

# 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

小出健太<sup>†</sup> 柴光輝<sup>†</sup> 谷垣宏一<sup>†</sup> 高山茂伸<sup>†</sup>

三菱電機株式会社 情報技術総合研究所<sup>†</sup>

## 1.はじめに

企業内情報システムは、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ることがある。このようなシステムでは、データの分散が生じていることが多く、システムの運用・保守コストの増大を招いている。そのため必要となるデータの統合作業には、データの対応関係洗い出しに多大なコストが必要となる。データ統合を省力化するため、データ項目の対応関係を自動推定する方式が存在し、既存の方式では属人的スキルに頼った作業が課題である。本稿では、属人的スキルに拠っていた作業を学習によって自動化する方式の提案により、課題の解決を図った。それと共に、実用化に向けた学習用データ量の参考値について評価し報告する。

データ項目の対応関係を推定する手法はスキーママッチングと呼ばれ、しばしば複数の個別類似度判定手法(個別マッチャ)のスコアを組み合わせることで、対応関係の有無を判定する<sup>[1]</sup>。例として、COMA(COMA++)<sup>[2][3]</sup>でも複数の個別マッチャの組み合わせから判定を行う。これら従来の手法では、個別マッチャの組み合わせを人手で実施するため、作業人員に高いスキルが求められ、人員確保に伴うコストが課題となる。COMAにおいても、スキーママッチングの判定結果を元に個別マッチャの組み合わせを試行錯誤することで精度を高めるため、同様の課題がある。そこで本稿ではベイジアンコンポジットマッチャという方式によって、属人的スキル依存の少ないスキーママッチング手法を提案する。

## 2.ベイジアンコンポジットマッチャ

本開発方式では異なるスキーマ間におけるデータ項目の対応有無を判定するため、複数の個別マッチャのスコアの組み合わせをベイジアンネットワーク<sup>[4]</sup>を用いて自動的に行う事で、

属人的スキル依存の少ないスキーママッチング手法を実現した。この方式はベイジアンコンポジットマッチャ(BCM)と呼ぶ。BCMの特徴は、個別マッチャのスコアの組み合わせを自動化した事に加え、図1に示すように従来ではマッチングの結果を見ながら個別マッチャの組み合わせを試行錯誤するという属人的スキル依存の作業に対し、過去のデータ項目の対応関係を学習することによる個別マッチャのスコアへの重み付けで代替する点にある。この特徴には2つのメリットが存在する。一つは過去のデータ項目対応関係を活用することで新たなデータを用意する必要がないため、学習用データを用意するにあたってのスキルやコストを抑えることができる。もう一つは新たなシステムでの対応関係を学習データとして再利用していくデータライフサイクルを構成することで学習データの更新や拡充の手間を抑えることができる。

以上のように、ベイジアンネットワークでの学習によるスキーママッチングの自動化を行ったが、実用上において必要な精度を出すための学習データ量が許容範囲であるかどうかは不明である。そこで評価を行い、必要となる学習データ量の参考値を得る。

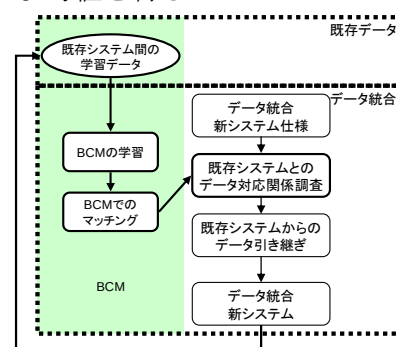


図1 BCMの特徴

## 3.評価

### 3.1 評価条件

評価データはCOMAと同じく [www.biztalk.org](http://www.biztalk.org) のXMLスキーマデータデータを使用する。評価スキーマのデータ項目対応関係の正解情報は、Leipzig大学より取得し、COMAと同条件としている<sup>[5]</sup>。評価スキーマの属性値を表1に示す。

The schema matching accuracy improvement effect evaluation report by study data volume

<sup>†</sup>Kenta Koide, Mitsuteru Shiba, Koichi Tanigaki, Shigenobu Takayama · Information Technology R&D Center Mitsubishi Electric Corp.

表 1 評価スキーマ属性値一覧

| No. | スキーマ名   | ノード<br>/パス数 | ノード数<br>/リーフ数<br>/共有ノード数 | MAX<br>/AVG<br>パス長 |
|-----|---------|-------------|--------------------------|--------------------|
| 1   | CDIX    | 27/34       | 7/20/7                   | 4/2.9              |
| 2   | Excel   | 32/48       | 9/23/11                  | 4/3.5              |
| 3   | Noris   | 46/65       | 8/38/18                  | 4/3.2              |
| 4   | Paragon | 59/77       | 11/48/13                 | 6/3.6              |
| 5   | Apertum | 74/134      | 22/52/24                 | 5/3.6              |

評価には、表 1 のデータを表 3 に示すように評価用スキーマペアと学習用スキーマペアに分けて行う。学習用スキーマペアの組み合わせを変える事で学習用データに含む項目ペアの量を変えて、精度がどのように変化するかを確認する。学習用スキーマペアとしては評価用スキーマペアのスキーマを除いた組み合わせを用いた。具体的な組み合わせを表 3 に記載している。

本評価では BCM で使用する個別マッチャとして表 2 に示す 20 種の手法を用いる。

表 2 個別マッチャ一覧

| Lexical Match<br>(名前判断)                                                                                                                                             | Synonym Match<br>(辞書判断) | Structure Match<br>(構造判断)                                                                 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|-------------------------------------------------------------------------------------------|
| AffixName, BiGramDistance, CosineSimilarity, HammingDistance, JaroMeasure, LevenshteinString, PathName, PrefixName, QuadGramDistance, SubStringDistance, SuffixName | Synonym,                | TriGramDistance, Name, ChildName, ChildPath, LeafName, LeafPath, SiblingName, SiblingPath |

### 3.2 評価結果

本評価の結果を表 3 および図 2 に示す。評価からは学習データ量の増加に従って精度向上効果が見られた。未学習時 60%の精度は、1,2+1,4 の学習データを用いることで 79%(最大値)までの向上が見られた。ただし、学習を行うデータのドメインによって精度向上効果にばらつきが見られる。特に、スキーマペア 1,2 の学習データを用いた場合に、無学習の場合と比較して精度が下がる結果を得ている。この点に関しては、今回の学習を最尤推定による過学習の影響と思われる、MAP 推定など他の手法の適用によって改善される可能性がある。

表 3 評価結果

| 組み合わせ          | ペア数  | 最良 F 値 | 適合率   | 再現率   | 最良閾値   |
|----------------|------|--------|-------|-------|--------|
| 1,2            | 2160 | 0.451  | 0.468 | 0.435 | 0      |
| 1,4            | 3200 | 0.750  | 0.800 | 0.706 | 0.1    |
| 2,4            | 4320 | 0.636  | 0.727 | 0.565 | 0.1    |
| 1,2+1,4        | 5360 | 0.790  | 0.861 | 0.729 | 0.1    |
| 1,2+2,4        | 6480 | 0.658  | 0.729 | 0.600 | 0.1    |
| 1,4+2,4        | 7520 | 0.745  | 0.838 | 0.671 | 0.1    |
| 1,2+1,4+2,4    | 9680 | 0.769  | 0.845 | 0.706 | 0.1    |
| 学習なし<br>(単純平均) | -    | 0.597  | 0.563 | 0.635 | 0.9692 |

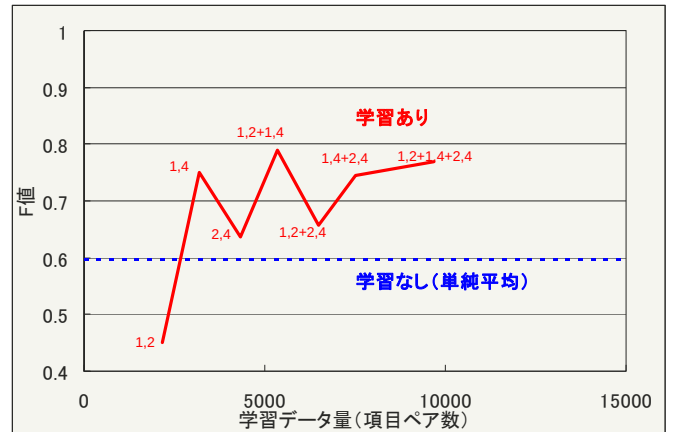


図 2 評価結果

学習用データ量の増加による精度向上効果のばらつきは、学習用データの増加に従って減衰し、10,000~15,000 件ほどの学習用データ量によって学習効果による精度向上効果のばらつきがほぼ収束することがわかった。収束までに必要な学習量はデータベースに換算して 100~120 カラムを有する 2 テーブル間の連携に相当し、過去のデータ項目対応関係から用意できる学習用データの期待量として許容範囲内であると考察する。

### 5. まとめ

個別マッチャの組み合わせ手法の問題に対し、BCM という手法の提案によって解決した。この BCM に求める精度は、実使用に耐えうる学習データ量で得られるのかについて評価を行った。評価結果から、実使用において必要な学習データ量の参考値は、実使用に耐えうるものであることを確認した。

### 参考文献

- [1] Erhard Rahm, Philip A. Bernstein : A survey of approaches to automatic schema matching (2001)
- [2] Hong-Hai Do, Erhard Rahm : Matching large schemas: Approaches and evaluation, ScienceDirect Information Systems 32, P857-885 (2007)
- [3] Do, H.H. and Rahm, E. : COMA - A System for Flexible Combination of Schema Matching Approaches, Proc. 28th Intl. Conference on Very Large Databases, VLDB (2002)
- [4] Lauritzen, S. L. The EM algorithm for graphical association models with missing data, Computational Statistics and Data Analysis 19: 191-201 (1995)
- [5] Sabine Massmann, Daniel Engmann, Erhard Rahm: COMA++: Results for the, Ontology Alignment Contest OAEI 2006, <http://oei.ontologymatching.org/2006/>