5B−9

電話帳ＤＢと地図ＤＢにおける結合方式の評価

唐沢裕明°　川辺秀樹

ＮＴＴ　情報通信網研究所

【１】　はじめに

近年、多種多様な用途のデータベースが構築されてきており、こうした複数のデータベースの関連付けを行う（本稿においては「結合」と呼ぶ）ことにより提供できる新しいサービスの分野が広がっている。電話帳ＤＢについても、地図ＤＢと結合することによって地図案内サービスなど［２］［３］を提供できる。しかしながら、これらのデータベースは構築者を異にするため、データベース項目の構成や格納形式が異なっており容易に結合してサービスを開始することはできない。そこで本稿では、電話帳ＤＢと地図ＤＢにおいて、名義等の言語解析技術と住所情報の拡大処理を組み合わせた結合アルゴリズムを提案し、その評価を提示する。

【２】　結合方式の検討

一般に市販されている電話帳ＤＢと地図ＤＢの主なレコード構成および結合後のデータベースのレコード構成の一例を図１に示す。

電話帳ＤＢと地図ＤＢにおける共通項目は名義，ビル名および住所であり、このうち住所だけはコード化されているため他の項目と性質を異にする。名義およびビル名は自然言語で記述されているため、その表現にゆらぎや誤りが存在する。

一方、住所はコード化されているため、自然言語にみられるような表現のゆらぎが存在しない代わりにコード化データの欠如（不完全性）や誤りが存在する。これらのデータベース項目の性質を考慮して結合を行なうための検討［１］を行なったが、その処理内容を図２にまとめる。

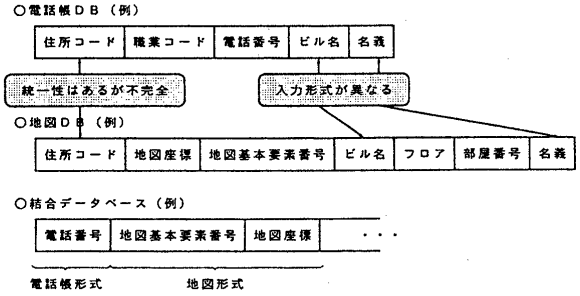

図１　データベースのレコード構成例

【３】　結合方式の実装

電話帳ＤＢと地図ＤＢとの結合方式の実装を行なった。本稿では電話帳ＤＢの電話番号から地図ＤＢの建物図形を指示するような結合ＤＢを想定しているためにビル名が同一と判断されたレコード同士も結合されたものとする。ここで実装した結合方式を、結合されるデータの候補を抽出するパートと、その抽出されたデータの候補群の中から結合されるべきデータを選択するあるいは正当性を確認するパートの２段階のパートに分けて述べる。

ＤＢ項目	表現形式	性　質	処　理
名義 ビル名	規格化されていない一般文字列	ゆらぎ 誤り	・固有情報（項目内容を一意に指示し得るような（会社名，姓情報等）を抽出して比較 ・字面のパターンマッチ
住所	コード	不完全性 誤り	・結合相手候補の絞り込み ・結合相手候補の範囲の拡大・縮小

図２　データベース項目処理内容

（１）結合候補抽出パート

対象となるデータベースから結合条件の緩和によって得られる結合候補レコードを抽出するパート

次の①と②は住所と名義，ビル名のそれぞれの結合条件の緩和を示す。

①　住所コードにより結合対象の住所エリアを拡大

②　①の住所エリア内の名義，ビル名の自然語処理により、以下のレベル順に結合候補を抽出

・名義，ビル名の完全一致
・正規化処理後の完全一致
・固有情報抽出後の完全一致
・固有情報の正規化処理後の完全一致

（２）正当性確認パート

結合候補抽出パートと組み合わせて適用し、一旦結合候補抽出を行った後、候補の中から結合されるべき正当なレコードを選択あるいは確認するパート

次の①と②の条件を適宜勘案して結合の相手レコードを決定

①　結合候補抽出時の自然言語処理による結合のレベル

②　住所コードの一致の度合い

これら２つのパートで行われる正規化処理や固有情報処理などの個々の処理の組み合わせを結合アルゴリズムと呼ぶことにする。今回、実装した結合処理アルゴリズムを図３に示す。

【４】　結合方式の評価法

結合方式の評価として、以下の２つの評価要素を導入することにより表現する。

結合の対象となるレコード数 U のデータベースにその結合アルゴリズムを作用させたときに結合できたレコード数を Q_s とする。さらに、そのうち人為的な判断によってもその結合が正しいとされたレコード数を Q_r とする。

（１）結合成功率：P_s

結合処理で抽出された結合レコードが１つでも存在したときに結合成功とした場合、全体のレコード数に対する結合成功数の割合。

An Evaluation on Record Matching Method between Different Databases

Hiroaki KARASAWA, Hideki KAWABE

NTT Network Information Systems Laboratories

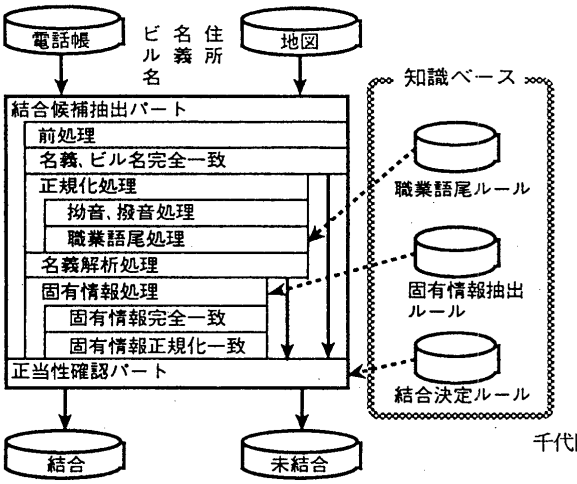

図3　結合アルゴリズム

$$P_s = \frac{Q_s}{U}$$

（2）結合信頼率：P_r

結合データベースの現実世界に対する信頼性は改めて調査を行う以外には評価することはできない。実際の結合処理は結合する対象となる複数のデータベースによって記述された世界に閉じており、結合処理における信頼性の判断は人間の判断に頼る他はない。従って、人為的な結合を行った際の人間によって結合されるレコードを結合信頼率１００％と考える。このとき、結合処理アルゴリズムによって結合成功した全レコード数に対する人間によっても結合すると判断したレコード数の割合。

$$P_r = \frac{Q_r}{Q_s}$$

この２つの評価要素は結合処理のアルゴリズムと対象となるデータベースに依存しており、この２つの総称を結合率P_mと呼ぶことにする。すなわち全体のレコード数に対する誤りのない結合数の割合を示す。

$$P_m = P_s \times P_r$$

これらの評価概念は、結合アルゴリズムで作成された結合データベースの結合の度合や信頼性を示している。また、基準となるデータベースを設け、それに対して結合を行なうことにより結合処理アルゴリズムの評価とすることができる。

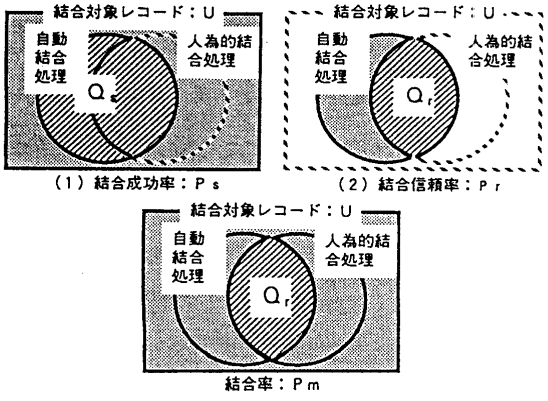

図4　結合方式の評価法

【５】　結合方式の評価

電話帳ＤＢに千代田区タウンページ６７，９９８件と市販の地図ＤＢを用いて、自動結合処理を行った評価結果を報告する。

上記のデータベースを、実装した結合アルゴリズムにより結合を行なった時、結合成功率を個々の処理単位に分類してグラフ化したものを図５に示す。

結合信頼率は結合条件を緩和するほど低下するという定性的な性質を考慮すると、固有情報正規化一致による結合レコードの信頼性は他の処理に比べて低いことがわかる。この結合信頼率の低い処理であってもサンプル調査により９０％以上の結合信頼率を示すことから、検討を行なった結合方式は有効であると言える。

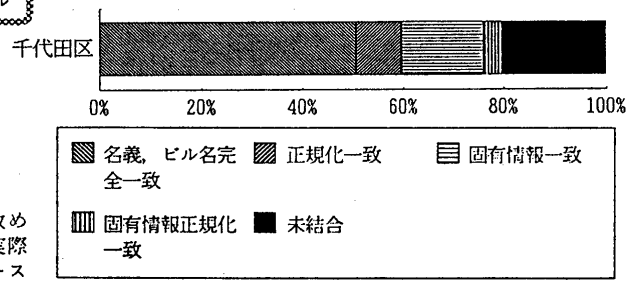

図5　自動結合処理の評価

【６】　結論

本稿では、電話帳ＤＢと地図ＤＢの結合方式の実装を報告した。実装したプログラムはＣ言語で９，０００ｓｔｅｐであった。また、結合の評価法を提案し、実装したプログラムにより自動結合処理の評価を行ったところ結合成功率約８０％、結合信頼率９０％以上を達成した。

【７】　おわりに

検討した処理の未実装な部分を実装し、結合アルゴリズムを構成する。その結合成功率の評価をとり、人為的な結合実験を併せて行い、それぞれのアルゴリズムにおける結合信頼率を算出する。これらの評価要素から電話帳ＤＢと地図ＤＢの結合に最適な結合アルゴリズムを見出すことができる。

参考文献

［１］唐沢：「異種データベース結合方式の検討」，平成３年第４３回情処学会全大，１Ｍ－６，４－１０５
［２］安田，松村，水町，唐沢：「電話・ＦＡＸを使った地図案内システム」，第３回機能図形シンポジウム講演論文，ｐｐ．８１－８６，１９９２
［３］川辺，唐沢：「オンライン経路案内システム」，平成４年第４５回情処学会全大，６Ｎ－４