情 報 学 基 礎 45－6
（1997. 5. 15 ）

文書画像データからの書誌情報の抽出とマッチング

高須 淳宏†、早川 公泉‡、片山 紀生†、大山 敬三†、安達 淳†

† 学術情報センター、‡ 東京大学

電子図書館は、全文データや画像データなどの大量のデータを持ち、関連のあるデータの間にリンクを張るなどのデータの加工が必要である。大量のデータに対して、このような作業を人手で行なうことはほとんど不可能で、なんらかのデータベース構築支援メカニズムが必要になる。本稿では、電子図書館において学術文献の参照関係に対応するハイパーリンクを作成するための書誌照合手法を提案する。この手法のポイントは、文書画像処理によって得られたエラーを含む書誌データに対して高速な照合を精度良く行なうところにある。提案される手法は、非連続 N-gram を用いて大規模書誌データベースから候補レコード集合を高速に選び出す近似書誌照合と最長共通部分文字列を用いた類似度に基づく候補レコードの評価の２つのステップから構成される。本稿では、提案される手法が OCR の各種の認識誤りに対応できることを示す。

Bibliographic Data Extraction from Document Images and Bibliographic Matching

Atsuhiro Takasu†, Kimimoto Hayakawa‡, Norio Katayama†
Keizo Oyama†, Jun Adachi†

† National Center for Science Information Systems,
‡ University of Tokyo

Digital libraries usually have large amount of fulltext and image data and many pieces of the data are linked each other. Since the data is too large to compile manually, database construction support mechanism is one of the key problem of digital libraries. This paper presents a method for constructing hyperlinks of references which is one of the most important links in digital libraries of academic journals. The objectives of this paper are to handle bibliographic data with errors that is obtained using a document understanding methodology and to build a fast algorithm for bibliographic matching. The presented method is two phases. It firstly finds a candidate record set from bibliographic databases by an approximate matching using non-consecutive n-grams. Secondly, it evaluates each candidate record based on the longest common substring. The paper shows the robustness of the presented approximate matching for various kinds of OCR errors.

1 はじめに

近年注目を集めている電子図書館は、各種の資料を電子化してネットワークを介して配布することによって迅速に資料を利用者のもとに届けることが可能になる。ネットワークを利用することにより物理的距離や時間の制約を軽減したり、大量情報の高速検索を可能にするなど、さまざまな可能性を持ったシステムと考えることができる。

電子図書館の実現には、著作権等の制度上の問題から技術的な問題に至るまで広範な問題が関係している。その1つに電子図書館のデータベースの構築支援の問題がある。図書館のデータベースは、これまで図書目録データが中心となっていた。しかし、電子図書館では、目録データに加え、文献の全文データや画像、動画像などのマルチメディアデータを扱う。また、ハイパーテキストの普及に伴い、各資料を有機的に結合する必要が生じてきている。その結果、データの大量化、構造の多様化、複雑化が進みデータベース構築をすべて人手で行うのはほとんど不可能になっている。

電子図書館で扱うデータには、本や雑誌などすでに印刷物として流通しているものも多く、これらの印刷物を電子化することによってデータベースの構築をかなり支援することができる。印刷物の電子化の過程で用いることのできる技術としてOCRをはじめとする文書画像処理[9]の研究が古くから行われてきた。電子図書館と関連する文書画像処理の先駆的研究として、Wang等による文書画像の領域分割手法に関する研究[18]、Nagy等による文書のレイアウト解析の研究[8]、雑誌の目次の解析システムRightPages[15]などがある。その後も文書画像のレイアウト解析に関する多くの研究がなされてきた。文書画像からのハイパーテキストの自動生成に関する研究も始められており、佐藤等はハイパーテキストの自動生成のフレームワークを提案している[13]。

本稿では、学術雑誌の文献を対象とする電子図書館において、文書画像から得られた参考文献に対する動的ハイパーリンクの自動的な生成方法を提案する。参照関係は文献間の重要な関連の1つであり、ハイパーテキスト化が有効に働く。参考文献は、分野や雑誌によってその構造がかなり異るため解析が単純でないことに加え、文書画像処理過程の認識誤りが解析をさらに難しくすることもあってこれまでわずかな研究[2), 11]しか行なわれていない。本稿で提案する手法は、電子図書館で扱われる大量データに対し高速かつ動的にハイパーリンクをつくり出すことを目的としており、文書画像データからの参考文献領域の抽出と解析、抽出された参考文献データに対する近似照合による効率的な候補書誌集合の抽出、類似度による候補書誌の評価によって構成されている。本稿では、書誌照合の手法を中心に述べる。

2 書誌照合の背景と関連研究

電子図書館で求められる参照関係のリンクは、文末の参考文献リストの個々の書誌データから文献自体へのリンクとなる。電子図書館では、新しく発行された雑誌の入力や遡及入力によってデータが逐次増加していく。特に、遡及入力されるデータに対する参照を考えた場合、リンク参照時に動的にリンクをつくり出すことが望ましい。一方、参考文献データの照合の対象としては、書誌データベースを考える。その理由は、まず第一に、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照合が必要になる。第二に、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照合を行なうことによって、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、参考文献の情報よりも詳細な書誌情報が得られる可能性がある。第三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ケースも考えられるため、書誌データベースとのリンクを考えることによって適切な文献の取り出しに柔軟に対応できる。このような書誌照合の枠組を考えた場合、データ入力時に文書画像からの参考文献領域の切り出し、OCRによるテキスト化、書誌照合のための書誌項目の抽出の処理を行ない、リンク参照時に書誌データベースと参考文献の照合処理を行なうことになる。

書誌照合の研究は、書誌データベースの重複レコードの削除問題として研究が行われてきた。O'Neill等は、OCLCのオンライン総合目録データベースからランダムに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るフィールドの傾向を調べている[10]。一方、重複レコードを自動的に発見する方法としてタイトルや著者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、これらの方法の比較を行っている[4]。またRidleyは、エキスパートシステムを用い

ることによって、重複レコード発見の精度をあげるとともに、重複レコードの中から優れたレコードを選びだすシステムを提案している[12]。

文書画像処理によって得られた参考文献の照合は、重複レコードの削除問題と比較し以下のような問題点がある。

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[6]。置換誤りは、例えば「情」を「惰」と誤って認識してしまうことを意味する。挿入および削除は、余計な文字が挿入されたり、一部の文字が認識されない誤りを示す。複合誤りは、複数の文字を1つの文字として認識してしまったり、1つの文字を複数の文字として認識するなど、複数の文字に係わる誤りを示す。このうち、置換誤り以外の誤りでは、本来の文字列と認識された文字列の間で1対1の対応がとれないため、照合処理が難しくなる。

OCRの認識誤りを含んだテキストに対するいくつかの単語照合手法が提案されている。文字クラスを用いた照合手法は、互いに誤りを起こしやすい文字のクラスを定義し、同一のクラスの文字は同じ文字として処理する方法であり、田中らはOCRを用いず画像データから直接文字クラスを取りだし照合する手法を提案している[16]。また、任意の文字の組合わせc_1とc_2に対して, c_1をc_2と誤認識する確率をあらかじめ求めておき、文字列の組に対してその類似度を確率的に求めるConfusion Matrixの手法も提案されている[5), 6]。藤澤等は、誤認識パターンに対応したオートマトンを生成して、照合処理を行う方法を提案している[3]。Myka等は、OCRを用いて得られた全文データベースに対する文字列検索問題において、これらの手法のいくつかを比較し、Confusion Matrixの精度が比較的良い結果を示すことを報告している[7]。

書誌構造の複雑さ　参考文献の記述は、雑誌や著者によってその記述がかなり異るため、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を前提とした手法が使えない。参考文献の場合、各構成要素はデリミタで区切られている。しかし、デリミタの認識誤りはその後の照合処理に大きな悪影響を与える。Belaid等は、文書画像処理によって得られた書誌データから書誌フィールドを抽出する方法を提案している[2), 11]。

表記の相違　参考文献のなかには、雑誌名の短縮形や著者の表記法の揺れなど、表記上の各種の相違がある。照合処理では、これらの表記の揺れを扱う必要がある。

処理効率　照合の対象となるデータベースは数百万件規模のデータベースであり、動的にリンクを作成するためには、全件をサーチするような処理は現実的でない。そのため、前処理によってインデックスを作成し、効率的な処理の可能な手法である必要がある。

3　書誌照合手法

大量のデータに対して高速に照合を行なうために、本稿で提案する手法は、書誌データの構成要素を抽出した後、タイトル文字列による近似照合および類似度による評価の2つのステップを経てデータベースのレコードとの照合を行なう。

3.1　書誌項目の抽出

画像データのレイアウト解析およびOCRの処理の結果得られる参考文献データに対して、Belaid等[2), 11]は、著者やタイトルなどの書誌の構成要素を精度良く抽出するために、著者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た。電子図書館のような多種類の雑誌を扱うシステムでは、このような規則を準備し維持していくことは大変な作業であり、雑誌数が多くなったときには現実的でなくなるものと思われる。そこで、本手法では、このような事前の規則をできるだけ少くすることを基本方針とした。

書誌照合の第一ステップでは、照合にタイトルのみを用いる。これは、多くの文献においてタイトルが参考文献中に含まれること、タイトルが他の項目と比較して長いためOCRの認識誤りへの対処が容易であること、著者や雑誌名のような表記の揺れが非常に少いことによる。照合の第2ステップでは、参考文献の個々の項目の比較を行なわず、参考文献中に現れる文字全体を用いて照合を行なう。従って、抽出すべき書誌項目はタイトルのみである。タイトルは、事前に用意されたタイトルのパターンとの照合によって得られる。本稿で提案する手法では、正規表現を用いてタイト

ルのパターンを表す。例えば、情報処理学会論文誌の場合、":" と"," に囲まれた部分がタイトルとなり、正規表現":.*," で表す。タイトル中に区切り記号が含まれる場合には曖昧になるが、このような場合には、すべての可能性について以下で述べる照合を試みる。

3.2 非連続 N-gram による近似照合

近似照合の目的は、参考文献と同じ書誌の候補レコード集合を高速に求めることにある。選択された候補レコードは、高い確率で正解のレコードを含むことと候補レコード数が少ないことが重要になる。本稿では、これらの精度を再現率と適合率で測ることにする。以下の議論において書誌データベースのタイトル文字列に誤りはないものと仮定する。まず、タイトル中の文字すべてを用いて照合処理を行う場合を考える。簡単のため、OCR の認識率がすべての文字において同一にαとし、誤りが各文字について独立に起こるものと仮定する。完全照合を行なう場合、データベース中のタイトルについてのインデックスを作成することによって、レコード数rに対して$O(\log r)$の照合が可能であるが、タイトル文字列の長さmに対して再現率はα^mとなる。この式は、タイトル文字列が長くなると、再現率が指数関数的に悪化することを示している。一方、再現率の高い Confusion Matrix[6] やオートマトン[3] のような近似マッチング手法では、データベース中のすべてのレコードに対して照合処理を行なう必要があるため、参考文献の照合問題に適さない。

タイトル文字列すべてを使うかわりに N-gram を使うことによって近似照合を高速に行なう手法を提案する。書誌データベース中のレコードのタイトル文字列を T、OCR によって処理されたタイトル文字列を Sとする。文字列 Sに対して、相対位置 p $(0 \leq p \leq 1)$ の文字とは、Sの $\lceil p|S| \rceil$ 番目の文字で $S[p]$ と表記する。また、相対位置のリスト $P \equiv (p_1, p_2, \cdots, p_n)$ に対して $S[P]$ は $S[p_1]S[p_2]\cdots S[p_n]$ を表すものとする。提案する手法は、あらかじめ定められた相対位置のリスト P に対して、N-gram $S[P]$ と $T[P]$ の完全照合によって近似照合を行なうというものである。連続した n 個の文字列を使わない理由は、この部分文字列が単語の一部となった場合、その単語をタイトルに含むすべての書誌が選択されてしまうために、候補レコード集合が大きくなってしまうことによる。

タイトル中の n 文字を用いることによってタイトル文字列すべてを用いるよりも再現率をあげることができるが、それでも再現率はα^nにとどまる。そこで、相対位置リストを k組用い、そのどれかで完全照合したものを候補レコードとすることによって再現率をあげることができる。

前節で述べたように OCR の誤りには挿入、削除、複合誤りのように文字列の長さを変えてしまう誤りがある。このような誤りに対応するために、OCR 処理されたタイトル文字列 Sについては、相対位置 p の隣接文字 $S[p - \frac{1}{|S|}]$ および $S[p + \frac{1}{|S|}]$ も考慮する。つまり、相対位置リスト $P \equiv (p_1, p_2, \cdots, p_n)$ に対して N-gram の集合 $S_P \equiv \{s_1 s_2 \cdots s_m \mid s_i = S[p_i - \frac{1}{|S|}]$ or $S[p_i]$ or $S[p_i + \frac{1}{|S|}]\}$ を考える。本稿で提案する書誌近似照合は、相対位置リスト $P_1, P_2, \cdots, P_k$と OCR 処理されたタイトル文字列 Sが与えられた時、以下の条件を満たす候補レコード集合を求めるものである。

> ある相対位置リスト P_i $(1 \leq i \leq k)$ に対して、$T[P_i]$ と S_{P_i}のある要素が完全に一致する。

ここで、Tはデータベース中のレコードのタイトルを表す。

まず、この手法による照合の再現率について議論する。OCR の認識誤りが独立に起き、認識率が均一であるという仮定のもとで、1つの相対位置リストで照合が失敗する確率は $1-\alpha^n$であるから、k組の相対位置リストすべてで失敗する確率は、$(1-\alpha^n)^k$となり、照合の再現率は以下の式で表される。

$$1 - (1 - \alpha^n)^k \ . \qquad (1)$$

例えば、$n = 4, k = 1, \alpha = 95\%$の時、近似マッチングの再現率は81.5%となってしまうが、$n = k = 4, \alpha = 95\%$の場合は、再現率を 99.4%まであげることができる。この議論では、認識誤りが独立におこり、認識率が均一であるという仮定をしているため、正確な解析にはならないが、N-gram の長さ n や相対位置リストの数 kに対する挙動は推測できるものと思われる。なお、n が極端に短い場合には、適合率が減少し、次節で述べる類似度の評価における計算時間が問題になる。

次に挿入、削除、複合誤りによって文字の 1:1 対応が壊れた場合の隣接文字の効果について議論す

る。もとの文字列の相対位置 p の文字が認識された文字列では位置 $f(p)$ に現れるとすると隣接文字を考慮することによって照合が成功するための必要十分条件は、相対リスト中の任意の相対位置 p に対して以下の条件が満たされることである。

$$\lceil p|S| \rceil - 1 \leq f(p) \leq \lceil p|S| \rceil + 1 \qquad (2)$$

たとえば、OCR の認識過程で削除誤りのみが起きる場合を考える。相対位置 p に対して $d(p)$ は位置 p の前で削除された文字の数を表すものとする。すると、$f(p)$ は $\lceil p|T| \rceil - d(p)$ となる。式 (2) より以下の条件が得られる。

$$\lceil p|T| \rceil - \lceil p|S| \rceil - 1 \leq d(p) \leq \lceil p|T| \rceil - \lceil p|S| \rceil + 1$$

ここで $\lceil p|S| \rceil$ が $\lceil p(|T| - d(1.0)) \rceil$ であることと ceil 関数の性質より以下の不等式が成り立つ。

$$\lceil p|T| \rceil - \lceil pd(0.1) \rceil \leq \lceil p|S| \rceil \leq \lceil p|T| \rceil - \lceil pd(0.1) \rceil + 1$$

これら 2 つの不等式から以下の不等式が成り立てば、隣接文字を考慮することによって近似照合が成功することがわかる。

$$\lceil pd(1.0) \rceil - 1 \leq d(p) \leq \lceil pd(1.0) \rceil \qquad (3)$$

式 (3) より、d 個の文字が削除された場合、タイトル文字列を d 等分し、各領域に 1 度づつ削除誤りが起きれば削除された文字を除く任意の文字を用いた照合が成功することがわかる。図 1(a) のグラフは $d(1.0) = 4$ の場合に式 (3) を満す $d(p)$ を示している。図 1(b) は、削除誤りが均等に起こる例を示している。一方、図 1(c) は最初の削除誤りが 6 番目の文字で起きたため 5 番目の文字については、式 (3) の条件を満さない。このような場合、式 (3) の条件を満さない位置から照合用の文字列が選びだされた場合に照合に失敗する。挿入誤りについても全く同様の議論が成り立つ。複合誤りの場合は、式 (3) の条件を満すことが照合に成功するための条件となる。一般に削除や挿入誤りが混在する場合、位置のずれを相殺する働きがあるため、各誤りが単独で現れるた場合よりも近似照合は成功しやすくなる。

最後にこの方法の時間計算量は、近似照合に使用される N-gram についてインデックスを作成することによって $O(\log r)$ で処理できる。隣接文字を考慮することから、各 N-gram は 3^n 個に展開されるため、時間計算量は式 (4) となる。

$$O(k3^n \log r) \qquad (4)$$

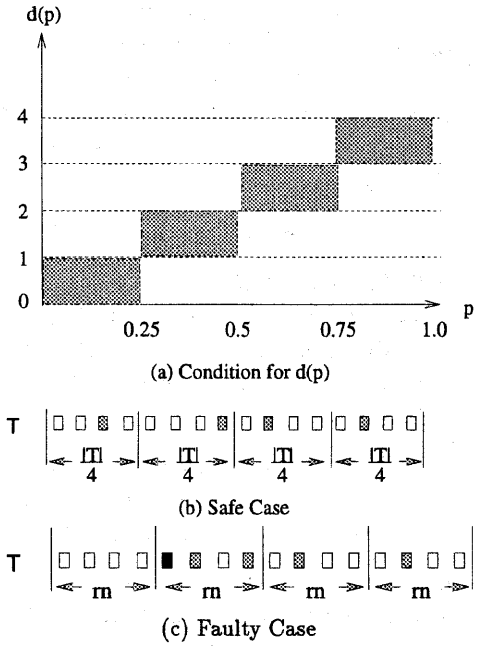

図 1: d(p) の条件と削除誤りの例

N-gram の長さについては指数関数となるが、データベースのレコード数に対して対数オーダの計算量となり、大規模書誌データベースに対しても高速に候補集合を選択できる。

3.3 類似度の計算と照合レコードの判定

書誌項目の列挙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、近似照合によって選択された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。この文字列を B_1 とし、文書画像処理によって得られる参考文献の文字列 B_2 との最長共通部分文字列[17] M を求め、$\frac{|M|}{|B_2|}$ を類似度として用いる。この類似度はおもに省略形に対する対応を考慮している。例えば「電子情報通信学会論文誌」と「信学論」との類似度は 1.0 となる。この類似度の有効性については文献[14]を参照されたい。

4 おわりに

本稿では、電子図書館において文献の参照関係に対応するハイパーリンクを動的につくり出すための書誌照合手法を提案した。この手法は、大規模書誌データベースのレコードとの照合を高速にお

こなうために、高速な近似書誌照合と類似度による評価の２つのステップより構成されている。近似書誌照合において、OCRの各種の認識誤りに対処するために非連続N-gramを用いるところに特徴があり、文字列長が変わってしまうような誤りに対しても、その誤りがほぼ均等に起これば正しく照合できることを示した。

参 考 文 献

1) F. H. Ayres *et al*: "The Universal Standard Bibligraphic Code (USBC): Its Use for Clearing, Merging and Controlling Large Databases", *Program*, Vol.22, No.2, pp.117–132, 1988.

2) A. Belaid *et al*: "Qualitative Analysis of Low-Level Logical Structures", *Intl. Conf. on Electronic Publishing*, pp.435–446, 1994.

3) H. Fujisawa and K. Marukawa: "Full-Text Search and Document Recognition of Japanese Text", In *4th Symposium on Document Analysis and Information Retrieval*, pp. 55–80, 1995.

4) P. Goyal: "Duplicate Record Identification in Bibiliographic Databases", *Information Systems*, 12(3):239–242, 1987.

5) S. Kahan *et al*: "On the Recognition of Characters of any Font Size". *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.9, No.9, pp.274–287, 1987.

6) Karen Kukich. "Techniques for Automtically Correcting Words in Text". *ACM Computing Surveys*, Vol.24, No.4, pp.377–439, 1992.

7) A. Myka and U. Guntzer: "Fuzzy Full-Text Searches in OCR Database". *Forum on Research & Technology Advances in Digital Libraries*, pp.87–100, 1995.

8) G. Nagy, *et al*: "A prototype document image analysis for technical journals", *IEEE Computer*, Vol.25, No.7, pp.10–22, 1992.

9) L. O'Gorman and R. Kasturi: *"Document Image Analysis"*, IEEE Computer Society, 1994.

10) T. O'Neill *et al*: "Characteristics of Duplicate Records in OCLC's Online Union Catalog", *Library Resources & Technical Services*, Vol.37, No.1, pp.59–71, 1992.

11) F. Parmentier and A. Belaid: "Bibliography References Validation Using Emergent Architecture", In *Intl Conf. on Document Analysis and Recognition*, pp. 532–535, 1995.

12) J. Ridley, M: "An Expert System for Quality Control and Duplicate Detection in Bibliographic Databases", *Program*, Vol.26, No.1, pp.1–18, 1992.

13) S. Satoh *et al*: "A Collaborative Supporting Method between Document Processing and Hypertext Construction", In *Intl Conf. of Document Analysis and Recognition*, pp. 533–536, 1993.

14) A. Takasu *et al*: "Approximate Matching for OCR-processed Bibliographic Data", In *Intl Conf. of Pattern Recognition*, pp. 175–179, 1996.

15) G. A. Story *et al*: "The RightPages Image-Based Electronic Library for Alerting and Browsing", *IEEE Computer*, Vol.25, No.9, pp.17–26, 1992.

16) Y. Tanaka and H. Torii: "Transmedia Machine and its Keyword Search over Image Texts". In *Proc. of RIAO 88*, pp. 248–258, 1988.

17) R. A. Wagner. "The String-to-String Correction Problem", *JACM*, Vol.21, No.1, pp.168–173, 1974.

18) K. Y. Wang *et al*: "Document Analysis System", *IBM Journal of Res. Develp.*, Vol.26, No.6, pp.647–656, 1982.