5C-5

同義性判定ルールを用いた重複レコード照合システム

齋藤 悠　立石 健二　久寿居 大
NEC インターネットシステム研究所

1. はじめに

顧客 DB 等の多くの DB では表記のゆれを含み、レコード登録の際、登録済みのレコードをうまく検索できず同じレコードを何度も登録してしまうことがある（重複レコードの発生）。DB に重複レコードがあると、例えば一つの宛先に何通も同じ請求書を送ってしまうなどの問題が起こる。そのため DB 内の重複レコードを検出・削除する必要がある。しかし、対象となる DB は多くの場合大規模で、表記のゆれを含むため、人手で照合作業を行うのは容易ではない。そこで、照合作業を支援するための技術やシステムが提案されている[1]。

重複レコードを照合するための技術として、例えば、類似度判定に距離関数を用いる方法や、同義語辞書や表記ゆれに対応したルールを用意する方法が提案されている。しかし企業名等の固有表現には新語が多いため、同義語辞書は更新・メンテナンスの負担が大きい。一方で、企業名等の同義語は、比較的一定の規則で同義語が生成されることが多い。例えば、「株式会社」と「(株)」という部分文字列の置き換え規則は多くの企業名の同義語に出現するため、この規則を用いると「株式会社日本電気」と「（株）日本電気」、「情報処理株式会社」と「情報処理（株）」は同じであると判定できる。このような置き換え規則は多くのバリエーションがあり、どのような規則が効果的なのかを判断することは難しい。

本稿では、複数の同義語グループ間に共通する省略/置き換え規則を同義性判定ルールとして自動抽出し、それらを用いた重複レコード照合システムを提案する。また、提案手法で抽出された同義性判定ルールが重複レコードの検出に有効であることを検証する。

2. 同義語辞書を用いた重複レコード照合システム

提案する重複レコード照合システムは、図 1 のように重複候補検出部、重複判定部から構成される。重複候補検出部は、照合対象 DB から同義語辞書と編集距離を基にレコード同士の類似度を計算し、類似度が閾値以上のレコードを重複レコード候補グループとして出力する。類似度は同義語辞書に登録されている部分は同一文字列として編集距離を計算することにより求める。例えば、「日本電気／東京都港区芝 5−1」「NEC／東京都港区芝 5−1」の二つのレコードは「日本電気＝NEC」という同義語が登録されている辞書を用いれば完全一致となるが、辞書がなければ住所のみが一致する文字列となる。

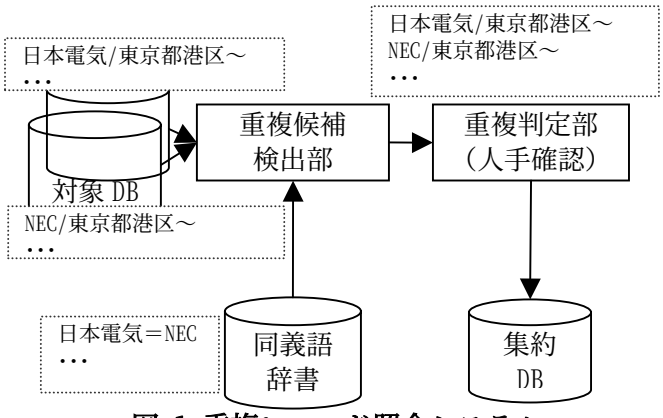

図 1 重複レコード照合システム

重複判定部では、重複候補検出部が出力した重複レコード候補グループ毎に、ユーザが目検し、重複の場合はそれらの重複関係を集約 DB に保存する。

3. 同義性判定ルール

3.1. 同義性判定ルールの抽出方法

例として、表 1のような 4 つの同義語グループを考える。以下、各処理の内容について例を用いて説明する。

Step1) 任意の二つの同義語グループを比較し、最長な共通文字列をルールとして抽出する。この処理をすべての同義語グループの組み合わせに対して行う。

例．表 1の任意の同義語グループからは、「(株)/株式会社」（Group:1&2、1&3、1&4、2&3、2&4 の比較による）と「工業株式会社/工業(株)」（Group:3&4 の比較による）がルールとして抽出される。

The System for Record Linkage with Synonymous Rules
Haruka Saito, Kenji Tateishi, Dai Kusui, Internet Systems Research Laboratories, NEC Corp.

表 1 同義語グループ

Group	同義語	
1	日本電気株式会社	日本電気(株)
2	株式会社情報処理	(株)情報処理
3	データ工業株式会社	データ工業(株)
4	言語工業株式会社	言語工業(株)

Step2) 抽出したルールの出現頻度を計算する。ルールの出現頻度はルール抽出元である同義語グループの数で計算する。
例. ルール「株式会社/(株)」は Group:1、2、3、4、から抽出されているので出現頻度 4 である。
ルール「工業株式会社/工業(株)」は Group:3、4、から抽出されているので出現頻度 2 である。
Step3) 出現頻度がある閾値以上のルールのみを採用する。

3.2.同義性判定ルールの例

表 2では、3.1節の方法で抽出した同義性判定ルールとその出現頻度の例を示している。抽出元の同義語辞書は、一部上場企業名の 14000 語(代表表記として 1600 個)を用いた。

4. 評価実験

4.1.実験内容

照合対象の DB として、企業名/住所をフィールドに持つ 160 万件の顧客 DB を用いて、同義語辞書および同義性判定ルールを適用した場合とそうでない場合とで検出される重複レコード候補数、精度、処理時間を比較する。各レコードは企業名フィールドと住所フィールドを持ち、フィールド毎の文字列類似度の平均を重複レコード検出に用いるレコードの類似度とする。住所フィールドについては編集距離に基づいた類似度計算を行う。企業名フィールドについては同義性判定ルールや辞書を用いる場合は表記を変換した後で類似度計算を行う。類似度 95%以上のレコード群を重複レコード候補とする。精度は、重複レコード候補をランダムに 100 グループ選び、重複レコードである割合で評価する。重複かどうかは評価者が目検で確認する。なお、同義性判定ルールは3.2節で抽出したルールのうち出現頻度 10 以上のものを採用する。

実験結果を表 3に示す。表中の(A)は同義語辞書なし、(B)は同義語辞書あり、(C)は同義性判定ルールありの場合について示している。「候補数」の項目で「+」記号に続く数値は、(A)同義語辞書なしで見つかった候補以外に新たに見つかった候補グループ数を示している。

4.2.考察

検出されるレコード候補が多く、かつ精度が高ければユーザは最小限の確認数で漏れを少なくできるので、重複レコード検出に有効であるといえる。

表 3の B(辞書)と C(ルール)に注目すると、ルールは辞書より少ないルール数で高い精度を保ったまま多くの重複レコード候補を検出できている。「(株)/株式会社」などの、適用範囲が広くかつ精度が高いルールが効果的であったと考えられる。処理時間の点からも、ルールの参照回数が少なく抑えられ効率的である。一方で検出結果を詳細に調べると、同義語辞書では検出できるがルールでは検出できない重複候補レコードが存在した。例えば、「日本電気」を含むレコードと「NEC」を含むレコードがある。以上のことから適用範囲の広い同義性判定ルールと同義語辞書とを重複レコード照合システムに組み込むことが、重複レコード検出に有効である。

表 2 同義性判定ルール抽出例

同義語ルール		頻度
株式会社	（株）	1235
工業株式会社	工業（株）	318
株式会社	業	292
ホールディングス	ＨＤ	21
建設（株）	組	15
海上	保険（株）	6
ソフトウェア（株）	ソフト	2
情報システム	ＩＳ	2
ファイナンス	信販	2

表 3 重複レコード照合システムでの精度

辞書	語数/ルール数	重複候補数[グループ]	精度[%]	時間[m's]
A(なし)	0	65700	100	4'18
B(辞書)	14000	+4051	99.8	10'23
C(ルール)	270	+21932	99.0	6'34

5. おわりに

本稿では、同義語辞書に頻出する文字列を同義性判定ルールとして抽出する方法を提案し、それらルールと同義語辞書を組み合わせることが重複レコードの検出に有効であることを検証した。

参考文献

[1]相澤彰子 高須淳宏 大山敬三 安達淳, "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", NII Journal, No. 8, pp. 43—51, 2004.