

DBシステム再構築におけるスキーママッチング適用

○鹿島 理華[†] 佐藤 彰洋[†] 永嶋 規充[†]

三菱電機株式会社 情報技術総合研究所[†]

1. はじめに

企業内の情報システムは、業務毎に独立したシステム、サブシステムを段階的に構築してきたため大規模で複雑なシステムとなり、データも各システムに分散し個別に管理されていることも多い。これは、同じ意味を示しているにもかかわらず名称が異なる項目名のばらつき（例:CORPNAME と CORP_NM）やデータの二重持ちなどのデータの品質の低下、メンテナンス負荷増加、分析などのデータの二次利用が困難といった問題につながる。

これに対し、既存システムへの影響少なくデータを統合するために、データ自身ではなく、メタデータと呼ばれる“データに関するデータ”だけを統合する DB システム再構築の一方式を提案した[1]。ここでは異なるシステム間で管理されているデータ項目の把握が必要であるが、そこにわれわれの持つスキーママッチング技術[2]を適用した場合の評価結果を述べる。

2. 課題

新規にサブシステムを追加する時のテーブル設計では、システム内に既に存在するデータ項目であれば、性能に影響がない限りデータの二重持ちを避け、既存のものを参照してシステム内のデータの品質を維持する必要がある。ここで、指定したカラム名と類似するものを対象システム内から見つけを推薦する設計支援の中で、スキーママッチング技術を用いる。

スキーママッチングは DB 間のカラム対応を、カラム名やデータ型、桁数などの定義情報の類似性に基づいて高精度に推薦する。

今回開発しようとしている設計支援では、選択したテーブル間でスキーママッチングをするという我々のこれまでのスキーママッチング技術の活用とは異なり、システム内の全カラムがスキーママッチング対象になり、システム内のカラム数の総計が多い場合、性能がリアルタイム処理に耐えられないという課題がある。

3. 類似推薦の応答性能向上の方式検討

課題解決のため、1つのカラム名に類似するカラムを、システム内の全項目から推薦するときの応答性能向上のための方式検討を行った。

3.1 スキーママッチングの入出力仕様

まず、スキーママッチングの入出力仕様について説明する。

スキーママッチングは、対象とする2つのテーブルの XML スキーマ定義をファイルで与え実行する。入力は XML スキーマ定義 (XMD) 形式、出力は2つのテーブルのカラム間の類似度 (0~1 の数値) を示すマトリックス表となる。

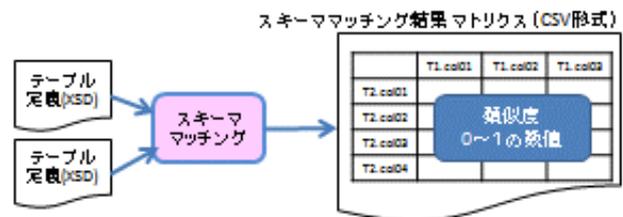


図 1 スキーママッチングの入出力情報

3.2 仮想テーブル方式

指定したカラムの類似カラムを推薦するとき、システム内の全テーブルとスキーママッチングを行う。そこで、方式案1は全テーブルの全カラムからなる仮想テーブルの XML スキーマ定義ファイル(A)をあらかじめ作成しておき、ユーザからの類似カラム名の推薦の処理要求時には、その1つのカラムのみを持つ仮想的なテーブルの XML スキーマ定義ファイル(B)を作成し、仮想テーブル(A)と(B)のスキーママッチングを行う方式とする。こうすることにより、スキーママッチングは1回のみ実行すればよいことになる。

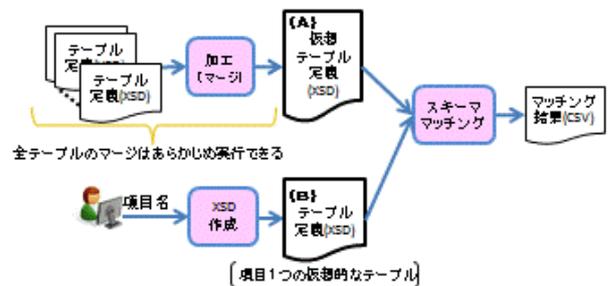


図 2 方式案 1

A Method to apply the Schema Matching in the Database System rebuilding.

[†]Rika Kashima, Akihiro Sato, Norimitsu Nagashima
Information Technology R&D Center, Mitsubishi Electric Corporation

3.3 マッチング結果格納方式

方式案2は、テーブルの定義は日々行われることはない想定し、あらかじめ全テーブル同士、つまり全カラム同士のマッチングをした結果をデータベースに格納しておき、ユーザからの処理要求時には、データベースに格納されたマッチング結果を参照するようにする方式とする。

スキーママッチングの結果は、図1で示したように、2つのテーブルのカラム間の類似度を示すマトリックスであるが、これを2つのカラムの各名称と類似度を1レコードにしてデータベースに格納しておく。イメージを図3に示す。

類似するカラム名の推薦をするときには、カラム1に指定されたカラム名、もしくはその一部を持ち、かつ類似度が指定値以上のものを選択する。

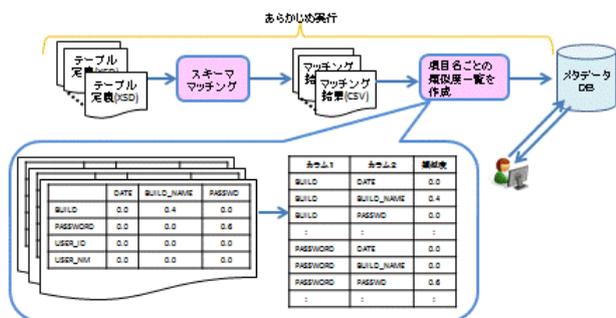


図3 方式案2

3.4 評価

設計支援ツールへの適用を想定しているため、ユーザツールのパフォーマンスに関する記事[3]を参考に、計算時間としてユーザが待ちきれなくなる10秒を応答性能の要件と定義し、マッチングの処理時間の評価を行った。

方式1のように、カラム数1のテーブル定義のXSDファイルとカラム数nのテーブル定義のXSDファイルを入力とし、nを変えスキーママッチングを実行した。表1に、結果を示す。

表1 マッチング性能

n	10	100	1000	2000	3000
時間(sec)	0.46	0.54	2.90	9.78	21.71

方式1は、スキーママッチングの起動が1回となり起動、後処理のオーバーヘッドが減るが、ユーザからの処理要求時に全テーブルの全カラムの個数分のマッチング処理を行うことには変

わりない。評価環境では、対象カラムの総数が2000までの場合応答性能要件を満たせるが、それを超えると要件を満たせなくなる。このため、大規模なシステムでリアルタイムの応答性能の要件を満たすためには、仮想テーブル(A)をサブシステム毎や何かのグループ毎に作成するなどの工夫が必要となる。

次に、方式2の評価として、データベースにマッチング結果を格納し、そこからある値以上の類似度を持つカラムを検索する処理を測定した。結果を表2に示す。表3に示すように今回の評価環境はサーバではなくPC、データベースにPostgreSQLを使ったが、レコード数が題材とするシステムのカラム数と同じ2万で0.02秒の応答時間を得た。このことからスペックの高いDBサーバとOracleの実環境では十分な性能を得ることができるといえる。ただし方式2では全く異なる単語に対しては類似度が格納されていないためヒットしないという問題点もある。

これらより、設計支援の内容により方式1と2を組み合わせる必要があると考える。

表2 検索性能

レコード数(万件)	2	10	100	200	400
時間(msec)	21	678	5664	11117	31961

表3 測定環境

プロセッサ	Intel Core i3 CPU 3.2GHz×2
メモリ	4GB
OS	Windows7
データベース	PostgreSQL

4. おわりに

今回の検討した方式のメタデータによるデータ統合における設計支援への適用を計画している。今後はこの実装開発を通し、技術適用効果の検証を行っていく予定である。

参考文献

- [1] 鹿島他 「辞書構築技術適用によるDBシステム再構築」情報処理学会第76回大会 5B-1
- [2] 小出他 「学習データ量によるスキーママッチング精度向上効果評価報告」情報処理学会第74回大会 6B-4
- [3] 「Excel 2010 のパフォーマンス：計算パフォーマンスの強化」 [http://msdn.microsoft.com/ja-jp/library/office/ff700515\(v=office.14\).aspx](http://msdn.microsoft.com/ja-jp/library/office/ff700515(v=office.14).aspx)