

大規模データベースを利用したリンクージシステムの提案と実装

Design and Implementation of a Linkage System with Large-Scale Databases

相澤 彰子[▼] 高久 雅生[▼] 大山 敬三[▼]
Akiko AIZAWA Msao TAKAKU Keizo OYAMA

本稿では、引用文献文字列などのテキストを大規模データベースのレコードに動的に対応付ける「リンクージシステム」の実現法を検討する。そして、(1) 入力テキストを単語列に変換するためのテキスト処理、(2) データベースに格納されたレコードからの候補検索、(3) テキストとデータベースレコードの同一性判定、の3つの処理を行うシステムを提案し、特に(2)の候補検索について、レコードに特徴的な N グラムを用いて効率的に候補を絞り込む検索法を提案する。また、実際に大規模な論文データベースを用いて開発した試作版リンクージシステムについて報告する。

This paper introduces a 'linkage system' that identifies records of large-scale databases being referred to in unformatted text such as reference lists of scientific papers. The proposed system contains the following three procedures: (1) text processing that converts input texts into sequences of words, (2) candidate selection that selects database records of possible match, (3) record identification based on the matching score between the input sequence and the candidate database record. In the paper, we particularly focus on (2) and propose an efficient searching scheme utilizing low-frequency N-grams characteristic to the target records. We also report our implementation of a prototype linkage system that is based on a real-scale bibliographic database.

1. はじめに

本稿では、「フォーマットを持たないテキスト」を、データベースに登録されたレコードに動的に対応づけるための手法を検討する。特に、書名や論文タイトルなど、比較的長い単位の文字列を属性として含む場合に焦点をあてて、大規模なデータベースに対しても効率的に照合を行う方式を提案する。具体的に想定するのは図1に例示されるような応用である。すなわち、入力テキストに対してデータベース上のレコードを参照している箇所をタグづけしたり(出力例 1)、参照されているレコードの内容を表示したり(出力例 2)する。このような対応付けは、類似の入出力インタフェースを持つ既存の検索システムと比較して、以下の点で特徴的である。

- (a) 入力テキスト全体ではなく、一部に照合するレコードが出力される。

- (b) 出力結果として得られるのは、類似レコードのランキングではなく、同一と判定されたレコードである。

たとえば図1の例では、入力テキストのうち論文を参照している箇所だけがレコードの同定に用いられる。もし入力テキストの複数箇所異なるレコードが参照されていれば、それぞれに対応するレコードの一覧が出力される。また、入力中には「情報提供サービス」を「データ提供サービス」、「No.5」を「No.3」とする誤りが存在するが、出力は正確な情報を含む唯一のレコードに絞り込まれる。本稿では、このような処理を実現するシステムを新たに「リンクージシステム」と呼び、その実現に向けた検討を行う。

以下、まず 2. でリンクージシステムの概要を述べ、特徴的な部分単語列を抽出してレコード候補を高速に数え上げるエンジンの実現法を説明する。次に 3. で、試作版リンクージシステムの実装を紹介し、簡単な性能評価を行った結果を述べる。さらに 4. で関連研究に触れ、5. でまとめを述べる。

入力:

国立情報学研究所「データ提供サービスの新たな展開」の記事が図書館雑誌 (vol.97, No.3) で紹介されていました。図書館にあるのでコピーを下さい。

出力例 1:

国立情報学研究所「データ提供サービスの新たな展開」の記事が図書館雑誌 (vol.97, No.3) で紹介されていました。図書館にあるのでコピーを下さい。

出力例 2:

ID	cinii:40005739948
URL	http://ci.nii.ac.jp/naid/40005739948/
収録誌	図書館雑誌 The Library journal. Vol.97, No.5 (2003/5) (通号 954) pp. 292~294 日本図書館協会 ISSN:03854000
書誌情報	国立情報学研究所・情報提供サービスの新たな展開(特集 大学改革と図書館) 国立情報学研究所開発事業部
所蔵図書館	...

図1 論文データベースを利用したレコード同定の例
Fig.1 Example of record identification using a bibliographic database.

2. 提案手法

2.1 提案リンクージシステムの概要

本稿で提案するリンクージシステムは、(1)入力テキストを単語列に変換するためのテキスト処理、(2)データベースに格納されたレコードからの候補検索、(3)テキストとデータベースレコードの同一性判定、の3つの機能で実現される(図2)。(1)のテキスト処理では、正規化や区切り文字による領域分割、形態素解析による分かち書きなど既存のツールを用いた処理を行う。(3)の同一性判定では、データベース分野におけるレコードリンクージの技術を適用し、編集距離に代表される文字列比較関数やSVM等の分類器を利用した判定を行う。そして(1)と(3)の間に位置する(2)の候補検索では、(1)を手がかりに照合可能性が高い候補を検索し、比較的計算コストが高い(3)の処理へと受け渡す。

ここで注意が必要なのは、(1)から(2)への出力は、入力テキストにもともと含まれる文字誤りや表記の揺らぎ、解析時の分割誤りなど多くのノイズを伴うことである。このため(2)で単純なキーワード一致による検索を適用しても、該当する候補が得られる可能性は低い。一方で、あいまい検索を適用

[▼] 正会員 国立情報学研究所/総合研究大学院大学
{aizawa, oyama}@nii.ac.jp

[◆] 情報・システム研究機構 masao@nii.ac.jp

する場合に、(3)の同一性判定は計算コストが高いことから、(2)で得られる候補レコードの数が多すぎるとシステム全体の応答性が低下する。これより本稿では、リンケージシステム実現に向けた第一のポイントとして(2)の候補検索を中心に検討を進め、可変長の N グラムをキーとする高速検索法を提案する[1]。

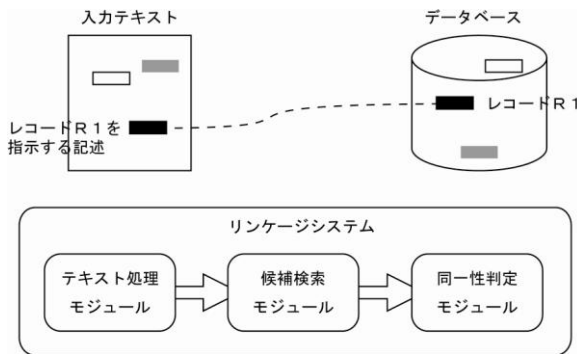


図 2 リンケージシステムの構成要素

Fig. 2 Components of the proposed linkage system.

2.2 提案手法の着眼点: 固有 N グラム

本稿で注目するのは、タイトルなど書誌情報の中に出現する特徴的なフレーズである。本稿では、このようなフレーズを「固有 N グラム」と呼び、「高々 B 個のレコードにしか含まれない任意個の連続した単語の並び」として定義する。固有 N グラムは、テキスト中でそのままの形で引用される可能性が高く、限定されたレコードだけに含まれることから単語や 2 単語の並びであるバイグラムと比較して、候補数の削減に効果的であることが期待される。

いま、レコード全体の集合を R として、任意のレコード $r_i \in R$ から得られる単語列を $w_{i1}w_{i2} \dots w_{in_i}$ とする。 n_i は単語列の長さである。この単語列の、長さ k ($1 \leq k \leq n_i$) の部分単語列 (すなわち単語 k グラム) の集合を $T_k(r_i)$ で表記する。たとえば、 $T_1(r_i) = \{w_{i1}, w_{i2}, \dots, w_{in_i}\}$ は r_i と対応づけられたユニグラムの集合、 $T_2(r_i) = \{w_{i1}w_{i2}, w_{i2}w_{i3}, \dots, w_{in_i-1}w_{in_i}\}$ は r_i と対応づけられたバイグラムの集合などである。このとき $T_k(r_i)$ の異なり要素数 $|T_k(r_i)|$ は、定義から以下となり、

$$|T_k(r_i)| \leq n_i - k + 1$$

これより r_i に含まれるすべての N グラムの数の上限値は次式で与えられる。

$$\sum_{k=1}^{n_i} |T_k(r_i)| \leq \frac{n_i(n_i + 1)}{2}$$

以下、 r_i から得られるすべての N グラムの集合を $T^*(r_i) = \cup_{1 \leq k \leq n_i} T_k(r_i)$ で表す。

さて、レコード集合 R および長さ n の任意の N グラム $W = w_1w_2 \dots w_n$ が与えられるとき、W を含む R の要素集合をレコードブロックと呼び、 $B(W)$ で表す。すなわち、

$$B(W) = \{r_i | r_i \in R \text{ かつ } W \in T_n(r_i)\}$$

である。このとき、 $B(W)$ の要素数を $|B(W)|$ として、固有 N グラムは、以下を満足する N グラムである。

$$|B(W)| \leq B$$

本稿では $|B(W)|$ を「ブロックサイズ」と呼ぶ。

候補検索では、入力テキスト I に対して、固有 N グラムを少なくとも 1 つ以上共有するレコードを選ぶ。すなわち、I から得られる単語列を $v_1v_2 \dots v_L$ として、レコードの場合と

同様に $T^*(I)$ を定めるとき、I に対する候補レコード集合 $C(I)$ は以下となる。

$$C(I) = \{r_i | T^*(r_i) \cap T^*(I) \neq \emptyset\}$$

ここで、前述のように、長さ L の単語列には高々 $L(L + 1)/2$ 個の N グラムしか含まれないから、すべてが固有 N グラムとなる場合でも、 $C(I)$ の大きさは $BL(L + 1)/2$ を超えない。すなわち提案手法では、候補数の上限値を入力テキストの長さから定めることが可能である。ただし、候補列挙の際の見落とし率の下限値を保証することはできないため、3.において、実データを用いた検証を行う。

2.3 リンケージシステムにおける同定処理手順

提案リンケージシステムにおける候補検索では、与えられた入力テキストと 1 つ以上の固有 N グラムを共有するレコードだけを候補として選択する。たとえば、「ガスタービンと教育・研究問題：ガスタービンを 100 倍面白くできるか」という論文タイトルに対して、「教育・研究問題」や「ガスタービン」などの単語列は多くのレコードに共通して出現するため候補検索には用いられないが、「ガスタービンを 100 倍面白くできるか」は特徴的なので固有 N グラムの条件を満足し、候補検索に用いられることになる (図 3)。以下、リンケージシステムにおける処理の流れを簡単にまとめる。

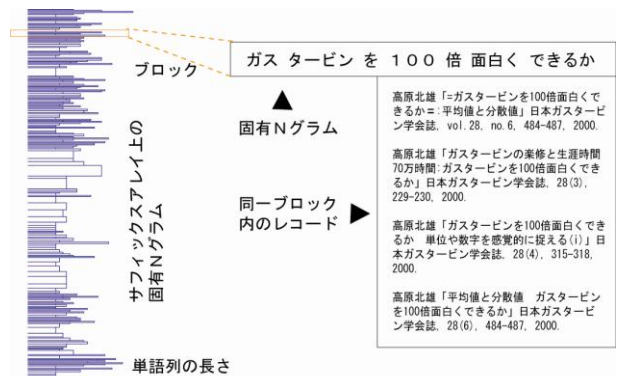


図 3 サフィックスアレイ上の固有 N グラムとブロック
Fig.3 Example of a Characteristic N-gram and a record block on a suffix array structure.

(1) データベースレコードの読み込み

まず、データベースに格納されたレコードを、「著者、タイトル、雑誌名、巻号、ページ、年」などの一般的な書式にしたがって、テキストに変換する。次に、和文テキストの場合には形態素解析ツールにより単語分割し、英文テキストの場合には空白文字を単語区切りとして、これを単語列に変換する。最後に、すべての単語列を対象としてサフィックスアレイを構築する。サフィックスアレイは、レコード集合中に含まれるすべての部分単語列の辞書順ソートとなっている。先頭 k 単語が一致する単語列は互いに隣り合うことから、任意のレコードブロックは、サフィックスアレイ上では連続する領域に対応する。固有 N グラムは B 個以下のレコードより構成されるブロックとして簡単に識別できる。

(2) 候補レコードの検索

まず、レコードと同様に形態素解析ツールや空白文字を使って入力テキストを単語列 v_1v_2, \dots, v_L に変換する。次に $v_1v_2, \dots, v_L, v_2, \dots, v_L$ のように先頭位置をずらしながらサフィックスアレイのバイナリサーチを行い、入力列に一致する N

グラムが、固有 N グラムのブロックサイズの条件（現在の実装でデフォルトは $B=50$ ）を満足するかどうかをチェックする。固有 N グラムが見つかったら、対応するブロック中のレコードを候補に追加する。

(3) 候補レコードの絞り込みと同定

まず、固有 N グラムにより得られた候補レコードに単純な類似度関数を適用して候補を絞り込む。具体的には、入力テキスト I と候補レコード r_i を構成する 2 つの単語列の間で DP マッチングを適用して、順番を考慮して両者で一致する単語 $\{u_1, \dots, u_m\}$ を求める。ここで m は一致語数である。次に、単語 $u_k (1 \leq k \leq m)$ のレコード頻度を $|B(u_k)|$ 、総レコード数を $|R|$ として、一致スコア

$$\text{score}(I, r_i) = \sum_{k=1}^m \left(-\log \frac{|B(u_k)|}{|R|} \right)$$

を計算し、上位から指定された件数のレコードを求める（デフォルトは 5 件）。最後に、これらの候補について、著者や出版年など属性を考慮した同定関数を適用して、同一と判定されたレコードを結果として返す。データベースに重複がなければ最終的に出力されるレコードは原則 1 件だけである。

3. リンケージシステムの実装と動作例

3.1 実験の概要

固有 N グラムによる検索は、「論文には意識的に他と区別できるような文字列が割り当てられているはず」という仮説に基づくもので、固有 N グラムの条件を満足しない N グラムはすべて無視されることになる。しかし、すべてのレコードが固有 N グラムを含むことは理論的には保証されないし、入力テキスト中のノイズの影響で失われる固有 N グラムも存在する。そこで、固有 N グラムの有効性を調べるために、プロトタイプシステムを実装して簡単な動作検証を行った。このプロトタイプシステムは、国立情報学研究所および協力機関がサービスとして提供する学術文献データベース [2] から約 2100 万件の書誌レコードを抽出して読込んだもので、簡単な HTML のフォームを使ってテキストをカット&ペーストすることで、オンラインで検索結果を表示できる。表示に要する時間は現在の実装で約 0.5~1 秒である。

3.2 ノイズを含む入力

実験では情報処理学会論文誌の引用文献を OCR で自動認識した引用文字列について [3]、リンケージシステムによる同定の正解率を調べた。まず、引用文献を出典が偏らないようにサンプリングしながら、各々についてデータベース中に正解が存在するかどうかを手手により調べた。具体的には公開されている CiNii 検索インタフェースを使って作業者が雑誌名や著者を組み合わせながら対応する文献を検索し、存在が確認できたものを 200 件選んだ。これらの引用文字列には、著者自身の入力時の誤りに加え、OCR の読取り誤りが存在する。200 件すべてを手手によりチェックをしたところ、全体の 55%にあたる 110 件に何らかの形の誤りが存在し、その内訳は、OCR の読取り誤り 46 件、出版年・著者名・ページ番号等の情報の誤り 44 件、タイトルやページ番号等の欠落 24 件（誤りの分類は重複を含む）であった。

次に、選択した 200 件をリンケージシステムに入力し、同一であると判定されたレコードの ID とあらかじめ人手で調べた正解レコードの ID を照合して、正解率を調べた。もし引用文字列に固有 N グラムが 1 つも含まれなければ、正解レコードは検索対象とはならないため、同定結果は不正解と

なる。逆に、同定結果が正解であれば、入力に固有 N グラムが含まれたことがわかる。ここで、人手で正解を調べる場合には、キーワードの組合せや文字誤り等を適宜修正しながら対話的に検索を進め正解を拾うが、リンケージシステムが受け取るのは、表 1 のような誤りを含む入力テキスト全体であり、それ以上の手がかりは与えられない。

表 1 書誌レコード同定の例
Table 1 Examples of identification results.

(a) 正解の例	
【入力テキスト】	4)前出,塩「吉井:正規化主成分特:微量を利用した物体抽出法とを定量的評価」,電子情報通信学会・論文誌,VgLJ 亨 5・D・II:・No.110, PP・1660-1672,(1992).
【出力レコード】	author="前田,英作 塩,昭夫 石井,健一郎", title="正規化主成分特微量を利用した物体抽出法とその定量的評価", journal="電子情報通信学会論文誌・D-II, 情報・システム, II-情報処理", ISSN="09151923", publisher="電子情報通信学会情報・システムソサイエティ", year="19921025", volume="75", number="10", pages="1660-1672"
【正解レコード】	出力レコードに一致
(b) 不正解の例	
【入力テキスト】	3)末田ほか:画像処理エキスパートシステム,東芝レビュー,Vol.40, No.5, pp.403-406(1985).
【出力レコード】	該当なし
【正解レコード】	author="末田,直道 and 三亀,和雄 and 片桐,政雄", title="画像処理エキスパートシステム (システム技術<特集>)", journal="東芝レビュー", ISSN="03720462", publisher="東芝技術企画室", year="1985/04", volume="40", number="5", pages="p403-406"

3.3 実験結果

リンケージシステムで同定を行った結果、200 件中 196 件で正しいレコードが出力された（正解率は 98%）。不正解であった 4 件のうち 3 件には入力に誤りが含まれており、また、4 件すべてについて論文タイトルに長音文字が含まれた。調べてみると、歴史的経緯から検索対象としたデータベースには長音文字が半角ハイフンとして記録されているレコードが存在することがわかった。特に表 1 (b) のように、論文タイトルが比較的短い場合に、長音と半角のミスマッチ（「エキスパートシステム」の長音が「エキスパートシステム」の半角ハイフン）により固有 N グラムが失われてしまうことが見逃しの原因であった。ためしに、長音をあらかじめ半角に変換するヒューリスティックな正規化ルールを組み込むと、不正解であった 4 件を含め、200 件すべてが正解となった。このような例外処理は必ずしも長音に限られるものではないが、確信度の高い同定結果から自動的に変換ルールを獲得する手だてがあれば、正解率の向上に寄与すると考えられる。

4. 関連研究

データベース分野においては、従来から効率的な重複レコード抽出のための手法が研究され [4]、伝統的には、特定のフィールドに注目して、「苗字の先頭 4 文字」などをキーとしてレコードをソート、同じキーの値を持つレコードどうしを照合の候補とする方法等が用いられてきた [5]。近年では、与

えられたレコードペアに対して、類似度関数の計算値が閾値以上となるかどうかを高速に判定するシグニチャ関数方式 [6] や選択的インデックス方式 [7] も提案されている。

ここで [6][7] の手法は、Jaccard 係数や編集距離などによる類似度の値が、与えられた閾値以上となる候補をすべて数え上げることを目的とするものである。しかしながら冒頭で述べたように、本稿では、入力テキストが複数のレコードを参照する場合に、その各々に対して同定を行うことを想定している。レコードはテキストの一部だけに一致するため、Jaccard 係数などの単純な類似度関数は必ずしも意味を持たない。このような応用において、本稿で提案した固有 N グラムは、入力テキスト長によらず一致候補レコードを高速に求めることが可能であるという利点を持つ。たとえば本稿のプロトタイプシステムにおいて、実験で用いた引用文字列を 5 件、10 件、20 件とまとめて入力した場合でも、各々に対する正解候補が検索できることを確認している。

一方で近年、特に情報検索分野からのアプローチとして、類似度ランキングに基づく候補選択法が提案されている。その代表例として、文字バイグラム的一致数をカウントする bigram indexing [8] や情報検索分野における tf-idf 重みを用いた canopy clustering [9][10] がある。

これらの手法では、入力テキストとの間で多くの N グラムを共有するレコードを候補として抽出する。「ブロックサイズ」が大きい N グラムを手がかりとするため、大規模なデータベースでは検索途中で読込まれる候補数が増加する。一方、提案手法では、ブロックサイズが固有 N グラムの閾値 B を超える N グラムはすべて無視するため、途中段階で考慮される候補の数は少ない。候補数を絞ることによる見逃しの可能性はゼロではないが、本稿では実際のデータを用いた実験により、入力テキストがある程度の長さを持つ引用文字列のような場合には、見逃し率は問題になるほどではなく、固有 N グラムによる検索が有効に働くことを示した。

5. おわりに

本稿では、フォーマットを持たないテキスト入力をデータベースのレコードに対応させるためのリンケージシステムの構成について述べ、特に低頻度の N グラムを活用して効率的に候補検索を行う方法について議論を行った。

文献 [11] では、大量の文書からなるテキストレポジトリと関係データベースを結付ける情報統合手法を提案している。このアプローチでは、テキスト文書からレコード抽出モジュールを用いていったん仮想的なデータベースを構築した上で、類似結合を適用する。これに対して提案手法は、まずレコードをテキスト形式に変換して入力文書を問い合わせ文として候補検索を行った上で、次に候補レコードの属性情報を手がかりに同一性の判定を行う。両者は相補的な面もあり、対象とするレコードや文書の性質に応じて使い分けたり、ブートストラッピング的に用いたりすることが考えられる。

なお、現在のリンケージシステムは、固有 N グラムで候補選別をした後に、同じデータ構造の上で DP マッチングを行う形で実装している。システムからは DP マッチング適用後のランキングだけが出力されるため、候補選択法に関する関連研究との性能比較は今回の実験には含めていない。今後は、評価項目や方法を検討しながら比較による評価を行いたい。また実験では、ある程度のタイトル長を含む引用文字列について良好な結果を得たが、タイトルが省略された場合の対処法についても今後の検討課題である。

[文献]

- [1] A. Aizawa and K. Oyama: "A Fast Linkage Detection Scheme for Multi-Source Information Integration," Proc. of the Int'l Workshop on Challenges in Web Information Retrieval and Integration, pp. 31-40 (2005).
- [2] 国立情報学研究所「CiNii Home (NII 引用文献情報ナビゲータ)」, <http://ci.nii.ac.jp/> (2004).
- [3] A. Takasu: "Bibliographic Attribute Extraction from Erroneous References Based on a Statistical Model," Proc. of ACM & IEEE Joint Conference on Digital Libraries, pp. 49-60, 2003. (2003).
- [4] 相澤彰子, 大山敬三, 高須淳宏, 安達淳: 「レコード同定問題に関する研究の課題と現状」, 電子情報通信学会論文誌, DI, Vol.J88-DI, No.3, pp. 576-589 (2005).
- [5] M. A. Jaro: "Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," J. of the American Statistical Society, 84 (406), pp. 414-420 (1989).
- [6] A. Arasu, V. Gant, and R. Kaushik: "Efficient Exact Set-Similarity Joins," Proc. of the 32nd Int'l Conf. on Very Large Data Bases, pp. 918-929 (2006).
- [7] R. J. Bayardo, Y. Ma, and R. Srikant: "Scaling Up All Pairs Similarity Search," Proc. of the 16th Int'l Conf. on World Wide Web, pp. 131-140 (2007).
- [8] P. Christen and T. Churches: "Febri - Freely Extensible Biomedical Record Linkage," Computer Science Technical Reports, TR-CS-02-05, Australian National University (2002).
- [9] A. McCallum, K. Nigam and L. H. Ungar: "Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching," Proc. of The Sixth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 169-178 (2000).
- [10] W. W. Cohen and J. Richman: "Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integration," Proc. of The Eighth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 475-480 (2002).
- [11] 張建偉, 石川佳治, 北川博之: 「トピックを考慮した大規模文書情報源からのレコード抽出」, 情報処理学会論文誌: データベース, Vol.48, No. SIG 14 (TOD 35), pp. 107-123 (2007).

相澤 彰子 Akiko AIZAWA

国立情報学研究所／総合研究大学院大学教授。言語テキストの処理を中心とした研究に従事。日本データベース学会、情報処理学会等会員。

高久 雅生 Masao TAKAKU

情報・システム研究機構 新領域融合研究センター プロジェクト研究員。情報検索、電子図書館の研究に従事。情報処理学会、情報知識学会等会員。

大山 敬三 Keizo OYAMA

国立情報学研究所／総合研究大学院大学教授。Web 情報アクセス技術を中心とした研究に従事。日本データベース学会、情報処理学会等会員。