冗長インデクスを用いたＯＣＲテキスト検索システム

7Ｌ－4

佐藤 研治　赤峯 享　村木 一至

NEC 情報メディア研究所

1．はじめに

インターネット/イントラネットの爆発的な普及に伴い、テキスト資産の共有の重要度がますます高まっている[1]。ローカルエリアでは情報を共有するためにグループウェアが用いられ、テキスト資産が活用されている。本研究所では、利用者のテキストに対する作業履歴やテキスト間の引用/参照関係を用いてテキストデータをインデクス化し、テキスト資産が増進的に共有/活用される環境の研究を行っている[2]。

ＯＣＲ認識技術は、英語では非常に高い認識率になるが、日本語では、現在はそれほど高い認識率ではない。ＯＣＲ認識を行った後に、言語的処理により辞書とマッチングし自動的に誤り訂正を行う技術もあるが、認識誤りを完全に除くことはできていない。ＯＣＲ認識後の人手による修正作業のコストは非常に高いため、テキスト資産のオンライン化、共有/活用を行う際の問題点となっている。

一般的な紙の文書を利用する場合を考えてみると、オープン世界のテキスト資産を活用する際の利用は、テキスト参照が主目的である（参照中のテキストの一部をカット＆ペーストして利用するというのは、紙の文書では無理である）。この点に着目し、低コストのＯＣＲテキスト認識システムを構築するため、本論文では、テキスト資産のイメージデータを正しいテキストに完全に変換するのではなく、検索が可能なようにイメージに対してインデクスを張り、そのインデクスを利用してテキストイメージを検索/提示する方法を提案する。

2．ＯＣＲテキストのインデクス化法

テキスト資産のイメージデータに対し、検索可能なインデクスを付与する手法として、データの内容によるインデクスと、データの利用コンテキストによるインデクスの異なる2方面のインデクス化方法を提案する。

a)　イメージデータをＯＣＲ認識させ、誤りを含む文字列を用いてそのままインデクス化

b)　検索者が検索結果を参照しつつ作成した文書で、参照されたテキストをインデクス化

インデクス化法a)は、ＯＣＲ認識させたテキストそのままの文字列でイメージをインデクス化する。このテキストデータは誤りを含むため、このインデクスに対しては検索する際に、ＯＣＲ認識誤りに応じて検索条件を展開する検索法と併せて用いる。この検索条件の展開法の例としては、検索条件の各文字を１つずつワイルドカードにして展開したもの（例：「コンピュータ」を「ンピュータ＋コ？ピュータ＋コン？ュータ＋コンピ？ータ＋コンピュ？タ＋コンピュー」とする）や、ＯＣＲが誤りやすい文字候補に展開する方法（例：「内蔵」を「内威＋内減＋…＋丙蔵＋…」とする）等がある。また、通常のテキスト検索と同様に、シソーラスによって単語を展開する方法もある。

インデクス化法b)は、検索結果文書を参照しながら作成した文書のタイトル部分だけを抽出したり、参照文書を表示している間に入力された文字列からキーワードを抽出して、これらの文字列を用いてインデクス化する方法である。インデクス化法b)は、文書が利用されるとリッチになる動的なインデクス法であり、インデクス化法a)を補う役割を果たす。インデクス化法a)で作成されたインデクスは固定的なものである。もし、インデクス化法a)によるインデクスのみしか用いなかった場合、ＯＣＲの誤りが上記で述べた「検索条件で展開する検索法」で対応出来ないもので会った場合、そのテキストはその検索条件では検索できない。しかし、b)のインデクス化法を併せて用いることにより、その文書が他の検索条件

Information Retrieval Supported by Rich and Redundant Indices for Erroneous OCR Text

Kenji SATOH, Susumu AKAMINE and Kazunori MURAKI

Information Technology Research Labs. NEC Corp.

で検索されて利用された際に新たなインデクスが付き、その後の利用者からは検索可能になっていくことになる。

これらの２つのインデクスは、相補的なインデクスである。インデクス化法a)だけでは、検索条件の展開誤りによる検索もれや余分な検索があるが、インデクス化法b)を併せて用いると、２つのインデクスが相補的に働き、長期的にみればＯＣＲ認識したテキストを人手で修正したり、自動的に修正してインデクス化した場合以上の検索のしやすさを持つテキスト資産データベースとなるであろう。

3. 評価

これらのインデクス化法の組み合わせで、どの程度の検索精度となるかを推定するため、インデクス化法a)に関して、検索条件の展開法によって再現率がどれほど向上するかの評価実験を行った。データは紙のデータをスキャナで読み込ませて用いたが、このデータに対する正解データがないため、紙のデータから目で単語を抽出し、その単語を検索条件として入力した際に、正しく抽出元データが検索されるかどうかで再現率を評価した。データは学会予稿集94件（188ページ）を用いた。各発表1件につき「２文字の単語」「３文字以上の単語」を１つずつ抜き出し、その単語による検索で抜き出し元の発表が検索されるかどうか見た。検索条件の展開法は「２文字単語」に「誤りやすい文字による展開」を、「３文字以上の単語」に「ワイルドカード展開」を適用した。この評価結果を表１に示す。

評価結果として、文字列展開を行わない場合に、２文字単語の方が３文字以上の単語よりも正しく検索できることが判った。（２文字単語は９１％、３文字以上の単語は７７％）。これは、単純に単語が長くなるとそのなかの１文字以上が誤る確率が高くなっていくためであろう。また、２文字単語の「誤りやすい文字によるによる展開」は２％しか再現率が上がらなかった。一方、３文字以上の「ワイルドカードによる展開」は１２％再現率が向上した。これは、もともと２文字単語の認識率がそれほど低くなかったことも関係していると考えられる。更に、より大きな要因として、誤りやすい文字候補を第１０候補まで利用したが、実際の誤りのパターンはもっとバリエーションに富んでおり、１０候補だけではあまり展開が効かなかったことが挙げられる。

表１：ＯＣＲテキスト検索システムの再現率

	２文字単語	３文字単語
展開なし	86/94 0.91	79/94 0.77
文字列展開	87/94 0.93	84/94 0.89

4. おわりに

ＯＣＲ認識したテキストに対し、人手による修正作業が不要なテキストのインデクス化および検索法を提案し、その検索時の検索条件の展開法で、認識誤りが含まれていてもある程度正しく検索可能であることを示した。

今後の課題としては、インデクス化法b)まで含めたテキスト資産活用環境を構築し、実際の運用の中で、どれほどの再現率になっていくのかを評価することがある。また、本論文で述べた手法で提示したイメージ上を範囲指定しコピーした際に、その座標情報よりＯＣＲ認識したテキストの対応部分をコピーすることで、誤りを含みつつも部分的にテキストをカット＆ペーストできるサービスを提供することも可能である。このサービスを利用した場合には、コピーした範囲のテキストの誤りを利用者がその後訂正することが考えられ、その訂正テキストでＯＣＲ認識テキストを置き換えることで、だんだんと正しいテキストにＯＣＲテキストを変えていくことも可能である。これらの手法を統合して用いることで、より再現率の高いインデクスを張ることができるテキスト資産活用環境を、今後構築していく予定である。

[1] Peter J. Nurnberg, Richard Furuta, John Leggett, Catherine C. Marshall, Frank M Shipmann III: Digital Libraries: Issues and Architectures, Digital Libraries '95, June 1995.

[2] Kenji SATOH and Kazunori MURAKI: Penstation for Idea Processing, Natural Language Processing Pacific Rim Symposium (NLPRS'93), Dec 1993.