似度をノード数の積に比例する時間で求めることができ漢字の複雑さを考慮しても，充分実用になる。
- 記号の種類をk個，列の長さをlとすれば，現れうる記号列の場合の数はk^lになり，意味のない場合を除いても非常に多い。（記号の数は実効的に10個．長さの平均は24.4なので$10^{24.4}$になる。）従って情報量も大きい。
- 記号列どうしは長さが異なっても，比較することができる。
- 状態遷移方式と比べて，記号内の順序による重みづけはなく，平等で，情報量を有効に使える。

§6　システムの評価

印刷漢字認識システム全体として，これまで認識実験を行なって来たのでその結果を示して，評価について述べたい。

認識実験は，入力と学習を通して作られた辞書とのLCSマッチングを行ない，類似度が0.7以上で最も大きいものを認識結果として判定する。

この際，類似度が0.7以上のものがなければ，棄却され，学習を行なう。

即ち，その入力の記号列を新しく，辞書の中に追加登録する。

また，誤認識を行なった場合は，データが異常でないことを確かめた上で同様に学習を行なう。

次々とこの認識及び学習をくり返した結果が図2のグラフである。

1. 棄却後正読率：棄却された文字は除いて正読率を計算．
2. 完全正読率：棄却は誤りとみなして計算．
3. 単純正読率：0.7以下の類似度でも強制的に判定させたもの．
4. 棄却率：棄却されたものの数．

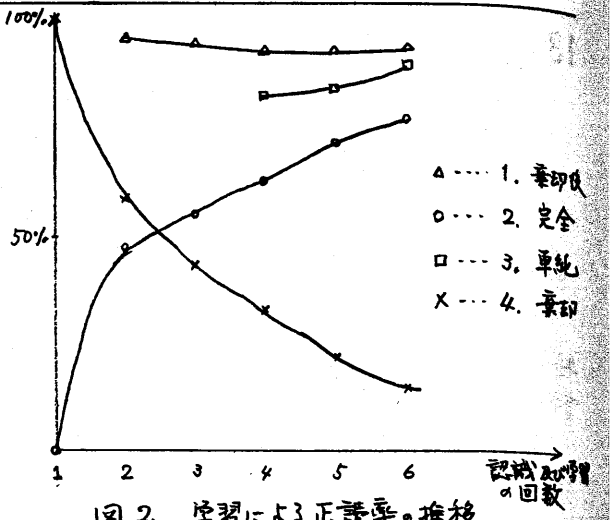

図2．学習による正読率の推移

この認識実験は，教育漢字881字ゴシック1号活字を対象に行なった。

漢字のplex構造には，各ノードの座標，長さなども含まれており，その情報を用いて，クラスター化を行なうことや，類似漢字に対する特殊処理など，認識システムとして完成する為には，まだ多くの改良すべき点を残しているが，単一のマッチング方式のみによる認識結果としては，満足すべき結果だと言うことができよう。

§7　今後の課題

以上のように，LCSマッチングは非常に有効な手法であることがわかったが，今後はこれを実用化するシステムとして，マイクロコンピュータなどを用いた処理方式，特に，パターン認識に含まれる並列性を活かした並列処理方式をalgorithm orientedに開発する予定である。なお，本研究に要した費用の一部は，文部省科学研究費補助金によった。

（参考文献）

(1) 元岡，田中，鈴木：印刷漢字の特徴抽出 昭和51年度オ17回全国大会講演論文集
(2) R. A. Wagner et.al. "The String to String Correction Problem" JACM, vol.21, No.1, '74, pp168-1
(3) D.S. Hirschberg, "A linear Space Algorithm for Computing MCS", CACM, vol.18, No.6, '75.
(4) Hunt et.al. "A fast algorithm for computing LCS", CACM, vol 20, No.5, '77