日本がん登録協議会
第31回学術集会
ポスター演題 R-1-4

リレーショナルデータベースシステムを用いない汎用的照合システムの開発

田渕 健[1)2)], 柿崎 裕則[2)], 下山 達[1)]
1)東京都立駒込病院, 2)東京都がん登録室

【利益相反】

【学術集会名】
日本がん登録協議会第31回学術集会

【筆頭演者氏名】
田渕 健(東京都立駒込病院)

【COI開示】
当演題に関し, 開示すべきCOIはありません

【背景】

- 個人同定を要するデータベースにおいて, 照合工程が, データの品質や統計精度を左右する
- 全国がん登録システムは本邦がん登録史上初めてがん登録業務の体系的な自動化が導入され, かつ, 照合精度や速度が向上し, 東京都のデータ規模でも実行可能性を担保されたが, 次のような改善点がある
- 登録業務においては, 照合によって判明した不整合を電子データとして出力する機能が装備されておらず, 照会調査に人手と時間がかかる
- 情報の提供における外部照合は, 工程が煩雑で時間がかかり, 他の業務への配慮が必要

【目的】

- 届出業務においては, 全国がん登録システムとは別にデータ加工レベルで, 事前に照合上の不整合・エラーを検出し, 殆ど人手を介さずに, 照会調査リスト出力の定型処理を行い, また, 照会調査回答結果の反映を自動処理を行った後, 全国がん登録システムに移行する
- 外部照合業務においては, 全国がん登録システムを用いずにデータ加工レベルで前処理に引き続いて一連の工程として照合業務を行う
- 以上の開発経緯を本演題では数式やプログラミングを用いずに紹介する

【方法】全国がん登録実務業務改善の一環

- 全国がん登録標準手順と安全管理措置に従い『データ加工』業務改善の一環として実施
- 厳密な比較検討は行っていない(ほぼ不可能)
- 外部非接続東京都がん登録室内LANシステムで実施
- データ加工で用いるデータ様式は, 全てテキストファイル(CSVファイルないしExcelファイル)である
- 固定的にはもちろん一時的にもリレーショナルデータベース(RDBMS)の機能は行わない
  - 全国がん登録データベースの正当性を優害していない
- 本開発に特化した人員及び資金は用いていない

【方法】データファイル

【方法】届出情報の事前照合のプロセス

【方法】外部照合のプロセス

【方法】照合アルゴリズム

1. 同一性の定義
   - 氏名漢字は普通漢字(R1-7参照), 住所については町丁レベルでの住所コード, 生年月日と性別はそのものの同一性を定義し, 同一性サブスコアは各組合項目に応じて重みを付けて設定
2. 類似性の定義
   - テキストは編集距離で類似性サブスコアを設定, 生年月日は完全選択ミスやタイプミス等を考慮した類似性サブスコア設定
3. 照合項目正規化・様式変換・データクレンジング, 同一性変換
   - 区市町村データでは独自外字のユニコード変換が特に必要
4. 照合項目(氏名, 生年月日, 住所, 性別等)をペア関に全国がん登録データと照合対象リストの共通部分を抽出し, ソートする
5. 上記ペア毎に, 共通項目には同一性サブスコアを適応, それ他の項目には類似性サブスコアを計算し, 合計してペア毎照合スコアを算出し, 閾値以上を同一とすることで照合結果を判定

【方法】スコアリングの設計

- 同一性に基づくサブスコアリング
  - 全国がん登録システムに採用されている方法に準拠(第24回本学術集会抄録1-11新野ら(国がん), 2015)
  - 照合指標の設定, 重み付けスコアリング
- 類似性に基づくサブスコアリング
  - 日付データ: 1-2個の数字のミスタイプや元号選択ミス等を考慮した日付の類似度判定を行った
  - 氏名: 氏名中の漢字は異体字・類似文字の代表文字に変換した上, 文字列の類似度判定を行った
  - カナ氏名: 文字列の類似度判定を行った
  - 文字列類似度判定にはpythonのdifflibライブラリを使用

【結果】一貫工程化による効率化

- 前処理・後処理などの周辺業務や一連の業務との一貫工程化を実現
- 各工程での予期しない不整合発生や暴走を防止するため, 工程間チェックポイントを設置(工程管理表)
- 各工程は完全電子化して, シームレスに自動処理を進めるため, がん登録個人情報の紙出力は不要(ペーパーレスの実現)
- 照合上の不整合を事前に検出することで, 届出票の票内エラーと票間エラーを一元的な「照会調査」の実施が可能となった

【結果】人手作業の軽減

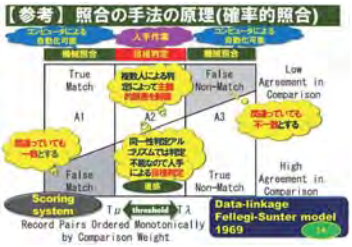
- 照合における目視同定は労働集約的な人手作業であるが, 提案手法では大幅な自動化によってこの様な作業を殆ど回避できるようになった
- 周辺業務との一貫工程が可能となったことで, 転記作業(コピペ)等の人手作業が不要となった
- 東京都のデータ規模において, 他のがん登録業務を含めても, 作業担当者1名分の人手作業に縮約された
  - 作業担当者は時間外業務を全く行っていない
  - がん登録業務の内, 全国がん登録システムでのデータ処理に必要な時間は週2日

【結果】東京都がん登録室における実務作業担当者数の推移と情報の提供実績

【結果】照合に係る業務時間の大幅短縮

- 純粋な照合時間はシステム環境と照合対象に依存
  - 以下に示す照合時間は参考値(照合対象によって異なるシステム環境で実施しているので単純な比較や一般化は困難)
- 届出票の事前照合時間は1分/500件
  - ファイル読込等の時間は一定であり, 当該環境下で約50秒
- がん検診精度管理外部照合時間は7-10分/10万件
  - 照合対象が カバーする地名が多いほど照合時間がかかる
  - がん検診精度管理では照合対象の住所が同一自治体在住者が対象である
- 照合そのものにかかる時間は, 全国がん登録システム(初代)よりも所要時間を1桁短縮

【参考】照合の手法の原理(確率的照合)



【結果】照合精度の向上の可能性

- 照合精度の向上とは, false-matchやfalse-mismatchを減らすことである
  - 照合精度については本来は比較検討を行うべきだが冒頭にも述べたように, 比較検討は不可能であるため, 事例ベースでの結果であることをお断りしたい
- 人手作業による同定判断ミスは大幅に減少した
- 目視同定で見逃している同定が可能となった
  - 類似度判定を導入したことが大きい
- アルゴリズムの例外処理が完全かどうかは断定できないため, 最悪系統的なエラーが発生する危険性がある

【考察】客観的な比較検討の困難性

- 厳密には, 業務手法についても比較試験を行うことがエビデンスに基づく方法であるのは言うまでもない
- しかし, 特に東京都がん登録室のようにデータ規模の多い場合, 現実に実施している業務を標準的な枠組みの中で, 中断することなく通常通り継続しながら, 業務改善を進める形式において, 客観的な比較を行うことは困難である

【考察】全国がん登録システム

- 全国がん登録システムの照合システムは, 地域がん登録標準DBSの照合システムに比べ, 機械照合の導入, 照合速度の短縮, 照合精度の向上が図られた
  - 東京都では標準DBSの運用は事実上困難だったが, 全国がん登録システムでは実行可能性が担保された
- 全国がん登録システムであっても労働集約的業務であり, 更に, 必要人員数はデータ規模に比例関係以上の人員が必要である
- 人員数が一定以上を超えると業務品質の維持が困難
  - どうしても職員間格差が生じ, 登会のように複互関係が重要な業務では「悪貨が良貨を駆逐する」ことが生じる

【考察】ルールベースの仕事

- がん登録はデータマネージメントの一種であり, 基本はルールベースの仕事であり, 自動化が可能ではないかと思われる仕事
- 実際には, 多くのデータマネージメントの仕事同様, 標準化文書には記載されていない暗黙の了解が多数含まれるため, 一般的には完全自動化は困難とされる
- 現実のがん登録実務の人手作業を観察していると, 形式的な作業に終始し, 人手作業に期待される判断力が生かされていない
  - 判断を要する作業が形骸化して, 誤った判断が増加する
  - データ規模が大きい東京都の傾向かもしれない

【考察】票間エラー判定自動化の困難度

- 当登録室では標準DBSでの運用困難状況を目の当たりにして実務の自動化を模索してきた
- 票内エラー判定の自動化に比べて票間エラー判定(照合・集約)の自動化は難易度が高かった
- 票間エラーは, ルールの組み合わせが複雑なのに加えて, データの組み合わせが膨大になる
  - ルールの組み合わせについては, 確慎れば全部書き下すことは不可能ではない
  - データ規模Nに対して, 最大N^2の処理量を必要とする
  - データ規模が増加するとある閾値以上で急激に処理量が増えて, 膨大なリソースと時間を必要とし, 処理が不能に陥る

【考察】データ組合せに対する処理提

- 提案手法ではデータ規模Nの組合せのN^2の処理量をいかに減らすかが課題である
- 一致も類似もしない組合せ計算は無駄な負荷となるので, 原データと照合対象の共通部分のみに縮減した
- システムでのデータ処理に負荷がかかるのは, 一致度(類似度)の計算であるので, 上記共通部分のソートを行った後, 近接部分のみについて計算する
- 照合指標の組合せの組について複数セット(a回)行うとし, ファイル読み出しなど一定処理分bを加えると, 処理量はaNlogN+bとなり, N^2より大幅に縮減される

【考察】データ組合せ処理の方法

- データの組合せ処理は, 理想的にはリレーショナルデータベースのインデックスを用いることで高速化が可能
  - ⇨全国がん登録のデータ加工においてリレーショナルデータベースを構築することは, 安全管理措置の一時ファイルでの処理とは言いがたく, かつ, 全国がん登録データベースのクローンが発生するという正当性の問題も生じる
- テキストベース(CSVファイル)での通常のデータ組み合わせ処理を行うと, 単純な一致でさえも時間がかかり, 類似検索に必要な処理時間は現実的ではない
  - ⇨本提案手法は類似検索を行いながら大幅な時間短縮を実現した

【考察】今後の課題

- 本提案手法のがん登録データ以外でもデータ組合せ処理を行うアルゴリズムとして一般化可能か
- ルールと知識をコンピュータに指示するルールベースの手法は, それ自体が労働集約的であり, ルールや知識の変化に対応できない
- ⇨プログラマが常に新しい知識にキャッチアップし続けるのは困難であるので, 深層学習の手法を取り入れて, 専門家の判断レベルの指示をコンピュータだけで実現する必要がある

【結語】

- RDBMSを用いないで, 類似検索を考慮しかつ実用的な速度で照合を行う手法を開発した
- 照合に当たって, 全ての組合せを網羅的に検証するのではなく, 厳い条件での一致条件データを抽出・ソートした特定の一数データの近傍だけで時間のかかる類似計算処理を行うことにより, 大幅な高速化が実現できた