

# 重複する顧客データを 高速に統合する名寄せ技術

従来の約10倍の処理速度を実現

「名寄せ」は、銀行がペイオフ（預金保険制度）のために同一預金者の全口座を確認する作業として広く知られていますが、企業においても、企業合併や企業内のICTシステムの統合等に伴い、顧客データに対する名寄せのニーズが高まっています。今回は、名寄せの高速化を実現した富士通研究所の新しい技術をご紹介します。

## 項目の類似性を1文字ずつ 計算して求める名寄せ

「名寄せ」とは、顧客データベースに登録されたレコードの中から、同一の顧客を表すレコードの集合を求めることです。例えば、「(株)富士通研究所」「富士通研」「富士研究ぢよ」という3つのレコードが登録されていた場合、2つ目は省略表記で3つ目はタイプミスだから、これらは1つのレコードに名寄せしていい、と人間の目にはわかりますが、コンピュータで自動判定させるには相当量の計算が必要になります。

名称や住所等、顧客を特定しうる項目が、完全に一致しなくてもほとんど一致している／類似性が高いといった判定を行う手法の一つに、「編集距離」<sup>①</sup>があります。顧客データの場合、1項目あたりの文字数は長くても数十文字程度なので、一文字ずつ比較し

て編集距離を計算しても計算量はそれほど多くありません。ところが、例えば100万件のレコードに対して実行すると、1項目あたりの計算量が1兆組（100万の二乗）も実行することになり、相当な時間を要します。

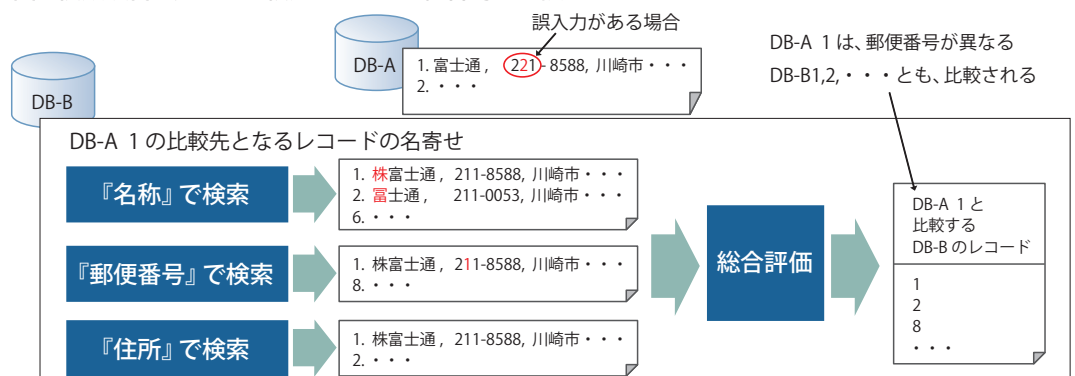
## 速度と精度の両立が課題

そこで従来は、郵便番号等を使ってデータベース上のデータを複数のグループに分割したうえで、グループ内のレコード同士で類似性を判定する方法がとられてきました。例えば、100万件を100件ずつ1万個のグループに分割すれば、全体の計算量を1億組（前述の1 / 10,000）に減らすことができます。しかし、郵便番号自体に誤入力があった場合、比較すべきレコードが同一グループにならないため類似性の判定が行われず、名寄せの

### ①編集距離

一方の値をもう一方の値にするために書き換える文字数。文字列間の類似性の判定に利用される計算手法の一つ。例えば「株富士つ研究」を「富士通研究所」に書き換えるには、「株」を削除、「つ」を「通」に置換、「所」を挿入のように3文字の書き換えが発生するため、編集距離は「3」となる。

■ 図1 複数項目を用いて比較先レコードを総合的に選択



見落としが発生します。

このように名寄せは、速度と名寄せの精度を両立することが困難とされ、大規模なデータに適用できる手法が課題となっていました。

## 顧客データの名寄せに特化した類似検索で高速化

今回、富士通研究所が開発した名寄せ技術は、顧客データの名寄せに特化した類似検索を用いることで、データベースをグループに分割せず名寄せを高速化することに成功しました。

本技術では、一つのレコードに対して、名寄せの可能性が高いレコードだけをデータベースから抽出して比較対象にします。この時、比較先対象レコードを、名称・郵便番号・住所等の複数の項目を用いて総合的に抽出することで、名寄せの見落としを軽減しています(図1)。

本技術の開発にあたり、富士通研究所は、顧客データに発生しやすい相違パターンを調査しました。その結果、変換ミスやタイプミス、略称、異体字(旧字体と新字体等)による相違が典型的ではあるものの、項目内で複数回発生する可能性は低く、名寄せの可能性が高いレコードの各項目は文字列の1カ所のみ異なっていることが多いことが分かりました。

そこで、例えば「富士通研究所 / (株)富士通研究所」「富士通研究所 / 富士通研究所」のように、1カ所のみ異なっている文字列を高速に検索できる類似検索技術を開発し、検索結果として「一方が文字を省略している」「先頭 / 末尾の文字列が異なる」「中間の文字列が異なる」というレコードの組み

合わせだけを抽出して編集距離による名寄せ実施の判定を行い、前述の計算量を大幅に減らしています(図2)。

■ 図2 顧客データで類似性が高いレコードの組み合わせ

### 例1 一方が文字を省略している場合

〔富士通研究所〕    〔富士通研究所〕  
富士通研                    (株)富士通研究所

### 例2 先頭 / 末尾の文字列が異なる場合

〔富士通研究所〕    〔富士通研究所〕  
富士通研究所            富士通研究ぢょ

### 例3 中間の文字列が異なる場合

〔富士通研究所〕  
富士つう研究所

## 企業内でも高まる名寄せニーズ

本技術を用いて顧客データ約200万件に対する名寄せを行ったところ、従来技術で15.5時間かかる処理を1.4時間で終わらせることができ、同等の見落とし率を保ちながら高速化し、約10倍の処理速度を実現しました<sup>②</sup>。

名寄せは、銀行や保険会社の業務、企業合併のみならず、企業内においても、「営業部門と出荷部門で別管理していた顧客データベースを統合したい」「担当毎に登録していた顧客の重複データを統合したい」「異なるデータベースをつないで顧客動向を分析したい」等、様々なニーズが高まっています。

富士通研究所では、2011年度中に顧客データの名寄せの実用化を目指すとともに、社内文書やWeb上のテキスト、画像・動画のタグ等、異種データの名寄せを実現し、様々な情報を連携させるサービスの提供につなげていきます。

② 10項目、約200万件の法人データに対する結果。名寄せ結果のうち約900件をランダムに抽出して見落とし率を評価し、従来技術で5.3%、本技術で5.2%とほぼ同じ値になる条件で処理速度を比較。