6E-05

データの組合せ可能性評価手法の提案とプロトタイプの開発

鬼頭　大介†　　屋代　聡†　　北原　圭†

(株)日立製作所　研究開発グループ　システムイノベーションセンタ†

1. はじめに

近年、政府や自治体等の公共機関が保有する公共データをオープンデータとして広く民間に公開する動きが活発化して来ている。企業や民間が持つデータとオープンデータを融合させることで新たなサービスの創出につながる可能性があり、経済活性化の効果が期待されている[1]。多様なデータを組合せるにあたり、従来はどのデータを組合せれば価値がありそうか、まず人間が仮説を立てて判断してから組合せるといったアプローチが多くとられていた[2]。しかしながら、オープンデータ等の膨大なデータを取り扱う場合、人手でデータの組合せ可否を判断する手法では負荷が高く、網羅できる組合せに限界があるという課題がある。上記課題に対し、本研究では、データの組合せ可能性を機械的に判断する手法を提案する。本稿では、提案手法の概要及びプロトタイプを実装して評価した結果について示す。

2. 従来技術と課題

大量データの探索において、従来の CRISP-DM 等の手法では、以下の手順が取られている。

①ビジネスの把握：事業上の目標や仮説の立案。
②データの準備：①に適合するデータを揃える。
③評価：分析モデルを構築し、仮説を検証。
④導入：仮説が有効な場合、事業に導入。

従来手法では、人間による仮説を前提とするため、網羅可能なデータ組合せには限界があり、また新サービスの発想には繋げにくい特徴がある。これに対し本研究では、

①’組合せて利用可能なデータを収集する。
②’①’のデータをもとに新サービスを発想。

というように上記手順①と②を変更する。本稿では、上記①’を支援する手段として、データの組合せ可能性を機械的に評価する手法の詳細について説明する。

Proposal of a data combination evaluation method and its prototype development
†Hitachi, Ltd., Research & Development Group, Center for Technology Innovation-Systems Engineering

3. データの組合せ可能性評価手法の提案

3.1 データ間の関係

データを組合せるとは、異なる二つのデータを関連付けて一纏まりのデータを作成することを意味する。例えば図 2-1に示すように、互いを関連付ける項目をキーとしてデータを連結する等に相当する。組合せたデータからは、単一のデータよりも多くの情報を得ることが可能になる。本研究では、オープンデータとして広く利用されている csv ファイルを対象として、データの組合せ可能性評価を行う。

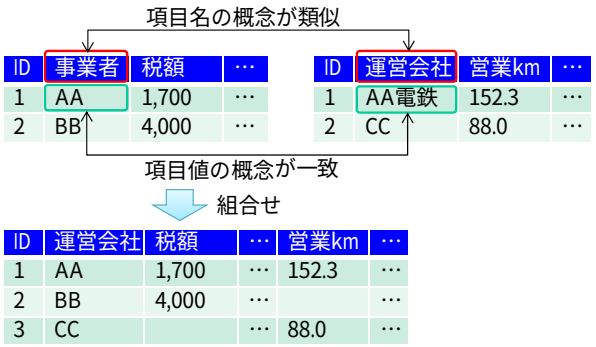

図 2-1 データの組合せ例

単純に値が一致する場合のデータ組合せについては、レコード統合等としてこれ迄に様々な検討がなされている[3]。本研究では、完全な一致以外に、共通部分が少しでもある類義や包含関係のものについても、データを組合せられる可能性があると考える。また、値が同じものを組合せる以外に、値が異なるものを組合せるといった組合せパタンも考えられる。例えば所定の項目について、異なる項目値を組合せてリスト化する等である。以上を踏まえ、互いを関連付ける項目を軸として値が一致するデータの組合せ以外に、値が一致しないデータの組合せも含めてデータの組合せ可能性を評価する。

3.2 データの組合せ可能性評価処理

提案するデータ組合せ可能性評価処理の概要を図 2-2 に示す。まず、組合せ可能性を評価する二ファイル間において、項目名レベルで一致や類似等の関係の有無の判定を行い、次に、関係があったものに対して、項目値レベルで同様

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.

に判定を行う。尚、項目名とは図 2-1 の例では、「事業者」等のカラム名が該当し、項目値とは上記項目の値、例えば、「AA」等が該当する。

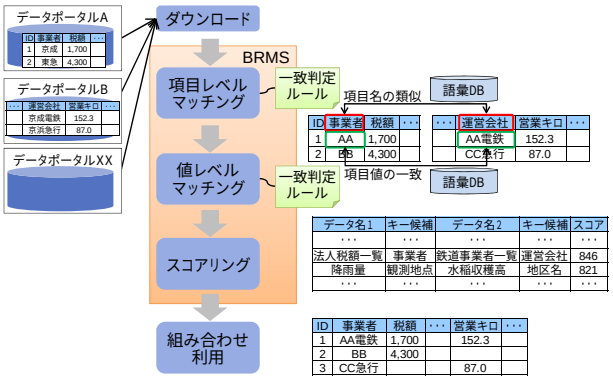

図 2-2 データ組合せ可能性評価処理の概要

上記の判定結果をもとに、ファイル間の組合せ可能性をスコアとして算出し、スコアの高いものほど組合せ可能性が高いと判断する。スコア付けは以下の表 2-1 に示す方針で行う。

表 2-1 スコア付け方針

#	内容
1	項目名及び項目値が、一致、類義、包含の何れかの関係を満たす場合にスコアを加算する。スコアは、一致＞類義＞包含 の順に高くする 【理由】類義や一致のものほどデータの関連付けを行い易いため
2	1を満たす項目名が多いほど、高スコアにする。但し、1を満たす項目名が3個以上の場合は、それ以上スコアを加算しない 【理由】組合せの軸として、大きく位置情報、時刻情報、その他の情報、の3種類が考えられ、それ以上の項目があっても組合せに使う可能性は低いため
3	1を満たす項目値の個数が多いほど、高スコアにする 【理由】時系列の情報など、データ数が多いほど意味のある情報になるものが存在するため
4	一方のファイルの項目名が他方のファイルの項目名と全て一致し、その項目値の一致割合も高いものは、スコアを加算しない 【理由】ファイル間で共通部分が多すぎる場合、同一のファイルまたは一方を更新しただけのファイルの可能性があり、組合せても意味のある情報が得られる可能性は低いため

データ組合せ可能性評価処理を行う上で必要なシステム構成について説明する。項目が類義や包含の関係にあるかを判断可能にするため、内部に定義した語彙情報を活用することとする。語彙情報は、例えば「氏名」と「名称」等、類似の用語を定義した類義語情報と、「都道府県」とその配下の「東京都」「神奈川県」等、用語間の包含関係を区別するためのカテゴリ情報で構成される。

また、組合せ可能性評価においては、表 2-1 のルールに従ってスコアを加算するが、ルールは今後追加や、評価対象とするデータに応じて変更するなど、様々な修正を施す可能性がある。そのため、ルール変更を容易に行える仕組みとすることが望ましい。上記を実現するため、ルールをアプリケーションから分離して管理・開発できるようにする。具体的には、ルールをプログラム内に定義するのではなく、BRMS（Business Rule Management System）の仕組みでプログラムと分けて管理する。

3.3 評価結果

データ組合せ可能性評価の動作を検証するため、プロトタイプの試作を行った。尚、横浜市の金沢区が公開するオープンデータ[4]を利用してデータ組合せ可能性の評価を行った。評価結果を以下の表 2-2 に示す。

表 2-2 データ組合せ可能性の評価結果

評価対象ファイル	重み付け (1)		重み付け (2)		重み付け (3)	
	スコア	順位	スコア	順位	スコア	順位
①	63	1	16.0	1	63	1
②	15	2	15.7	2	0	2
③	6	3	12.4	3	0	2
④	1	4	7.5	4	0	2
⑤	0	5	0	5	0	2

①～⑤は、地域の施設情報や家庭保育福祉員の情報等を格納したファイルであり、それぞれについて、地域の公園情報のファイルとの組合せ可能性を評価した。重み付け(1)は、表 2-1 のルールをもとに算出したスコアを示し、重み付け(2)(3)は、表 2-1 の#3 のルールの一部を変更（（2）項目値の一致割合が高いほど高スコア、（3）項目が主キーのもののみスコアを加算）して算出したスコアを示す。重み付け（1）と（2）では組合せ可能性順位が同じである一方、重み付け（3）ではファイル①のみが組合せ可能性ありと判断されている。重み付け方法によって若干の順位の違いが生じてくることから、重み付け方法の調整も重要であることがわかる。

4. まとめ

オープンデータを対象として、データの組合せ可能性の判断を機械的に行えるようにする手法の立案及びプロトタイプ開発を行った。組合せ可能性評価においては、スコアの重み付け方法が評価結果に影響を与えることが明らかになり、今後、様々なデータを題材として重み付け方法の調整を行い、さらなる手法改善を図る。

参考文献

[1] 平成 24 年度電子経済産業省構築事業（空間位置情報に関連する公共データの活用実証事業）調査報告書
[2] Chapman, P. et al., CRISP-DM 1.0 Step-by-step data mining guide, CRISP-DM Consortium, 2000.
[3] 相澤他, 「レコード同定問題に関する研究の課題と現状」, 電子情報通信学会論文誌 Vol.J88-D-I, 2005/3
[4] 横浜市金沢区データポータルサイト, http://www.city.yokohama.lg.jp/kanazawa/kz-opendata/kz-opendata.html

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.