レビュー

異種データベース間でのレコード照合に関する研究動向

Record Linkage of Multi-source Databases: Research Trends

相澤 彰子
国立情報学研究所
Akiko AIZAWA
National Institute of Informatics
高須 淳宏
国立情報学研究所
Atsuhiro TAKASU
National Institute of Informatics
大山 敬三
国立情報学研究所
Keizo OYAMA
National Institute of Informatics
安達 淳
国立情報学研究所
Jun ADACHI
National Institute of Informatics

要旨

異なるデータベースの統合においては，互いに重複するレコードを検出し排除することが必須である．しかしながら，このレコード間の照合は一般にコストのかかる困難な作業となる．というのも多くの場合，データベースどうしは共通のレコード識別子を持たず，レコードの属性や値どうしにも厳密な対応関係が存在しないためである．特に，長期間にわたり分散化した環境のもとで構築されてきた大規模なデータベースにおいて，信頼性の高い照合判定を実現することは容易ではない．また近年では半構造化データの扱いという新たな課題も出現している．そこで本論文では，重複レコードの検出と削除を行うためのデータクリーニング技術について概観する．

ABSTRACT

Detecting and eliminating duplicate records is crucial in integrating multiple source databases. However, the task of identifying linkages between records is often costly and hard due to the lack of common record identifiers, and also the variations of notations of attributes and values with no explicit correspondences. Specifically, when dealing with large databases with long and distributed maintenance histories, highly reliable record linkage is difficult. There has also been a new and emerging aspect of the problem that is the manipulation of semi-structured data. Based on the background, we present an overview of data cleaning techniques for detecting and eliminating duplicate records in this paper.

[キーワード]

レコード照合, **重複排除**, **データクリーニング**, **ブロッキング**, Fellegi-Sunter **モデル**

[Keywords]

Record Linkage, Deduplication, Data Cleaning, Blocking, Fellegi-Sunter Model

1 はじめに

異なる情報源の間で共通するレコードを照合する問題は歴史が古く，すでに 1959 年には Newcombe らにより，計算機による “record linkage” の自動化に関する論文が発表されている[1]．また同じ文献によれば，“record linkage” という言葉はさらに古く（この場合は人手による同定処理を指す），その起源は 1940 年代後半の Dunn や Marshall による文献までさかのぼることができる[2][3]．

当初 “record linkage” は，ばらばらに記録された “vital records”（出生・死亡・婚姻・離婚などの記録）を用いて様々な統計分析を行うための前処理として認識されていた．しばらくは主に疫学調査の分野で研究が行われていたが，その後，長い歴史の中で次第に適用範囲を広げ，様々な方面から研究されることになった．このことは，同種の問題が今日，論文ごとに異なる呼び名で呼ばれていることにも反映されている．参考のためその例を列挙すると，record matching, data cleaning, data cleansing, data scrubbing, entity reconciliation, entity identification, merge/purge problem, duplication check, duplication identification, duplication elimination, deduplication, hardening soft databases, reference matching, object consolidation, named entity co-reference determination 等である．本論文では以下，これらを総称して「レコード照合問題」と呼ぶことにする．[1]

レコード照合問題の代表的な適用例として,「人物の照合」および「書誌の照合」の２つをあげることができる.「人物の照合」とは，具体的には異なる２つの病院の患者記録の対応をとる等であり，医療データや国勢調査等の分析に欠かせない技術として，主に統計処理的な立場から研究されてきた．一方,「書誌の照合」とは，例えば文献データベース中の重複エントリの検出等であり，オンラインカタログの品質維持のため，主に図書館情報学の立場から検討されてきた．いずれの場合についても，データベース構築が長期にわたり分散した環境のもとで行われるため，レコード照合による品質の管理が必須となることが背景にあったと考えられる．

一方，データベース分野においては，レコード照合問題はデータクリーニング技術の１つとして認識されてきた[4]．特に近年では，データウェアハウス構築や Web マイニングといった話題の盛り上がりを受けて，タグ付きのいわゆる半構造化データの統合を目的とするクリーニング技術が注目を集めている．直感的にいえば,「Web 上の人物情報の照合」や「Web からの書誌情報の収集」といった身近なタスクにまで，レコード照合問題の範囲が広がったのである．また，従来の統計的なアプローチに加えて，機械学習や情報検索分野における研究成果の適用が試みられはじめたことも，近年の顕著な動きとして見逃せない．

ここで，レコード照合問題を考える上で特に注意が必要となる点を２つ述べる．第一は，値の欠落や誤りの多いデータを扱う場合には，レコードの誤り修正とレコードの照合が，しばしば不可分な処理となることである．すなわち，レコード照合の機能は単独のモジュールとして存在するのではなく，値の整合性チェックや異種データベース間でのスキーマ変換等を含む，より広範なデータクリーニング技術との連携ではじめて実現される．注意点の第二は，レコード照合問題では処理の品質を高い水準に保つことが最優先されることである．機械学習や情報検索では，領域に依存しない手法を用いて誤りを最小にするベストエフォート型システムの構成が最終的な目標となるが，レコード照合問題では品質を保証するために，領域知識の実装や維持管理，人手による判定処理までが必要となる．すなわちレコード照合問題の本質は，単なる学習や知識獲得アルゴリズムの適用なのではなく，現実世界のデータを前提とした総合的なシステム設計なのである．

本論文では，上記の背景を踏まえて，レコード照合問題に関するこれまでの研究動向を概観する．以下，2 節でレコード照合問題の分類を示し，3 節で一般的なモデルを紹介する．4 節では，レコード照合問題の要素技術について述べ，代表的な手法を簡単に紹介する．さらに 5 節で役に立つサーベイ論文を紹介し，最後に 6 節でまとめを述べる．

2 レコード照合問題の分類

まず，本論文における「レコード」の定義について述べる．データベース分野におけるレコードとは，データベースに蓄積される情報の単位であり，ある論理的な意味付けのもとに，ひとまとまりにされたデータを指すが[5]，本論文におけるレコードとは,「あらかじめ定められた属性を持ち，実世界上の特定のエンティティを明示的あるいは非明示的に参照するタグづけされた情報」であるとする．すなわち，レコードは広義に以下の２つを含む．

(1) データベース上のレコード

[1] レコード照合問題はすでに述べたように長い伝統と歴史を持つ分野であるが，我が国における研究発表はあまり多くない．和訳の存在しない用語も多く見受けられるため，本文中では技術用語については英文表記を原則として，必要に応じて括弧で和訳を示すことにした．

(2) 半構造化データ

ただしここでの「半構造化データ」とは，一定書式のテキスト等から抽出した情報であり，不特定多数の情報発信者が独立に記述する場合等を想定している．たとえば，(1) が顧客 DB，(2) が個人のアドレス帳に対応する，あるいは，(1) が書誌データベース，(2) が著者による引用文献リストに対応するなどである．

ここで，レコード照合という問題設定のもとでは，(1) と (2) の間には質的な違いが存在する．すなわち，(1) データベース上のレコードは，データベース内で一意に付与されるレコード識別子によって，明示的に現実世界のエンティティを参照している．レコードは，予め定義されたデータモデルにしたがって登録されており，属性値が欠落することはまれである．また，各々のデータベースは，エンティティとの対応が１対１となるよう設計されていることから，誤りが混入する場合を除き原則として重複レコードが存在することはない．一方，(2) 半構造化データは多くの場合，一意性が保証された識別子を持たず，エンティティの参照は非明示的である．半構造化データは，属性値の欠落や誤り等を含む可能性が高い．また，同一エンティティを参照するレコードが多数存在することは，重要な情報は多くの人から参照されるという原則に照らして，むしろ当然であるといえる．

上記を前提として本論文では，レコード同定問題を以下の３つのタイプに分類する（図 1）．

(A) レコードの重複排除

単一あるいは複数のデータベース間で重複するレコードを抽出して，ひとつにまとめる問題.

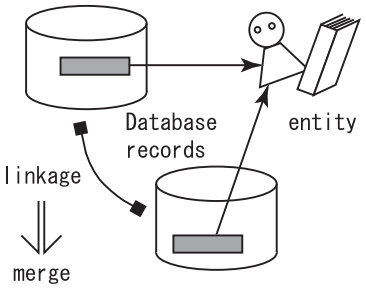
(B) レコード参照先の同定

半構造化データの参照先を，指定されたデータベース中の登録レコードの中から見つける問題.

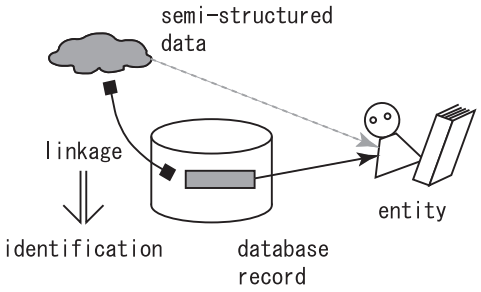
(C) レコード共参照関係の分析

同一のエンティティを参照する半構造化データどうしを，１つのグループにまとめる問題.

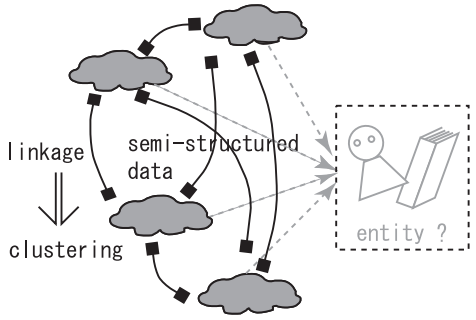
(A) はレコードの統合（merge），(B) はレコードの同定（identification），(C) はレコードのクラスタリング（clustering）にかかわるものである．単一データベース上ではレコードの重複はないものとすると，考慮するべきレコード間の対応関係は，(A) では１対１，(B) では１対多，(C) では多対多となる．



(A) レコードの重複排除



(B) レコード参照先の同定



(C) レコード共参照関係の分析

図 1: レコード照合問題の分類

3 レコード照合問題のモデル

今日，多くの研究で参照されている代表的なレコード照合問題のモデルは，1969 年の Fellegi & Sunter の文献[6] によるものである．以下では，最新の文献[7][8]の記法にしたがって，Fellegi-Sunter の確率モデルを紹介し，近年における機械学習の適用との関係について触れる．

3.1 Fellegi-Sunter モデル

まず，照合の対象となる２つの情報源 A, B を考える．A が n_a 個のレコードを，B が n_b 個のレコードを含むとする．さて，B のすべてのレコードが A のすべてのレコードの照合候補だとすると，match（照合）／ non-match（非照合）の決定が必要なレコード対は $n_a \times n_b$

個存在する．[2] これに基づき，A と B の直積 $A \times B$ を 2 つの排他的な集合 M, U に分割し，$A \times B$ の要素は，これらが照合関係にあれば M，不照合であれば U に属するものとする．ここでのレコード照合の目的は，$A \times B$ の要素に対して，以下のいずれかのラベルを付与することである．

A_1： match（照合）
A_2： possible match（照合可能性あり）
A_3： non-match（非照合）

レコード照合における一般的な考え方は，A_1 および A_3 を自動判定し，A_2 の場合については人手による判定を行うというものである．

さて，A，B の要素を a（$\in A$），b（$\in B$），各々に関する登録情報を $\alpha(a)$，$\beta(b)$ と表記するとき，候補対 $(\alpha(a), \beta(b))$ の一致の度合いを示すベクトル γ（$\in \Gamma$）を agreement vector（一致ベクトル）と呼ぶ．たとえば，フィールドごとに一致度を計算する場合には，フィールド数を k として，γ は k 次元ベクトルとなる．ここで，与えられたレコード対で照合が成立する場合に，agreement vector の値が γ となる確率を $m(\gamma)$ とする．すなわち，

$$m(\gamma) = P(\gamma | (a, b) \in M) \tag{1}$$

とする．同様にレコード対で照合が成立しない場合に，agreement vector の値が γ となる確率を $u(\gamma)$ とする．すなわち，

$$u(\gamma) = P(\gamma | (a, b) \in U) \tag{2}$$

とする．

このとき，A_1，A_2，A_3 のカテゴリ付与の誤りとして，次の 2 種類が存在する．

- False matches (Type I errors)

  本来異なるレコードを誤って同一のものとみなしてしまう誤り

- False non-matches (Type II errors)

  本来同一のレコードを誤って異なるものとみなしてしまう誤り

[2]実際には，A と B の対応関係は任意ではない場合が多い．たとえば B の要素が互いに異なるものである場合には，A の要素は B の複数の要素に同時に対応することはない．Cohen らは，このような問題を constrained matching problem と呼んでいる[9]．また，Gu らはこのような問題を 1-1 record linkage，その他の場合を 1-many record linkage と呼んでいる[8]．これは，本論文の図 1 で示したレコード照合問題の分類 (A)，(B) にそれぞれ対応すると考えられる．

ここで，Γ をすべての可能な agreement vector の集合とすると，両者の確率は次式となる．

$$P(A_1|U) = \sum_{\gamma \in \Gamma} u(\gamma) P(A_1|\gamma) \tag{3}$$

$$P(A_3|M) = \sum_{\gamma \in \Gamma} m(\gamma) P(A_3|\gamma) \tag{4}$$

Fellegi-Sunter モデルによる最適な照合戦略とは，false match の確率 $\mu = P(A_1|U)$ および false non-match の確率 $\lambda = P(A_3|M)$ が決められたとき，人手判定を必要とするレコード対 A_2 の数を最小にするものである．具体的には，$|\Gamma| = N_\Gamma$ として N_Γ 個の $\frac{m(\gamma)}{u(\gamma)}$ の値を降順に並べ，$\sum_{i=1}^{n} u(\gamma_i) = \mu$，$\sum_{i=n'}^{N_\Gamma} m(\gamma_i) = \lambda$ となるよう n，n'（$n < n'$）を決める．すると，n, n' における値 $T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$，$T_\lambda = \frac{m(\gamma_{n'})}{u(\gamma_{n'})}$ を判定の境界値として，最適戦略は以下のようになる．

$$\begin{aligned} (a, b) \quad &\in \quad A_1 \ if \ T_\mu \leq \frac{m(\gamma)}{u(\gamma)} \\ &\in \quad A_2 \ if \ T_\lambda < \frac{m(\gamma)}{u(\gamma)} < T_\mu \\ &\in \quad A_3 \ if \ \frac{m(\gamma)}{u(\gamma)} \leq T_\lambda \end{aligned} \tag{5}$$

ここで，各フィールドの値が互いに独立である場合には，$log\frac{m(\gamma)}{u(\gamma)} = \sum_{j=1}^{k} log\frac{m(\gamma^j)}{u(\gamma^j)}$ となる．図 2 は Fellegi-Sunter モデルによる A_1，A_2，A_3 の 3 つのラベルと 2 つのタイプの判定誤りを図示したものである[11]．

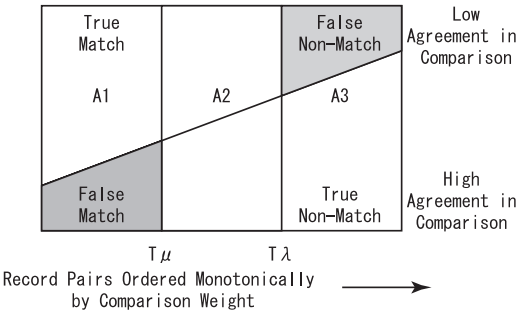


図 2: Fellegi-Sunter モデルにおける判定ラベルと判定誤り[11]

上記の枠組のもと，レコード照合問題は式 (1) (2) の $m(\gamma)$ および $u(\gamma)$（あるいは対数尤度比 $log\frac{m(\gamma)}{u(\gamma)}$）を推定する問題となる．当初の Fellegi-Sunter モデルでは，この値を (i) あらかじめ既知である，(ii) 判定済の事例における観察値に設定する，のいずれかとしていた[6]．その後 Jaro らは，各フィールドの独立性を仮定した上で，EM アルゴリズムを適用して対数尤度比（$log\frac{m(\gamma^j)}{u(\gamma^j)}$）を推定する方法を示した[10]．近年では Verkios らが，ラン

ダムに得られる観測値に基づき先験的分布を更新するベイズ決定理論の枠組みを示している[11].

3.2 Fellegi-Sunter モデルと機械学習の関係

近年では機械学習を直接適用して match（照合） / non-match（非照合）の２値分類問題を解く場合もあり，高い識別能力が報告されている[12][13]．機械学習自体は possible match（照合可能性あり）を判定する能力を持たないが，図 2 において，対数尤度比 $log\frac{m(\gamma)}{u(\gamma)}$ を機械学習によるスコア値に置き換えると，Fellegi-Sunter モデルと類似の判定方法が適用できる．

たとえば図 3 は，実際の書誌同定タスクにおけるスコアの分布を示したものである．横軸にはスコア値を，縦軸にはスコア値を 1 区切りでヒストグラム化した場合のレコード対の数を対数目盛りで示している．実線が照合する事例，点線が非照合事例である（見やすさのため縦軸を対数目盛りで示しているが，実際には照合，非照合の両者では後者の数が圧倒的に多い）．いま，人手判定の閾値 T_μ，T_λ が与えられたとすると，Fellegi-Sunter モデルの場合と同様に，閾値判定ラベルと誤りタイプが図中に示したように定まる．

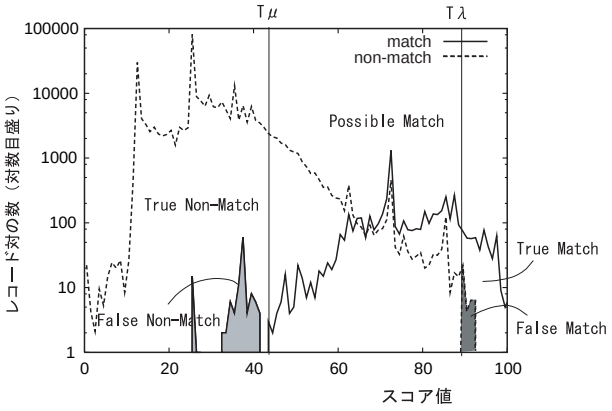


図 3: 実問題における照合スコアの分布の例

ただし，サポートベクタマシンなどの機械学習を適用する場合には，スコア値が確率的な解釈を持たないため，誤り確率 μ，λ をスコア値から推定することができない．このため人手判定の閾値 T_μ，T_λ については，別途経験的に設定することが必要である．これは具体的には，図 3 において，Fellegi-Sunter モデルによる agreement vector をヒストグラムの１区間に対応させた場合に，区間の並び方が必ずしも match / non-match の対数尤度比の順番にはならないことを意味している．

4 レコード照合のための要素技術

レコード照合における技術的なポイントは，(1) 大規模なデータを扱うため高い効率性が求められること，および (2) データベース品質維持のため高い信頼性が求められること，の２点である．しかしながら両者はトレードオフ関係にあり，すべてのレコード対を単純に比較するだけでは (1)(2) を同時に満足することはできない．このためレコード照合システムでは，候補選別，自動判定，人手による判定などを含む多段階の処理が必要となる．以下では，レコード照合システムの要素技術をまとめ，代表的な手法を簡単に紹介する．

4.1 レコード照合システムにおける処理要素

レコード照合システムの設計は研究毎に様々であり，文献により用語も異なるのが現状である．たとえば，レコード照合システムに必要な処理要素として Elfeky[31] らは，Standardization（正規化），Blocking/Searching（ブロック化/探索），Comparison（比較），Decision Model（判定モデル），Measurement（評価指標）の５つを示している．また Gu ら[8]はこれらに Database Management System（DBMS）および Graphical User Interface（GUI）を加えた 7 層モデルを提案している．また，Lee ら[14]はデータクリーニングにおいて必要となる操作を pre-processing，processing，validation and verification の３段階に分け，それぞれで実行する処理として Abnormalities Detection（不整合データの検出），Automatic Merge/Purge（自動統合/削除），User Manipulation（人手による判定）等をあげている．

本論文では，近年の動きであるテキストや半構造データからのデータ抽出，判定済のデータを訓練用事例とする知識獲得等も踏まえ，レコード照合システムの標準的な処理の流れを図 4 のようにまとめる．以下，それぞれの処理機能について簡単に説明する．

(1) セグメンテーション（segmentation）

入力テキストを解析して，レコード毎のフィールド値を抽出しデータベースにロードする．(例: “1999.8” を “year⇒1999, month⇒8” など）

(2) 正規化（normalization）

辞書や変換ルールを用いて，入力時の表記の揺れや用語の不統一を修正する．(例： “1999 年” と “’98” など）

(3) 選別（selection）

効率のよい方法を用いて，互いに重複する可能性

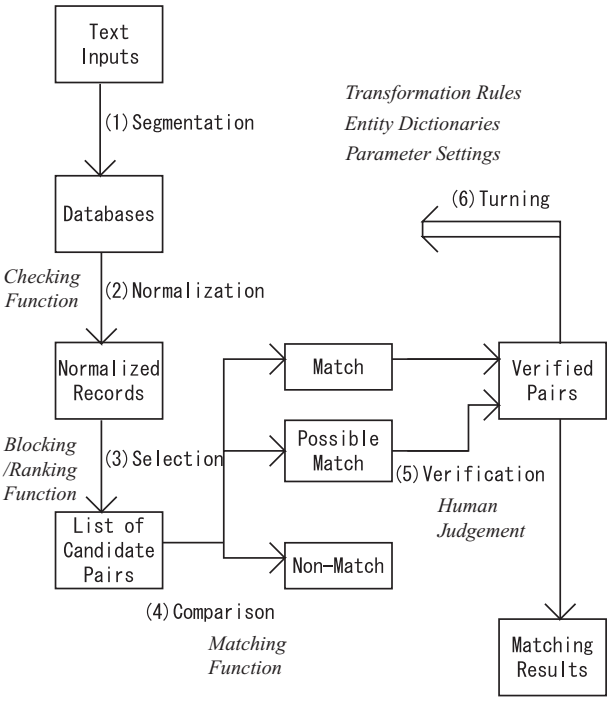


図 4: レコード照合システムにおける処理の流れ

があるレコード対の候補を数え上げる．

(4) 比較（comparison）

候補にあげられたレコード対の照合スコアを求め，match（照合）/possible match（照合可能性あり）/non-match（非照合）のいずれかに分類する．

(5) 検証（verification）

分類結果を受けて，必要に応じて人手による追加判定を行い，最終的な照合結果を出力する．

(6) 調整（tuning）

照合結果に基づきシステムの性能を評価し，選別や分類等におけるパラメタを調整する．また，必要に応じて，正規化のための辞書やルールを獲得する．

上記のうちで，レコード照合問題の中心的な検討課題となるのは，(2) の候補選択および (3) の候補比較である．これらは両者とも，いかにしてレコード間の照合スコア（類似度）を求めるか，という問題に帰着させて考えることができる．前者では，精度は低くてよいが取りこぼしがなく，大量のレコードを現実的な速度で処理できるような高速な手法が求められる．後者では，コストが高くても信頼性の高い手法が求められる．以下では，候補選択と比較に関する代表的な手法を簡単に紹介する．

4.2 代表的な候補選択手法

候補選択はしばしば文献中で blocking，filtering，detection，search 等の用語で参照される．blocking という用語を使う場合の基本的な考え方は，あらかじめレコードをいくつかのグループに分割し，照合候補の選択をグループ内のメンバーに限定して効率化をはかることである．これに対して search は，任意のレコードに対して，候補集合全体を調べて候補リストを作成する場合を含むものと考えられる．ただし歴史的に，これらの用語は厳密な区別なく使われる場合が多いようである．

候補選択の伝統的な方法として第 1 にあげられるのは standard blocking[10]である．この方法では，たとえば「苗字の先頭 4 文字」など，キーとなる属性（あるいは属性の組みあわせ）を 1 つ決めて，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．また，複数のキーに対してこのような操作を行う場合を multiple pass blocking と呼ぶ．

伝統的な方法の第 2 は sorted neighbourhood method (SNM) [15][16] である．この方法は，あらかじめ定めたキーの値にしたがってすべてのレコードをソートし，そのリスト上で固定長ウィンドウの範囲内にある近接レコード群を照合候補とする．SNM にはあらかじめグループ化したレコード群の中でソートを適用する clustering SNM，複数のキーでソートを繰り返す multi-pass SNM など，幾つかのバリエーションが存在する．また，リストを走査する際に，最新の代表的なレコードだけを特別なキューに入れて効率化をはかる priority queue method[19] も提案されている．

一方，近年になり提案された新しい候補選択法として，bigram indexing[17] と canopy clustering [18][9] をあげることができる．Bigram indexing は，キーの値を文字単位バイグラムの集合に変換し，ある一定の閾値以上でバイグラムが一致するレコードを候補とする方法である．閾値を設定することで，効率とあいまい性のトレードオフを可能にしている．一方，canopy clustering は，情報検索システムで広く用いられている tf-idf を類似度尺度として用いる方法である．レコードをランダムに順次選択しながら，tf-idf による距離が一定値以下であるようなレコード群をまとめて，canopy cluster と呼ばれるグループを構成する．最後に，同一の canopy cluster 内のレコードどうしを照合候補とする．

4.3 代表的な比較手法

候補の比較では，与えられた 2 つのレコードを比較して，match，possible match，non-match のいずれかに分

別する．候補比較は，レコード照合の中で最もコストが高い計算であり，対象領域に特化あるいは適応した手法が必要とされる．Sung らは候補比較の方法を，(1) ルールに基づく方法，(2) 類似関数に基づく方法，の２つに大別している[20]．

ルールに基づく方法では，たとえば「姓が等しく名が１文字違いならば同一人物とみなす」などの IF-THEN 型のルールを，適用領域にあわせて人手で作成してシステムに組み込む[15]．ここで性能向上のため，ルールに確信度を与える方法も提案されている[14]．

類似関数に基づく方法では，文字列比較の汎用的な関数である編集距離を用いることが一般的に行われる[21]．多くの場合，編集距離を計算するためのパラメタ（挿入，置換などの編集操作のコスト）は適応領域にあわせて人手により調整するが，機械学習を利用して適応的に距離関数を獲得する方法も提案されている[13]．その他の方法として，token, field, record の３つのレベル毎に順番に類似度を計算する record similarity[22]，tf-idf を適用する方法[23]，Q-gram を用いる方法[24] などが提案されている．

5　関連文献と代表的なシステムの例

レコード照合問題に関するサーベイとして，まず，U. S. Census Bureau 主催の２回のワークショップ論文集をあげることができる．第１回の論文集[25]（Section I, Selected Background Papers: 1959-1983）には，1959 年から 1983 年までの代表的な論文が集められており，その歴史を概観するのに役立つ．また同文献（Section II ～ V）には，1985 年の時点での技術概観，理論，応用分野等に関する包括的な情報もまとめられている．さらに第２回の論文集[26]（Chapter 11）には，1986 年から 1997 年までの代表的な論文，および 1997 年の時点での最新の技術動向が集められている．

さらに最近のサーベイとして，2000 年の IEEE Transaction on Data Engineering のデータクリーニングに関する特集号[4]に掲載された Monge による記事[27]，Winkler による U. S. Census Bureau のテクニカルレポート[28]，2003 年発刊の本に掲載された Sung らによる概説[20]，Gu らによる 2003 年の論文[8]等がある．また近年では，VLDB (International Conference on Very Large Databases)，ACM SIGMOD (International Conference on Management of Data)，ACM SIGKDD (International Conference on Knowledge Discovery and Data Mining)，WWW (International World Wide Web Conference）等の著名な会議において発表がされており，レコード照合問題に関する盛り上がりがうかがえる．

最後に，レコード照合を実装したプロトタイプシステムとしては，AutoMatch[29]，AJAX[30]，IntelliClean[14]，TAILOR[31]，Febrl[17] 等が知られている．また，Gu らの概説論文[8]では代表的な無料・商用ソフトウェアや利用可能なベンチマーク用データもあわせて紹介している．さらに，1997 の論文集[26]（Chapter 13）では，レコード照合ソフトウェアを評価するためのチェックリストが提供されており興味深い．

6　おわりに

レコード照合問題に関する研究は，1950 年代までさかのぼる長い歴史を持つ一方で，近年では，半構造化データやデータウェアハウスの登場を背景に，データベース，機械学習，情報検索等の分野からのアプローチが活発化している．しかしながら我が国においては，今回調査した範囲では発表がほとんど見受けられず，独立した研究分野としてあまり認識されていないようであった．レコード照合問題は，領域固有の知識を多く必要とすることから，学術的な一般化がむずかしいという側面もある．また我が国では戸籍制度が整備されており国勢調査等での人物名同定の必要性があまりなかったという社会的背景もあろう．しかしレコード照合は，今日の社会に氾濫する情報を整理し，役に立つ形で発信して行く上で不可欠な技術である．広くは，一般にテキスト情報を実世界上の実体に対応付ける操作とみなすこともできる．単純に諸外国の商用ソフトを導入しただけでは実用に耐えるレコード照合システムの構築は困難であると考えられ，我が国においても今後の研究の発展が望まれる．

謝辞

本研究動向調査は，国立情報学研究所が事業サービスを開始している文献情報ナビゲータ（CiNii）開発の一環として行いました．研究開発および運用に携わる富士通研究所，富士通株式会社，国立情報学研究所開発・事業部アプリケーション課の方々に感謝いたします．

文献

[1] H. B. Newcombe, J. M. Kennedy, S. J. Axford and A. P. James, Automatic Linkage of Vital Records, Science, 130 (3381), pp. 954-959, 1959

[2] Halbert L. Dunn, Record Linkage, American Journal of Public Health, 36, pp. 1412-1416, 1946

[3] J. T. Marshall, Canada's National Vital Statistics Index, Population Studies, 1 (2), pp. 204-211, 1947

[4] Erhard Rahm and Hong Hai Do, Data Cleaning: Problems and Current Approaches, IEEE Transaction on Data Engineering, 23 (4), pp. 3-13, 2000

[5] 有澤博, データベース理論, 情報処理学会, 1981

[6] Ivan P. Fellegi and Alan B. Sunter, A Theory for Record Linkage, Journal of American Statistical Association, 64 (328), pp. 1183-1210, 1969

[7] S. Gomatam, R. Carter, M. Ariet and G. Mitchell, An Empirical Comparison of Record Linkage Procedures, Statistics in Medicine, 21, pp. 1485-1496, 2002

[8] Lifang Gu, Rohan Baxter, Deanne Vickers and Chris Rainsford, Record Linkage: Current Practice and Future Directions, CMIS Technical Report, CSIRO Mathematical and Information Sciences, 03/83, 2003

[9] William W. Cohen and Jacob Richman, Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integration, Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (KDD2002), pp. 475-480, 2002

[10] Matthew A. Jaro, Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Society, 84 (406), pp. 414-420, 1989

[11] Vassilions S. Verykios, George V. Moustakides and Mohamed G. Elfeky, A Bayesian Decision Model for Cost Optimal Record Matching, The International Journal on Very Large Databases, Vol. 12, pp. 28-40, 2003

[12] Sunita Sarawagi and Anuradha Bhamidipaty, Interactive Deduplication using Active Learning, Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (KDD2002), pp. 269-278, 2002

[13] Mikhail Bilenko and Raymond J. Mooney, Adaptive Duplicate Detection Using Learnable String Similarity Measures, Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (KDD2003), 2003

[14] Mong Li Lee, Tok Wang Ling and Wai Lup Low , IntelliClean: A Knowledge-Based Intelligent Data Cleaner, Proceedings of the 6th ACM International Conference on Knowledge Discovery and Data Mining (KDD2000), pp. 290-294, 2000

[15] Mauricio A. Hernandez and Salvatore J. Stolfo, The Merge/Purge Problem for Large Databases, Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data (SIGMOD 1995), pp. 127-138, 1995

[16] Mauricio A. Hernandez and Salvatore J. Stolfo, Real-world Data is Dirty: Data Cleansing and the Merge/Purge Problem, Journal of Data Mining and Knowledge Discovery, 1 (2), 1998

[17] Peter Christen and Tim Churches, Febrl – Freely Extensible Biomedical Record Linkage, Computer Science Technical Reports, TR-CS-02-05, Australian National University, 2002

[18] Andrew McCallum, Kamal Nigam and Lyle H. Ungar , Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching, Proceedings of the 6th ACM International Conference on Knowledge Discovery and Data Mining (KDD2000), pp. 169-178, 2000

[19] Alvaro E. Monge and Charles P. Elkan, An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records, Proceedings of the ACM-SIGMOD Workshop on Research Issues on Knowledge Discovery and Data Mining, 1997

[20] Sam Y. Sung, Zhao Li and Tok W. Ling, Clustering Techniques for Large Database Cleansing, in "Clustering and Information Retrieval", W. Wu, H. Xiong and S. Shekhar eds., Kluwer Academic Publishers, pp. 227-259, 2003

[21] Alvaro E. Monge and Charles P. Elkan, The Field Matching Problem: Algorithms and Applications, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 1996), pp. 267-27, 1996

[22] M. L. Lee, H. Lu, T. W. Ling and Y. T. Ko, Cleansing Data for Mining and Warehousing, Proceedings of the 10th International Conference on Database and Expert Systems Applications (DEXA 1999), pp. 751-760, 1999

[23] William W. Cohen, Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity, Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data (SIGMOD 1998), pp. 201-212, 1998

[24] L. Gravano, P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan and D. Srinivasta, Approximate

String Joins in a Database, Proceedings of the 27th International Conference on Very Large Data Bases (VLDB 2001), pp. 491-500, 2001

[25] Beth Kilss and Wendy Alvey eds., Record Linkage Techniques – 1985: Proceedings of the Workshop on Exact Matching Technologies, Statistics of Income Division, Internal Revenue Service Publication 1299-2-96 (available from http://www.fcsm.gov/working-papers/RLT __ 1985.html), 1985

[26] National Research Council, Record Linkage Techniques – 1997: Proceedings of an International Workshopthe Workshop and Exposition, National Academic Press (available from http://www.fcsm.gov/working-papers/RLT __ 1997.html), 1997

[27] Alvaro E. Monge, Matching Algorithms Within a Duplicate Detection System, IEEE Transaction on Data Engineering, 23 (4), pp. 14-20, 2000

[28] William E. Winkler, The State of Record Linkage and Current Research Problems, Technical Report RR/200/06, Statistical Research Report Series, U.S. Bureau of the Census, 2000

[29] Matthew A. Jaro, Software Demonstrations, Proceedings of an International Workshop and Exposition - Record Linkage Techniques, 1997

[30] Helena Galhardas, Daniela Florescu and Dennis Shasha, AJAX: An Extensible Data Cleaning Tool, Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD 2000), pp. 590, 2000

[31] Mohamed G. Elfeky, Vassilios S. Verykios and Ahmed K. Elmagarmid, TAILOR: A Record Linkage Toolbox, In Proceedings of the 18th International Conference on Data Engineering (ICDE 2002), 2002