学術文献画像の書誌情報の近似マッチング法

高須淳宏[†] 片山紀生[†] 大山敬三[†]
安達淳[†] 影浦峡[†]

本論文は，学術雑誌を対象とした電子図書館において文献の引用関係を文書画像から抽出する方法を提案する．この手法は，これまで大量に紙媒体に蓄積されてきた学術論文を効率良く電子化するための一要素技術と位置づけられる．提案手法のポイントは，画像認識された参考文献データと書誌データベースの高速な近似マッチング法にあり，文献タイトルの部分文字列を用いたインデクスを用いることによって高速近似マッチングを実現した．2 百万件規模の書誌データベースを用いて実験したところ，約 87.2%の精度を得ることができた．

An Approximate Matching Method for Bibliographic Data in Academic Article Images

Atsuhiro Takasu,[†] Norio Katayama,[†] Keizo Oyama,[†]
Jun Adachi[†] and Kyo Kageura[†]

This paper proposes a method that extracts reference relationship among articles in digital library of academic journals. The proposed method is regarded as one of fundamental techniques to capture articles efficiently from printed journals. The key of the proposed method is a fast approximate matching of reference data obtained from scanned article images and records in bibliographic databases. The indices consisting of title substrings enable fast retrieval of candidate records from large database. The proposed method achieved 87.2% accuracy in the experiment using a bibliographic database containing about 2 million records.

1. はじめに

大規模なデータベースを構築するうえで重要な問題の 1 つにデータの収集と形式変換の問題がある．大規模データを効率良く収集するためには，事前に定められたスキーマに従って新たなデータを作り出すとともに，既存のデータを効率良く活用するメカニズムが必要になる．たとえば，電子図書館を考えた場合，新規に作成されるデータを収集するメカニズムとして，電子ジャーナルのようにデータの作成から蓄積，活用までのプロセスを統一的に扱うことが試みられている．このような新規データの収集と並行して既存のデータをデータベースに取り込むことによって大規模データベースを構築することが可能になる．

既存データの活用方法については，これまで情報が主に紙に印刷されて蓄積されてきたこともあり，OCR をはじめとする文書画像解析の研究が行われてきた[3]．この研究では，文字認識や領域分割といった文書画像処理の基本要素技術の研究と並行して，章や節といった文書の構成要素を取り出す研究が進められた．これらの文書画像の構造解析技術は，SGML 文書などの自動生成へと発展しており，データベース構築の観点からも重要な技術と考えられる．

しかし，データベース中の文書を活用することを考えると，文書の構造情報だけでは不十分で，ハイパーテキストのような文書間の関連情報も必要になる．文書間の関連性を抽出するためには，文書画像解析によって得られるデータ間のマッチング技術が必要になる．本論文では，文書間の最も基本的な関連の 1 つとして，学術雑誌の参考文献と文献本体との引用関係を取り上げ，文書画像から引用関係を抽出するためのマッチング法について検討する．文書画像における参考文献領域の解析に関する研究は，これまであまり行われておらず，筆者らが知る限りでは Belaid らによる参考文献領域の抽出と著者名や文献タイトルなどの書誌項目の抽出の研究[2),13)] にとどまっている．本論文では，文書画像解析によって抽出された書誌データ

† 国立情報学研究所
National Institute of Informatics

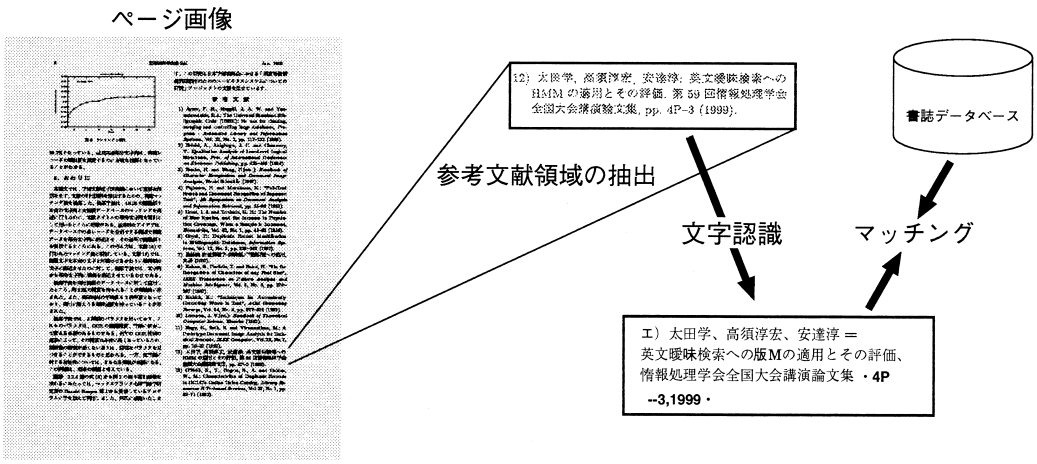


図 1 書誌マッチング過程の概要
Fig. 1 Outline of bibliographic matching process.

を用いたマッチング法を提案する．

以下では，2 章で本論文が扱う書誌マッチング問題の概要を述べ，その問題点について議論する．3 章では，画像解析によって得られたデータに対する書誌マッチング法を示し，4 章で高速にマッチングを行うためのインデキシング法を提案する．5 章では，学術論文を用いた実験結果を示す．

2. 書誌マッチング問題の概要

文献の引用情報を抽出するために，画像解析によって得られた文献データどうしのマッチングを考えることもできるが，ここでは，書誌データベースのレコードと画像解析によって得られた参考文献データのマッチングを考える．その理由は，まず第 1 に，学術文献の電子図書館においてはシステムが含む文献の書誌データベースが作成されるため，このデータベースとのマッチングが必要になる．第 2 に，これまで多くの書誌データベースが構築されてきており，これらのデータベースとのマッチングを行うことによって，仮に目的とする文献が電子図書館に含まれていなかったとしても，参考文献のより詳細な情報を得られる可能性がある．第 3 に分散電子図書館では，同一の文献が複数のサーバに含まれることもあり，書誌データベースとのリンクを考えることによって柔軟な引用関係の取扱いが可能になる．**図 1** は，学術文献の画像データと書誌データベースのレコードとのマッチング過程の概要を示している．本論文では，この過程の中で，文字認識された参考文献文字列と書誌データベースとのマッチングについて述べる．

書誌マッチングは，これまで書誌データベース中の重複レコードの削除問題として研究されてきた．O'Neill らは，OCLC の総合目録データベースに重複して登録されているレコードを抽出し，重複レコードにおいて値の異なるフィールドの傾向を調べている[12]．一方，重複レコードを自動的に発見する方法として文献タイトルや著者名などの代表的なフィールドの値を用いてレコードをコード化する手法が提案されている[1]．Goyal は，いくつかの重複レコード検出方法を比較している[6]．また Ridley は，エキスパートシステムを用いることによって，重複レコードの発見の精度を上げるとともに，重複レコードの中から優れたレコードを選び出すシステムを提案している[14]．

本論文が対象とする文書画像解析によって得られた参考文献のマッチングには，以下の問題点がある．

OCR の文字認識誤り

書誌データベースにおける重複レコードの削除問題では，入力ミスに関する検討も含まれているが，それほど注意は払われていない．OCR を利用する場合には認識誤りが避けられず，エラーの扱いが中心的な課題になる．OCR の認識誤りには，置換，挿入，削除，複合誤りの 4 種類がある[9]．このうち，置換誤り以外の誤りでは，文字認識の過程で文字列長が変わってしまい，本来の文字列と認識された文字列の位置の対応をとることが難しくなる．本論文では，このように文字列長を変える誤りを非置換誤りと呼ぶことにする．なお，本論文が対象としている参考文献は，日本語と英語の混在の度合いが高く OCR の認識誤りが比較的起きやすい文字列となっている．

書誌構造の多様さ

参考文献の表記法は，雑誌によって異なるため，書

誌データベースの重複レコード処理のようにタイトルや著者などの書誌項目が正しく得られていることを前提とした手法は適していない．

表記の揺れ

参考文献には，雑誌名の短縮形などさまざまな表記上の揺れがある．

処理効率

マッチングの対象となるデータベースは通常数百万件から数千万件のレコードを含んでおり，特に動的に引用関係を求める場合には，効率良くレコードを抽出する必要がある．

Belaid らの研究[2),13)]は，文書画像から書誌項目を抽出することを目的としている．そこでは，文字認識誤りと書誌構造の多様性に対処するために，著者名の辞書や雑誌名の省略形のリストを用意するとともに，多様な参考文献の構造を表す文法を用いている．一方，本論文では，書誌データベースが存在することを前提に，OCR の認識誤りに頑健で，大規模書誌データベースに対しても効率的に処理可能な書誌マッチング法を提案することを目的としている．

3. 書誌マッチング

本論文で提案する書誌マッチング法は，文書画像に文書構造解析および文字認識を適用して得られる参考文献データに対して，以下の処理を行う．

step 1 文献タイトルの抽出

step 2 部分文献タイトル文字列を用いた候補レコードの選択

step 3 最長共通部分文字列に基づいた候補レコードのランキング

候補レコードの選択において文献タイトルのみを用いるのは，多くの文献において文献タイトルが参考文献中に含まれること，文献タイトルは他の書誌項目と比較して長い文字列を持ち OCR の認識誤りへの対処が比較的容易であること，著者名や雑誌名のような表記の揺れが少いことによる．step 3 では，参考文献データ全体を使って類似度をはかることによって表記の揺れの影響を相対的に軽減させている．本章では，step 1 および step 3 について述べ，4 章で大規模データベースに対する高速マッチングの鍵となる step 2 について述べる．

3.1 step 1：文献タイトルの抽出

文書画像解析の結果得られる参考文献は，以下の下線部のように文字認識誤りを含んだ文字列となる．

エ）太田学，高須淳宏，安達淳＝英文曖昧検索への版M の適用とその評価，惰報処理学会全国大会講演論文集・4P ——3,1999・

一般に，参考文献中の書誌項目は，区切り記号で区切られている．たとえば，情報処理学会論文誌では，文献タイトルはコロンとカンマで区切られている．そこで，本手法は，正規文法によってパターンを記述することによって文献タイトルを切り出す．たとえば，上記の参考文献に対しては，以下の文字列が文献タイトルとして抽出される．

英文曖昧検索への版M の適用とその評価

なお，上記の例の「＝」のように区切り記号が誤認識される可能性がある．そこで，各区切り記号に対して OCR が誤認識する可能性のある文字をあらかじめ列挙しておき，それらの文字も区切り記号として扱うように正規文法を拡張することによって区切り記号の認識誤りに対処する．文法の曖昧さから，複数の文字列がタイトル文字列として抽出される可能性があるが，その場合はすべての文字列を step 2 で用いる．つまり，step 1 で求められた各タイトル文字列に対して候補レコードを求め，その和集合を候補レコードとして，step 3 においてランキングを行う．

3.2 step 3：類似度の計算とランキング

step 3 は，step 2 で得られた候補レコードと文書画像解析の結果得られた参考文献データを比較し，文字列の類似度に基づいてランキングを行う．参考文献における著者名や文献タイトルなどの書誌項目の順序は雑誌ごとにおおよそ固定されている．そこで，各雑誌について書誌項目の出現順序を調べ，step 2 で得られる候補レコードのフィールドをその順序に従って並べ変える．この文字列と参考文献の文字列の最長共通部分文字列を求め，その文字列長を類似度として用いる．

4. タイトル部分文字列を用いた近似マッチング

4.1 マッチング法

OCR によって得られる誤りを含んだ文字列のマッチングには，従来から，任意の文字の組合せ c_i，c_j に対して OCR が c_i を c_j と誤認識する確率を用いる Confusion Matrix 法[8)]，OCR の誤りパターンをオートマトンで表す方法[4)]，OCR の誤りパターンを隠れマルコフモデルで表す方法[11)]などが提案されてきた．しかし，これらの手法をタイトル文字列のマッチングに適用するためには，データベース中のすべてのレコードに対して，参考文献のタイトル文字列とのマッチングを行う必要がある．そのため大規模な書誌データベースに対して，これらの手法を適用することは処理時間の観点から望ましくない．

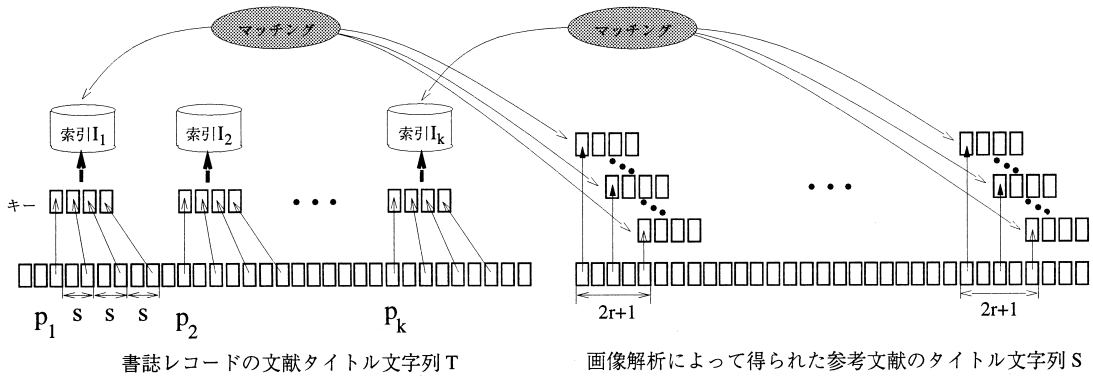


図 2 タイトル文字列に対する索引
Fig. 2 Index for title strings.

そこで，タイトルの部分文字列をキーとする索引を複数個作成し，いずれかの索引で参考文献タイトルと一致するキーを持つ書誌レコードを候補レコードとして抽出する近似マッチング法を提案する．**図 2** にこのマッチング法の概要を示す．

まず，以下で用いる記法を定義する．文字列 T に対して $|T|$ は T の長さを表す．また，文字列 T に対して相対位置 p $(0 < p \leq 1)$ の文字とは，$\lceil p|T| \rceil$ 番目の文字を意味する．

提案する近似マッチング法は，書誌データベースに対して，タイトルの部分文字列をキーとする k 個の索引 $I_1, I_2, \cdots, I_k$ を用いる．索引 I_i には，相対位置 p_i が割り当てられ，書誌レコードの文献タイトル文字列において開始位置が相対位置 p_i から始まる長さ l の部分文字列がキーとして用いられる．ここで，キーの各文字は s おきにとられる．つまり，索引 I_i は，書誌データベース中の各文献タイトル文字列 T の $\lceil p_i|T| \rceil, \lceil p_i|T| \rceil + s, \lceil p_i|T| \rceil + 2s, \cdots, \lceil p_i|T| \rceil + (k-1)s$ 番目の文字によって構成される．以後，データベースの索引のキーとして用いられる部分文字列を索引文字列と呼ぶ．たとえば，$l = 3$，$k = 2$，$s = 2$，$p_1 = 0.1$，$p_2 = 0.6$ とし，文献タイトル文字列「英文曖昧検索への HMM の適用とその評価」の索引文字列を考える．まず，タイトル文字列長が 19 であるため，相対位置 0.1 は 2 番目の文字に対応する．したがって，索引 I_1 に対する索引文字列は「文昧索」となる．同様に，索引 I_2 に対する索引文字列は,「の用そ」となる．

一方，文書画像解析の結果得られる参考文献のタイトル文字列 S に対しても同様に部分文字列を作成する．ただし，ここでは，この文字列の先頭位置を前後に幅 r の範囲でずらしたものも用いる．つまり，先頭位置として $\lceil p_i|S| \rceil$ に加えて $\lceil p_i|S| \rceil - 1, \lceil p_i|S| \rceil - 2, \cdots, \lceil p_i|S| \rceil - r$ および $\lceil p_i|S| \rceil + 1, \lceil p_i|S| \rceil + 2, \cdots, \lceil p_i|S| \rceil + r$ も用いる．したがって，1 つの索引に対して $2r + 1$ 個の部分文字列を作成することになる．以後，参考文献のタイトル文字列から作られる部分文字列を検索文字列と呼ぶ．たとえば，先頭位置の幅 r を 1 とし，3.1 節のタイトル文字列「英文曖昧検索への版 M の適用とその評価」を考える．$p_1 = 0.1$ の索引に対しては,「英曖検」,「文昧索」,「曖検へ」の 3 つの検索文字列が，また $p_2 = 0.6$ の索引に対しては「M 適と」,「の用そ」,「適との」の 3 つの検索文字列が作成される．

書誌データベースに対する検索結果は，各索引に対する検索によって得られたレコードの和集合となる．索引 I_i に対する検索文字列を $S_{i,-r}, \cdots, S_{i,0}, \cdots, S_{i,r}$ とし，$S_{i,j}$ と一致する索引文字列を索引 I_i に持つ書誌レコードの集合を $T_{i,j}$ とすると検索の結果は，

$$\bigcup_{i=1}^{k} \bigcup_{j=-r}^{r} T_{i,j} \tag{1}$$

となる．

部分文字列をキーとして用いるねらいは，OCR の認識誤りを避けることと B-tree などの高速なインデキシング法を適用することによって大規模なデータに対して高速な検索を実現することにある．キーの長さを短くすることによって，検索の精度つまり認識誤りを避けて正しいレコードを選択する確率を高めることができるが，候補レコード数が増加する．一方，索引数を増やすことによっても同様な効果がある．そこで，キーの長さと索引数を調整することによってトレードオフ関係にある精度と候補レコード数のバランスをとることが必要になる．キーに使う文字の距離を s (≥ 1) としたのは，同一長の部分文字列でも離れた文字を用いることによって候補レコード数をおさえる効果を期待できるためである．

部分文字列をキーとして用いるときの問題点は，非置換誤りによって，データベース中の文字列と認識された文字列の位置の対応がずれることにある．たとえば，3.1 節の例では,「HMM」が「版 M」と認識されたことによってこれ以降の文字位置の対応がずれている．相対位置を使うことと，検索文字列の先頭位置に幅を持たせることによってこのような問題にある程度対処することが可能になる．

以下では，トレードオフの関係にあり，提案手法のパラメータのチューニングにおいてポイントとなる精度と候補レコード数に関する提案手法の特性について議論する．

4.2 精　　度

本節では，まず，OCR の認識誤りが各文字独立に起こり，認識誤りは置換誤りだけであるという仮定のもとで部分文字列長 l および索引数 k の近似マッチングの精度を示す．次に，非置換誤りと検索文字列の先頭位置の幅との関係を議論する．ここで，精度とは，提案手法によって得られる候補レコード集合の中に正しいレコードが含まれる確率を表す．正しいレコードとは，画像解析によって得られた文献と同一の文献タイトルを持つレコードを意味する．

OCR の認識誤り率を ϵ とすると検索文字列が誤認識された文字を含む確率は，$1-(1-\epsilon)^l$ となる．ここで OCR の認識誤り率が十分低ければこの確率は $l\epsilon$ と近似できる．k 個のすべての検索文字列が誤認識文字を含む，つまり正解レコードの索引文字列と一致する検索文字列がない確率は $(l\epsilon)^k$ となるため，精度は以下の式で表される．

$$1-(l\epsilon)^k \tag{2}$$

式 (2) を導くにあたって非置換誤りは生じないという仮定をおいたが，実際には検索文字列よりも前で起きた非置換誤りによって検索文字列と対応する索引文字列の先頭位置がずれる可能性がある．そこで，以下では，検索文字列の先頭位置に幅を持たせることによって対処可能な非置換誤りのパターンを示す．

文献タイトル T が OCR によって文字列 S として認識されたとする．T において相対位置 p の文字が S において位置 $f(p)$ に現れるとする．先頭位置に幅を持たせることによって T の相対先頭位置 p の索引文字列が，S において検索文字列として用いられるための必要十分条件は以下のように表すことができる．

$$\lceil p|S|\rceil - r \le f(p) \le \lceil p|S|\rceil + r \tag{3}$$

ここで，非置換誤りによって生じる相対位置 p における文字の位置のずれ $d(p) \equiv f(p) - \lceil p|T|\rceil$ を考える．$p=1$ の場合，$d(p)$ は OCR によって認識されたタイトル文字列長ともとのタイトル文字列長の差 $|S|-|T|$ を表すことに注意されたい．以下ではこの値 $d(1)=|S|-|T|$ を δ で表すことにする．式 (3) は，$d(p)$ を用いて以下のように表される．

$$\lceil p|S|\rceil - \lceil p|T|\rceil - r \le d(p) \le \lceil p|S|\rceil - \lceil p|T|\rceil + r \tag{4}$$

ceil 関数の性質より，$\lceil p|S|\rceil - \lceil p|T|\rceil$ は以下の範囲に収まることが分かる．

$$\lceil \delta p\rceil - 1 \le \lceil p|S|\rceil - \lceil p|T|\rceil \le \lceil \delta p\rceil \tag{5}$$

不等式 (4) と (5) より以下の不等式 (6) が成り立てば，先頭位置に幅を持たせることによって非置換誤りによって生じる文字位置のずれに対処できることが分かる．

$$\lceil \delta p\rceil - r \le d(p) \le \lceil \delta p\rceil + r - 1 \tag{6}$$

δ が正の場合，$\lceil \delta p\rceil$ は相対位置 p に対して幅 $1/\delta$ 高さ 1 のステップ関数となる．したがって，不等式 (6) は，相対位置 p における文字位置のずれ $d(p)$ が，幅 $1/\delta$，高低差 $2r-1$ の階段状の領域に入っていれば先頭位置に幅 r を持たせることによって索引文字列と検索文字列の位置をあわせられることを意味している．δ が負の場合も同様の議論が成り立つ．以下では，不等式 (6) を満たす相対位置 p を有効であると呼ぶ．

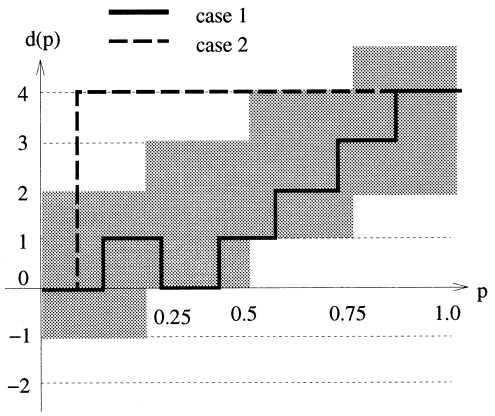


図 3　$d(1)=4$，$r=2$ における $d(p)$ の条件
Fig. 3　Condition of $d(p)$ when $d(1)=4$ and $r=2$.

図 3 は，$\delta=4$，$r=2$ の場合に $d(p)$ が不等式 (6) を満たす領域を示している．$d(p)$ が図 3 の case 1 のような場合，つまり，非置換誤りによる文字位置のずれが大きな偏りなく累積していく場合には検索文字列の先頭位置に幅を持たせることで，ほぼ全域で相対位置は有効になる．一方，case 2 のように非置換誤りがバースト的に起こる場合は，その位置よりも前の部分および終端部のみが有効になる．

式 (2) は，索引数 k を増やすことによって近似マッチングの精度が向上することを示している．非置換誤りが生じると，精度向上に寄与する索引は，索引文字列と検索文字列中のすべての文字で位置のずれがないものに限られる．提案手法は，先頭位置に幅 r を持せることによって先頭位置のずれに対してのみ対策を講じている．先頭以外の位置に対しても同様の幅を考えることは可能であるが，その結果，検索文字列に使われる文字位置の組合せが増加し，処理効率の低下を招く．また，先頭位置のずれと比較しそれ以外の位置の相対的なずれの可能性は低くあまり効果が期待できない．このような理由で，本提案手法では，先頭位置のずれに対してのみ幅を持たせている．なお，非置換誤りを考慮した精度は，式 (2) の k を精度向上に寄与する索引の数に置き換えたものとなる．

4.3 候補レコード数

本節では，提案する近似マッチング法で得られる候補レコード数について議論する．まず，1 つの索引文字列によって得られる候補レコードの平均値について考察する．データベース中のレコード数を n としてデータベースに含まれる異なる索引文字列の集合を W とする．すると，1 つの索引文字列によって得られるレコード数の平均値は，$n/|W|$ となる．したがって，データベース中に含まれる異なる索引文字列の数が多いほど 1 つの索引文字列によって得られる候補レコード数の平均値は小さいことになる．

ここで，索引に使われる文字の集合を C として長さ l の文字列の集合 C^l を考える．データベースのレコード数に対して索引文字列長 l が小さく，$W = C^l$ の場合，つまり，すべての長さ l の文字列が索引文字列として現れている場合，それ以後データベースのレコードの増加に比例して平均候補レコード数も増加することになる．一方，$W \subset C^l$ の場合は，データベース中のレコード増加にともなって W に含まれない文字列が索引文字列として現れることになり，平均候補レコード数の増加もおさえられる．したがって，将来データベースのレコードが増加しても $W \subset C^l$ となるように索引文字列の長さ l を選ぶことが望ましい．

ここで，文字列 $s \in C^l$ がある文献において索引文字列として用いられる確率は，文献によらず一定であるという仮定のもとで，1 つの索引文字列によって得られる平均候補レコード数の期待値を検討する．上記の仮定は，n 個の索引文字列からなる索引 I は，長さ l の文字列の集合 C^l からある確率分布に従って独立に n 個の文字列を抽出した結果であることを意味する．

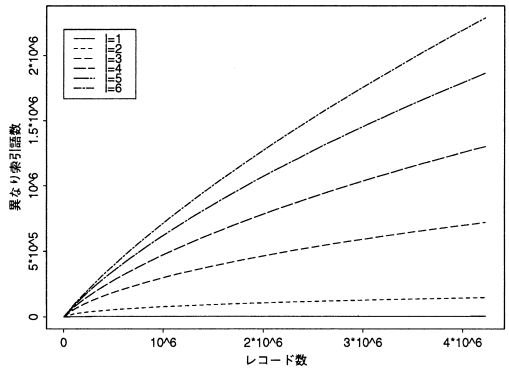
まず，データベースのレコード数が n のときの異なり索引文字列数を w_n，その索引中での出現頻度が f である索引文字列の異なり数を $v_{n,f}$ で表し，その期待値をそれぞれ $E[w_n]$，$E[v_{n,f}]$ で表すことにする．データベースのレコード数が λ 倍になったときに，索引に含まれる異なり索引文字列数の期待値 $E[w_{\lambda n}]$ は，サイズ n のデータベースに含まれる異なり索引文字列数の期待値 $E[w_n]$ と頻度分布 $E[v_{n,f}]$ $(f \geq 0)$ を用いて以下のように表されることが Good ら[5]によって導かれている．

$$E[w_{\lambda n}] = E[w_n] - \sum_{f=0}^{\infty}(-1)^f(\lambda-1)^f E[v_{n,f}] \tag{7}$$

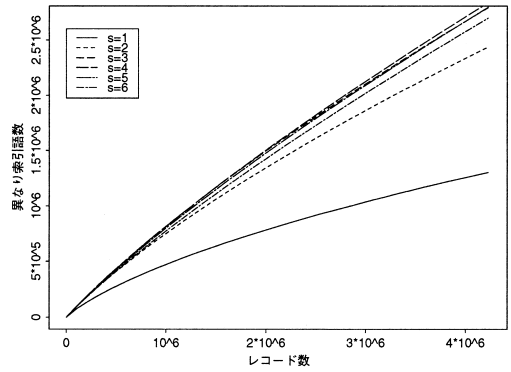
この式は，データベースのレコード数が n のときの異なり索引文字列数の期待値 $E[w_n]$ と頻度 f の索引文字列数の期待値 $E[v_{n,f}]$ からデータベースの大きさが λ 倍になったときに，異なり索引文字列数を予測するための式となっている．式 (7) は，計量言語学における語彙数の推定や計量情報学における論文執筆数の推定などにも用いられている[7]．

図 4 は，5 章で述べる約 2 百万件の書誌データベースから得られた索引文字列の頻度分布を式 (7) の $v_{n,f}$ の期待値と見なして導かれたレコード数と異なり索引文字列数の関係を表している．図 4 (a) は，書誌データベース中の各タイトルの先頭から 1 文字おきに長さ l の索引文字列を作成した場合のレコード数に対する異なり索引文字列数の推移を表している．このグラフでは，レコード数が 100 万件以上では，長さ 1 および 2 の異なり索引文字列数は横ばいとなっている．これは，100 万件規模のレコードを含むデータベースにはすべての部分文字列が出現していることを示しており，その後はデータベースのレコード数 n に比例して候補レコード数の平均値 n/w_n が増加することを示唆している．一方，部分文字列の長さが 3 以上の場合，400 万件程度のデータベースに対しても，レコード数に応じて異なり索引文字列数が増加することを示しており，レコード数の増加に対する平均候補レコード数の増加をおさえることができることを示唆している．

一方，図 4 (b) は，索引文字列として用いられる文字の間の距離 s による異なり索引文字列数の変化を表している．$s = 1$ の場合と比較して $s > 1$ の場合のほうが異なり索引文字列数が多くなり，索引文字列あたりの候補レコード数が小さくなることが分かる．これは，連続した文字を索引として用いる場合に文字の依存関係によって一部の索引文字列の出現確率が高



(a) 索引文字列の長さ vs. データベースのレコード数
(a) Index word length vs. DB record size



(b) 文字間の距離 vs. データベースのレコード数
(b) Inter-character distance vs. DB record size

図 4 索引文字列の長さと異なり索引文字列数の関係
Fig. 4 Relationship between index string length and number of index types.

くなり，結果として異なり索引文字列数が減少することによると考えられる．一般に，文字間の距離 s を大きくすることによって異なり索引文字列数をおさえることができるが，索引文字列に使われる文字列長が長くなり，その間に非置換誤りが生じる可能性が高くなる．また，$s > 1$ の場合の異なり索引文字列数にそれほど大きな差がないので，$s = 2$ が適していると考えられる．

5. 実験結果

5.1 実験の概要

情報処理学会論文誌，電子情報通信学会和文論文誌 D II および人工知能学会誌に掲載された論文と国立情報学研究所で提供されている書誌データベースを用いて本論文が提案する書誌マッチング法の実験を行った．実験に用いた書誌データベースは，2,121,707 件の本および雑誌に掲載された論文の書誌データを含んでいる．画像解析の対象とした論文は，1995 年度に発行された上記学会誌に掲載された論文で，全体で 8,465 件の参考文献を含んでいる．このうち，1,575 件の引用文献が書誌データベースに含まれていた．実験では，文書画像に文献 15) の文書画像解析を適用した後，人手によって参考文献領域の修正を行い正しい領域を抽出した．そして，索引および検索文字列長（l），索引数（k），検索文字列の先頭位置の幅（r），索引および検索文字列に用いる文字の間隔（s）を変化させ，精度，候補レコード数，処理時間を測定した．なお，索引文字列および検索文字列の開始位置は，タイトル文字列からできる限り重複なく文字が選ばれるように索引数に応じて均等に割り当てた．また，検索には，書誌データベースに正解が含まれる 1,575 件の引用文献を用いた．以下では，まず 4 章で述べたタイトル部分文字列を用いた近似マッチング法の性能を示し，続いて 3 章の書誌マッチングの性能を示す．さらに，書誌マッチングに要する処理時間とパラメータチューニングに関して述べる．

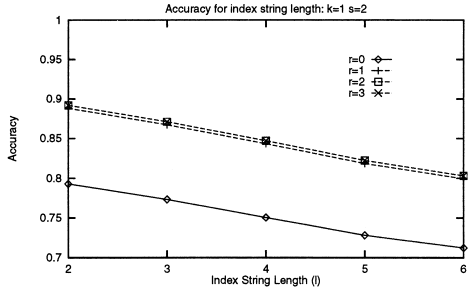
5.2 部分文字列近似マッチングの性能

図 5 は索引文字列長に対する提案手法の性能を表している．この実験では，文字の間隔 s を 2 に，また索引数 k を 1 に固定した．図 5 (a) は，検索文字列の先頭位置の幅 $r = 0, 1, 2, 3$ に対する索引文字列長と精度の関係を表している．まず，検索文字列の先頭位置の幅については，$r \geq 1$ とすることで，$r = 0$ と比較しておよそ 10%程度の精度の向上が見られる．しかし，$r > 1$ については精度についてほとんど差がない．これは，非置換誤りに対して，検索文字列の先頭位置に幅を持たせることが有効であること，およびそれほど大きな幅を設定する必要がないことを示している．

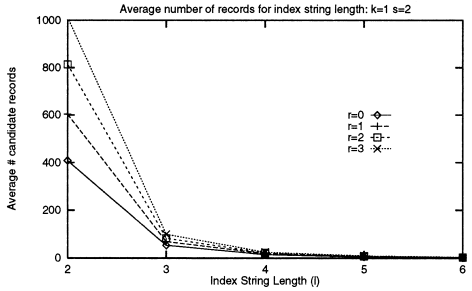
式 (2) は，$k = 1$ の場合，精度は索引文字列長の線形関数になることを示しているが，実験でも，精度が索引文字列長に対して，ほぼ線形に減少することが示されている．

図 5 (b) は，索引文字列長と候補レコード数の関係を表しており，索引文字列長 $l > 3$ のときに，候補レコードの絞り込みが可能であることが分かる．これは，4.3 節の図 4 (a) と対応している．

図 6 は索引数に対する提案手法の性能を表している．この実験では，文字の間隔 s および検索文字列の先頭位置の幅 r をともに 2 に固定し，索引文字列長 3 および 4 に対する索引数と性能の関係を測定した．

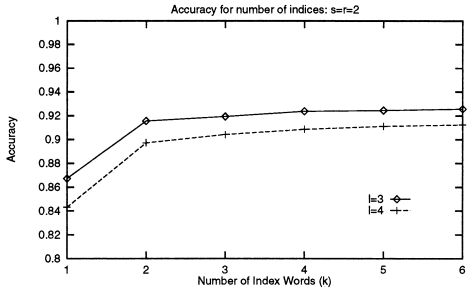


(a) 索引文字列長 vs. 精度
(a) Index word length vs. accuracy

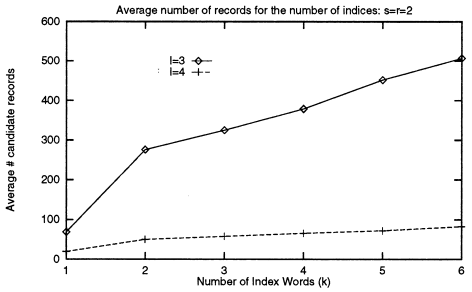


(b) 索引文字列長 vs. 平均候補レコード数
(b) Index word length vs. average candidate record size

図 5 索引文字列長と精度，平均候補レコード数の関係
Fig. 5 Relationship among index word length, accuracy, and average candaite record size.



(a) 索引数と精度
(a) Number of indices vs. accuracy



(b) 索引数と平均候補レコード数
(b) Number of indices vs. average candidate record size

図 6 索引数と精度，平均候補レコード数
Fig. 6 Relationship among number of indices, accuracy, and average candaite record size.

図 6 (a) は，索引数と精度の関係を表している．式 (2) は索引数を増加させると精度が指数関数で 1.0 に漸近することを示しているが，実際にはある一定の精度に漸近する．これは，文献タイトル文字列の抽出の失敗，書誌データベース中のタイトルの誤り，検索文字列の先頭位置の幅で扱いきれない非置換誤りなど式 (2) では考慮されていない問題に起因するものである．実験からは，索引数を 2 以上とした場合は，精度にそれほど大きな変化はなく，索引文字列長 3，索引数 6 の場合に 92.6%の精度を得られた．

図 6 (b) は，索引数と候補レコード数の平均値の関係を表している．候補レコード数は，索引数に比例して増加することが予想される．図では，索引数 2 の場合を除いて，候補レコード数は，索引数に比例して増加している．索引数 2 での増加量が大きいのは，索引数 2 で使用された索引文字列の開始位置において，索引文字列あたりの候補レコード数が他の位置の索引文字列に比べて多かったことによる．なお，この索引はタイトル末の部分文字列を用いたもので，実験に用いた書誌データでは,「‥による評価」,「‥の自動生成」

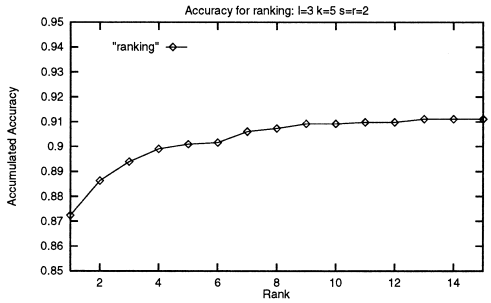


図 7 ランキングの効果
Fig. 7 Effect of ranking.

などのように論文タイトル末に一部の文字列が偏って用いられていたことによる．

5.3 書誌マッチングの性能

図 7 は書誌マッチングのランキングの性能を示している．この実験では，索引に使われる文字の間隔 s と検索文字列の先頭位置の幅 r を 2，索引数 k を 5，索引文字列長 l を 3 とした場合の結果を示している．図では，トップ m までにランクされたレコード中に正しいレコードが含まれている割合を表してい

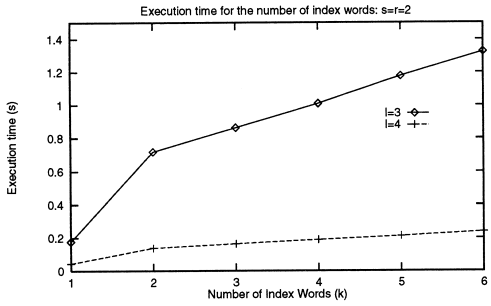


図 8 処理時間
Fig. 8 Processing time.

る．上記パラメータに対する候補レコード選択の精度は 91.24%であり，候補レコードに正解が含まれている場合の 95.6%は正解レコードがトップにランクされている．またトップ 10 にランクされた候補レコードの中に正解が含まれる割合は 99.7%となっている．このグラフから最長共通部分文字列は，書誌レコードの類似度を測定するのに有効な指標となっていることが分かる．

5.4 処理時間

本論文で提案する書誌データのマッチング法は，候補レコードの抽出と文字列の類似度計算により構成される．候補レコードの選択では検索文字列と一致する索引文字列を含むレコードを抽出することになる．この処理には B-Tree などのように効率の良いインデキシング法を用いることができ，データベースのレコード数 n に対して $O(\log n)$ で処理することができる．選択された候補レコードのランキングには，最長共通部分文字列長が用いられる．この計算には効率の良い計算アルゴリズムが数多く提案されており，文字列長 t に対して $O(t \log t)$ で求めることができる[10]．書誌データの場合，その文字列長は数十から数百程度であり，類似度の計算よりもディスク上にある書誌データを読み込む時間が支配的であると考えられる．

図 8 は，画像解析によって得られた参考文献中の書誌データのうち書誌データベース中に正解データが含まれる 1,575 件の書誌について，参考文献 1 件あたりに候補レコードの選択とランキング処理にかかった時間の平均値を表している．ここでは，文字間隔 s および先頭位置の幅 r を 2 とし，索引文字列長 l を 3 および 4 の場合の処理時間の平均値を示している．図 6 (b) の候補レコード数と比較すると分かるように候補レコード数にほぼ比例した処理時間がかかっており，処理時間の多くは候補レコードの読み込みにかかったものと予想される．なお，実験では，Ultra Sparc II（360 MHz）でメモリ 1 G バイトのコンピュータを用いた．平均処理時間は 2 秒以下であり，候補レコード数を数百件程度に絞り込めれば，実用上問題ない性能を持っていることが示されている．

5.5 パラメータチューニング

以上の実験結果から，提案手法では以下のパラメータ値が適していると考えられる．まず，検索文字列および索引文字列に用いられる文字間隔 s については，4.3 節の図 4 (b) より 2 とすれば十分であり，それ以上長くしてもあまり効果が期待できない．また，検索文字列の先頭位置の幅 r は，図 5 (a) より 1 とすれば十分であることが分かる．索引文字列長 l に関しては，2 の場合は，図 5 (b) より索引あたりの候補レコード数が約 600 となり，複数の索引を使った場合，図 8 より数秒から十数秒の処理時間が予想される．また図 4 (a) より書誌レコードの増加に比例して候補レコード数が増加することが予想されるため，$l > 2$ が望ましい．図 6 (a) より，索引文字列長 $l = 3$ で索引数 k が 3 以上が望ましい値となる．なお，パラメータチューニングにあたっては，OCR の特性や文献に含まれる言語などによって適した値が異なることが予想されるので，対象となるデータに応じたチューニングが必要となる．

6. おわりに

本論文では，学術文献電子図書館において重要な役割を果たす，文献の引用関係を抽出するための，書誌マッチング法を提案した．提案手法は，OCR の認識誤りを含む参考文献データと大規模書誌データベースのマッチングを高速に行うために，文献タイトルの部分文字列を索引として用いたところに特徴がある．

提案手法を実用規模のデータベースに適用したところ，87.2%の精度を得られることが実験的に示された（図 7）．また，マッチングの結果を人間が確認することが許されるような環境では，検索結果 10 件程度の候補を示すことで精度は約 90.9%となる．処理時間の平均値も 2 秒程度となっており，実用に耐えうる処理速度を持っていることが示された．

提案手法では，4 種類のパラメータを用いており，これらのパラメータは，OCR の認識精度，言語に依存して変える必要のあるものである．近年の OCR 技術の進歩によって，その精度は非常に高くなっているため，原画像の画質が悪くない限りは，適切なパラメータを見つけることができるものと思われる．一方，他の言語に対する有効性については，さらなる実験が必要になる．この問題は，将来の課題と考えている．

謝辞 4.3 節の式 (7) から図 4 の異なり索引文字列

数を求めるにあたっては，マックスプランク心理言語学研究所の Harald Baayen 博士のプログラムを利用しました．利用を許可してくださった同氏に感謝いたします．この研究は日本学術振興会における「高度分散情報資源活用のためのユービキタスシステムについての研究」プロジェクト（JSPS-RFTF96P00602）の支援を受けています．

参 考 文 献

1) Ayres, F.H., Huggill, J.A.W. and Yannakoudakis, E.J.: The Universal Standard Bibliographic Code (USBC): Its use for clearing, merging and controlling large databases, *Program – Automated Library and Information Systems*, Vol.22, No.2, pp.117–132 (1988).
2) Belaid, A., Anigbogu, J.C. and Chenevoy, Y.: Qualitative Analysis of Low-Level Logical Structures, *Proc. International Conference on Electronic Publishing*, pp.435–446 (1994).
3) Bunke, H. and Wang, P.S.P. (Eds.): *Handbook of Character Recognition and Document Image Analysis*, World Scientific (1997).
4) Fujisawa, H. and Marukawa, K.: Full-Text Search and Document Recognition of Japanese Text, *Proc. Symposium on Document Analysis and Information Retrieval*, pp.55–80 (1995).
5) Good, I.J. and Toulmin, G.H.: The Number of New Species, and the Increase in Population Coverage, When a Sample is Increased, *Biometrika*, Vol.43, No.1, pp.45–63 (1956).
6) Goyal, P.: Duplicate Record Identification in Bibliographic Databases, *Information Systems*, Vol.12, No.3, pp.239–242 (1987).
7) 影浦 峡：計量情報学—図書館/言語研究への応用，丸善 (2000).
8) Kahan, S., Pavlidis, T. and Baird, H.: On the Recognition of Printed Characters of any Font and Size, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.9, No.9, pp.274–288 (1987).
9) Kukich, K.: Techniques for Automtically Correcting Words in Text, *ACM Computing Surveys*, Vol.24, No.4, pp.377–439 (1992).
10) Leeuwen, J.V. (Ed.): *Handbook of Theoretical Computer Science – Algorithms and Complexity*, Elsevier (1992).
11) 太田 学，高須淳宏，安達 淳：英文曖昧検索への HMM の適用とその評価，第 59 回情報処理学会全国大会講演論文集，pp.4P–3 (1999).
12) O'Neill, E.T., Rogers, S.A. and Oskins, W.M.: Characteristics of Duplicate Records in OCLC's Online Union Catalog, *Library Resources & Technical Services*, Vol.37, No.1, pp.59–71 (1992).
13) Parmentier, F. and Belaid, A.: Bibliography References Validation Using Emergent Architecture, *Proc. IAPR International Conference on Document Analysis and Recognition*, pp.532–535 (1995).
14) Ridley, M.J.: An Expert System for Quality Control and Duplicate Detection in Bibliographic Databases, *Program – Automated Library and Information Systems*, Vol.26, No.1, pp.1–18 (1992).
15) Takasu, A.: Probabilistic Interpage Analysis for Article Extraction from Document Images, *Proc. IAPR International Conference on Pattern Recognition*, pp.932–935 (1998).

(平成 12 年 6 月 20 日受付)
(平成 12 年 9 月 28 日採録)

（担当編集委員 仲尾 由雄）

高須 淳宏（正会員）

1984 年東京大学工学部卒業．1989 年同大学大学院工学系研究科博士課程修了．工学博士．同年学術情報センター助手．学術情報センター助教授を経て現在国立情報学研究所助教授．データベースシステム，文書画像処理，機械学習に関する研究に従事．電子情報通信学会，人工知能学会，ACM 各会員．

片山 紀生（正会員）

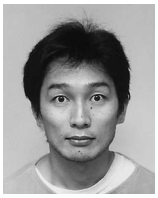
1990 年東京大学工学部卒業．1995 年同大学大学院工学系研究科博士課程修了．工学博士．同年学術情報センター助手．2000 年国立情報学研究所助手．同年同助教授，現在に至る．データベースシステムに関する研究に従事．電子情報通信学会，IEEE，ACM 各会員．

大山　敬三（正会員）

1980 年東京大学工学部卒．1985 年同大大学院工学系研究科博士課程修了．工学博士．同年東京大学文献情報センター助手，1986 年学術情報センター助手，1987 年同助教授，1998 年同教授，2000 年国立情報学研究所教授．現在に至る．文書処理システム，情報検索システム，電子図書館等の研究開発に従事．電子情報通信学会，情報メディア学会各会員．

安達　　淳（正会員）

1981 年東京大学大学院工学系研究科博士課程修了．工学博士．東京大学大型計算機センター助手，文部省学術情報センター研究開発部助教授，教授を経て現在国立情報学研究所教授．オンライン情報システム，分散処理システム，情報検索，電子図書館システム等の開発研究に従事．電子情報通信学会，IEEE，ACM 各会員．



影浦　　峡

1986 年東京大学教育学部卒業．1993 年マンチェスター大学 PhD．1996 年シェフィールド大学客員研究員．現在国立情報学研究所助教授．東京大学大学院教育学研究科助教授併任．専門用語の研究に従事．Association for Computational Linguistics，International Quantitative Linguistic Association，言語処理学会各会員．