統計的手法による文字誤りテキスト検索

5P－10

太田 学
東京大学大学院工学系研究科

片山 紀生　高須 淳宏　安達 淳
学術情報センター研究開発部

1 はじめに

OCR(光学的文字読取装置) を用いると大量の印刷文書のDBへの入力作業が大幅に省力化される。そこで大量の印刷文書を画像で入力し、OCRを使って全文DBを構築する試みもある[1]が、その場合OCRの誤認識への対処が必要不可欠である。現在までに著者らは、この誤認識を訂正するのではなく検索段階で吸収する手法について検討し、類似文字テーブル及び単語部分照合を用いた手法の提案を行なった[2]。

本稿ではさらなる検索効率の向上のために、統計的に得られる文字の連接情報 (2-gram 確率) を用いる。

2 検索の手順

本稿でいう検索は全文検索を念頭においており、テキストは原則として1つの長大な文字列として記憶されている。このテキストに対して入力された検索文字列で検索を行なう手順を以下に示す (図1参照)。

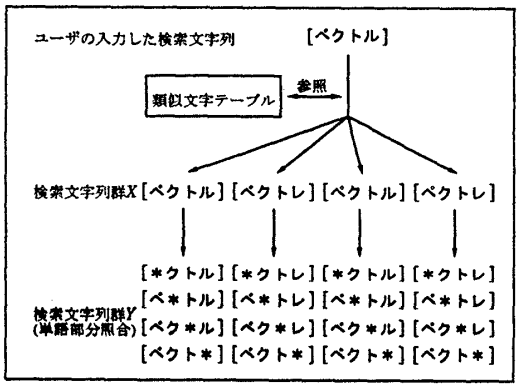

図 1: 検索文字列群の作成

1. 類似文字テーブルを参照してその入力文字列の各文字についてOCRの誤認識の可能性のある全ての文字を用いて新たに検索文字列群 X を作成する。
2. 新たに作成された検索文字列群 X の各検索文字列中の1文字を前から順にワイルドカード (長さ1) に置き換えた単語部分照合用の検索文字列群 Y を作成する。
3. 検索文字列群 X, Y を用いてテキスト検索を行ない、その後2-gram 確率を考慮して確信度を計算し各検索結果の適否を判断する。

Statistical Approach to Text Retrieval Containing Miss-Recognized Characters
Manabu OHTA[1], Norio KATAYAMA[2], Atsuhiro TAKASU[2], Jun ADACHI[2]
[1]Graduate School of Engineering, The University of Tokyo
[2]Research & Development Department, National Center for Science Information Systems

2.1 類似文字テーブル

検索に用いる類似文字テーブルとは、OCRの出力したテキストとそれに対応する誤りのないテキストを比較して得られるもので、OCRが誤る可能性のある文字が全てその確信度とともに格納されている。ここでいう確信度とは、正しいテキストにおける文字が i である事象を A_i、それに対応するOCRのテキストにおける文字が j である事象を B_j とするとき、OCRのテキストにおける文字 j が正しいテキストにおいて文字 i である (と確信できる) 確率 $P(A_i|B_j)$ のことである。この $P(A_i|B_j)$ は、Bayesの定理から次式で表される。

$$P(A_i|B_j) = \frac{P(A_i)P(B_j|A_i)}{\sum_{k=1}^{n} P(A_k)P(B_j|A_k)}. \quad (1)$$

2.2 2-gram 確率の利用

今 $B^{012\ldots n\ldots}$ というOCRの出力したテキストに対して $A_k^1\ldots$ で検索する場合、本手法では類似文字テーブルを参照することで、OCRのテキストにおいては $B^1\ldots$ という文字列で検索する。このとき本手法では次式の確率に注目する。

$$P(A_k^1|B^0) = \sum_i P_2(A_k^1|A_i^0)P(A_i^0|B^0). \quad (2)$$

つまりOCRのテキストにおいてその前の文字が B^0 であるときの次の文字が正しいテキストにおいて A_k^1 である確率である。この式の右辺の $P_2(A_k^1|A_i^0)$ が2-gram確率で $P(A_i^0|B^0)$ がOCRのテキストにおいて B^0 である文字が正しいテキストにおいて A_i^0 である確率 (確信度) である。

類似文字テーブルのみによって求まる $P(A_k^1|B^1)$ を $m_1(A_k^1)$、式 (2) で求めた $P(A_k^1|B^0)$ を $m_2(A_k^1)$ とおく。この2つの確率を統合した新たな確率 $m(A_k^1)$ は、Dempsterの結合規則を用いると次のように求まる[3]。

$$m(A_k^1) = \frac{m_1(A_k^1)m_2(A_k^1)}{1 - \sum_i m_1(A_i^1)(1 - m_2(A_i^1))}. \quad (3)$$

本手法では、この $m(A_k^1)$ を新たな確信度 $P(A_k^1|B^1)$ として用いる[3]。

2.3 文字列の確信度と部分確信度

本手法では、OCRのテキストを検索して得られる文字列が正しい確率を、各文字が正しい確率の積と仮定する。つまり、OCRの文字列 $B^{012\ldots n}$ が正しい文字列 $A^{012\ldots n}$ に対応する確率を、

$$P(A^{012\ldots n}|B^{012\ldots n}) = P(A^0|B^0)\ldots P(A^n|B^n). \quad (4)$$

[3]正確には $P(A_k^1|B^0B^1)$ と書くべきである。

と仮定する。

定義 1 式 (4) の $P(A^{012...n}|B^{012...n})$ のことを、文字列の確信度と定義する。

定義 2 文字列の部分確信度 $PP(A^{012...n}|B^{012...n})$ を以下のように定義する。

$P(A^k|B^k) = \min\{P(A^0|B^0), \ldots, P(A^n|B^n)\}$ のとき、

$PP(A^{012...n}|B^{012...n}) =$

$P(A^0|B^0)...P(A^{k-1}|B^{k-1})P(A^{k+1}|B^{k+1})...P(A^n|B^n)$.(5)

3 検索効率の評価

前節の仮定に基づいて OCR の出力したテキストにおいて検索を行ない、再現率[4] ・適合率[5] の評価を行なった。今回の実験では、(株) リコーの IMAZONE 日本語活字 OCR(公称認識率 99.2%) を使用し、その他の諸条件は以下の通りである。

- 類似文字テーブルの作成には、学習用テキストとして、情報処理学会論文誌 1994 年度の No. 1〜No. 5 の書誌データ (テキストファイルの大きさは約 80kbyte) を用いた。
- open data の検索用テキストとして電子情報通信学会論文誌 1994 年度の Vol. J77-A No. 1〜No. 2 の書誌データ (約 30kbyte) を用いた。
- 2-gram 確率は情報分野の学術論文雑誌のテキスト約 500 万文字について頻度統計をとり、実験における flooring は 10^{-5} とした。
- 検索用単語としては、検索用テキストに含まれる文字列で、名詞のものを無作為に 50 単語を抽出して用いた。検索効率の値は、それぞれその 50 単語で検索・評価した平均である。
- 検索して得られた各文字列には、第 2.3節で定義した従来の確信度 (P) 及び部分確信度 (PP) の他に 2-gram 確率を考慮した新たな確信度 (m) が付けられていて、その値が閾値を越えていれば検索結果とし、さもなくば棄却する。

表 1に closed data における検索結果を示す。表 1からは、OCR の認識率が比較的良い場合はどの検索条件を用いてもあまり差がみられないことが分かる。

表 1: 検索効率 (closed data 1)

検索条件	再現率 (%)	適合率 (%)
完全照合	96.62	100.0
$P \geq 0.01$	99.70	100.0
$(P > 0) \cap (PP \geq 0.01)$	99.70	100.0
$m \geq 0.01$	99.70	100.0

次に同じ closed data に対して、検索用単語として OCR の認識が弱い片仮名の名詞を無作為に 50 抜き出して検索した結果を表 2に示す。各検索条件の閾値を等しくしたところ、条件 ($m \geq 0.01$) で検索した場合の再現率が最も高くなった。これは、検索すべき文字列の確信度が 2-gram 確率を考慮することによって、引き上げられたことを示している。

表 2: 検索効率 (closed data 2)

検索条件	再現率 (%)	適合率 (%)
完全照合	86.17	100.0
$P \geq 0.01$	94.11	100.0
$(P > 0) \cap (PP \geq 0.01)$	98.13	100.0
$m \geq 0.01$	99.13	100.0

類似文字テーブルの学習と 2-gram 確率の学習のどちらにも用いなかったテキストに対する検索における本手法の有効性を示す (表 3参照)。完全照合において適合率が 100.0%でないのは、OCR の誤認識によってある文字列が既存の別の単語になってしまったからである。

表 3: 検索効率 (open data)

検索条件	再現率 (%)	適合率 (%)
完全照合	96.47	99.60
$P \geq 0.01$	96.63	99.60
$(P > 0) \cap (PP \geq 0.01)$	96.77	98.67
$m \geq 0.01$	99.60	99.15

4 まとめと今後の課題

本稿では、OCR の出力したテキストにおける検索手法において統計情報を用いることを試み、確信度の値が改善されることを確認した。しかしこの改善された確信度も、検索効率に対して有効となるか否かは閾値次第なので、閾値の設定という問題が残る。

今後は、このような問題を踏まえた上で文字の挿入や欠落への対処及び計算量の評価を行なっていく予定である。また最近では大容量テキストから n-gram 統計 ($n \geq 3$) をとることも可能になってきており[4]、そのような n-gram 確率 ($n \geq 3$) を用いることも検討してみたい。

参考文献

[1] Andreas, M. and Ulrich, G.: "Fuzzy Full-Text Searches in OCR Databases," Advances in Digital Libraries, pp.87-100, SPRINGER-VERLAG (1995).

[2] 太田学, 高須淳宏, 安達淳: "誤りを含むテキストにおける検索の一手法," 情報処理学会第 51 回全国大会, 7E-8(1995)

[3] 小林邦勝, 鈴木伸明, 根元義章, 佐藤利三郎: "Dempster と Shafer の確率理論に基づく情報量に関する一考察," 電子通信学会論文誌, Vol.J68-A, No.8, pp.741-747(August 1985).

[4] 長尾眞, 森信介: "大規模日本語テキストの n グラム統計の作り方と語句の自動抽出," 情報処理学会研究報告 93-NL-96, pp.1-8(July 1993).

[4] 再現率 = $\frac{OCRのテキストを検索して得た単語のうち正しいものの数}{正しいテキストを検索して得た単語数}$

[5] 適合率 = $\frac{OCRのテキストを検索して得た単語のうち正しいものの数}{OCRのテキストを検索して得た全単語数}$