pitney bowes

Spectrum™ Technology Platform

バージョン 2019.1.0

Smart Data Quality ガイド

目次

1 - はじめに

2 - 検索条件の生成

1 - はじめに

このセクションの構成

Smart Data Quality の概要

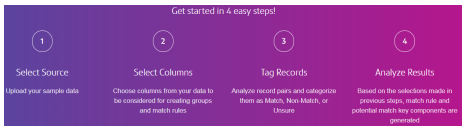
Spectrum™ Technology Platform Smart Data Quality は *Machine Learning* を利用したソリューションであり、エンティティ解決処理のための最初のマッチ ルールとマッチ キー コンポーネント候補を作成するのに役立ちます。データ*品質*処理に Machine Learning 機能が加わったことで、マッチング手順が大幅に簡素化され、データの潜在的能力を最大限に活用できます。

マッチングのアルゴリズムとしきい値は、ユーザのマッチング シナリオに基づいて自動的に学習されます。最初のマッチ ルールとマッチ キー コンポーネント候補は、入力およびタグ付けの指定によって生成されます。

このシステムを使用してマッチ ルールとマッチ キー コンポーネントを生成するには、*サンプル* データをアップロードします。サンプル データは、レコードのあらゆるバリエーションを網羅したコレクションでなければなりません。その後、マッチングを実行する列を選択し、それらを大ざっぱなグループにまとめて、異なるバリエーションのレコード ペアを選択できるようにします。さらに、各自のマッチング シナリオに従ってレコードにタグ付けし、サンプル データを通じて学習されたマッチ ルールと共にマッチ キー コンポーネント候補を取得します。

マッチ ルールとマッチ キー コンポーネント候補を生成するための手順については、「タスク フロー」とそれ以降のセクションを参照してください。

タスク フロー



1. 最初に、*ソースからファイルを選択*します。選択したファイルには*サンプル* データが含まれている必要があります。このサンプル データは、レコードのあらゆるバリエーションを実際に網羅して表すものでなければなりません。
2. サンプル データをアップロードした後、データからマッチングを実行する**列を選択**します。このステップで選択した列は、次のステップで**グループの生成**に使用します。
3. このステップでは、**グループ強度**の値を指定します。この強度は、類似のレコードのグループを生成するのに役立ちます。次のステップに進む前に、生成されたグループを確認してください。
4. グループを生成して確認した後、表示されたレコード ペアに **[一致]**、**[アンマッチ]**、**[不確定]** のようにタグ付けします。これらのタグは、正確なマッチ ルールとマッチ キー コンポーネント候補を生成するのに役立ちます。

5. 最後に、生成された結果を表示して分析します。確認が終わったマッチ ルールは **Enterprise Designer** でマッチ *ルール リポジトリ*にエクスポートし、マッチング *ステージ*で使用できます。マッチ ルールの詳細については、マッチ ルールを参照してください。

   確認が終わったマッチ キー コンポーネントは、**Enterprise Designer** の **Match Key Generator** ステージで使用できます。Match Key Generator の詳細については、マッチ キーの定義に関するテクニックを参照してください。

ログイン

以下では、Web ブラウザを使用して **Spectrum™ Smart Data Quality** にアクセスする手順を示します。

1. Web ブラウザを開きます。
2. `http://server:port/data-quality` という URL にアクセスします。ここで、*server* は Spectrum™ Technology Platform サーバーの名前または IP アドレスで、*port* は HTTP ポートです。デフォルトの HTTP ポートは 8080 です。
3. 有効な**ユーザ名**と**パスワード**を入力します。
4. **[サイン イン]** をクリックします。

   Smart Data Quality ホーム ページが表示されます。**[はじめに]** ボタンをクリックして、新しいプロジェクトを作成するか、**[プロジェクト]** タブをクリックして、既に作成されているプロジェクトとその進捗のリストを表示します。

2 - 検索条件の生成

このセクションの構成

プロジェクトの作成と表示

検索条件の生成を開始するには、プロジェクトを作成する必要があります。このセクションでは、新規プロジェクトを作成し、作成済みのプロジェクトを表示する方法について説明します。

新しいプロジェクトの作成

*新しいプロジェクト*を作成するには、次の手順を実行します。

1. **[Smart Data Quality]** ホームページで、**[はじめに]** ボタンをクリックします。
   [プロジェクトの作成] ページが表示されます。
2. **[プロジェクト名]** および **[プロジェクトの説明]** を入力します。
3. **[保存]** ボタンをクリックしてプロジェクトを **[プロジェクト]** ページに表示するか、または **[保存して続行]** ボタンをクリックして、次のステップに進みます。

プロジェクトの表示

作成済みのプロジェクトおよびその進捗を表示するには、**[Smart Data Quality]** ホームページの **[プロジェクト]** タブをクリックします。**[プロジェクト]** ページに以下の詳細が表示されます。

- **[プロジェクト名]** - 入力したプロジェクト名
- **[プロジェクトの説明]** - 入力したプロジェクトの説明
- **[作成者]** - プロジェクトの開始したユーザ
- **[最終更新]** - プロジェクトが最後に更新された日時
- **[ソース]** - アップロードされたソース ファイルの名前
- **[進捗]** - プロジェクトの現在のステータス (*[ソースを選択]*、*[列の選択]*、*[グループの生成]*、*[レコードのタグ付け]*、または *[マッチ ルールと生成されたキー]*)

[追加] アイコンをクリックすると、新しいプロジェクトを作成できます。**[編集]** または **[削除]** アイコンをクリックすると、任意のプロジェクトを編集または削除できます。

> **注 :** プロジェクトを表示して、現在のステージから続行するには、**プロジェクト名**をクリックします。

ソースからのファイルのアップロード

一致条件を生成するには、*サンプル データ*をアップロードする必要があります。*サンプル データ*は、*マッチ*、*アンマッチ*、*重複*、*ユニーク*、*各種フィールドについて視覚的に同じまたは異なるフィールドの両方*など、数値的な多様性を持つすべてのデータを実際に表現したものでなければなりません。

以下に、ファイルのアップロード手順を示します。

1. **[ソースを選択]** ページで、  アイコンをクリックしてデータ ファイルが置かれているパスに移動します。
2. **[OK]** ボタンをクリックします。
   データ ファイルが **[データ プレビュー]** セクションにプレビュー表示されます。
3. アップロードされたデータに従って、**[文字エンコード]**、**[フィールド区切り文字]**、**[テキスト修飾子]**、**[ライン区切り文字]** の各フィールドがあらかじめ設定されます。必要な場合、これらはユーザによって上書きされます (次の表を参照)。

フィールド名	説明
文字エンコーディング	テキスト ファイルのエンコーディング。次のいずれかを選択します。 テキスト ファイルのエンコーディング。次のいずれかを選択します。 **CP1252** このエンコーディングは Windows-1252 文字セット、または単に Windows 文字セットとも呼ばれています。これは ISO-8859-1 の上位クラスであり、128 ～ 159 のコード範囲を使用して、ISO-8859-1 文字セットに含まれていない追加の文字を表示します。 **UTF-8** すべての Unicode 文字をサポートし、かつ ASCII との下位互換性があります。UTF の詳細については、**unicode.org/faq/utf_bom.html** を参照してください。 **UTF-16** すべての Unicode 文字をサポートします。しかし、ASCII との下位互換性はありません。UTF の詳細については、**unicode.org/faq/utf_bom.html** を参照してください。 **US-ASCII** 英語のアルファベット順に従う文字エンコーディング。 **UTF-16BE** ビッグエンディアン UTF-16 エンコーディング (下位アドレスが上位バイトとなるようにシリアル化)。 **UTF-16LE** リトルエンディアン UTF-16 エンコーディング (下位アドレスが下位バイトとなるようにシリアル化)。 **ISO-8859-1** 主として西ヨーロッパの言語で使われる ASCII 文字エンコーディング。Latin-1 とも呼ばれます。 **ISO-8859-3** 主として南ヨーロッパの言語で使われる ASCII 文字エンコーディング。Latin-3 とも呼ばれます。 **ISO-8859-9** 主としてトルコ語で使われる ASCII 文字エンコーディング。Latin-5 とも呼ばれます。 **CP850** 西ヨーロッパの言語を書くための ASCII コード ページ。 **CP500** 西ヨーロッパの言語を書くための EBCDIC コード ページ。 **Shift_JIS** 日本語のための文字エンコーディング。 **MS932** NEC 特殊文字、NEC 選定 IBM 拡張文字、IBM 拡張文字を含めた Microsoft の拡張版 Shift_JIS 文字コード。 **CP1047** Latin-1 文字セット全体を含む EBCDIC コード ページ。

フィールド名	説明
フィールド区切り文字	区切り記号付きファイル内のフィールドを区切るのに使用する文字を指定します。 例えば、次のレコードではパイプ (|) がフィールド区切り文字として使われています。 `7200 13TH ST|MIAMI|FL|33144` フィールド区切り文字として使用可能な文字は次のとおりです。 • カンマ • セミコロン • パイプ (|) • タブ • スペース • ピリオド (.)
Text qualifier	区切り記号付きファイル内のテキスト値を囲むのに使用する文字。 例えば、次のレコードでは二重引用符 (") がテキスト修飾子として使われています。 `"7200 13TH ST"|"MIAMI"|"FL"|"33144"` テキスト修飾子として定義できるのは次の文字です。 • 一重引用符 (') • 二重引用符 (")
ライン区切り文字	順次ファイルまたは区切り記号付きファイル内の行 (ライン) を区切るのに使用する文字を指定します。 使用できるライン区切り文字の設定は次のとおりです。 **Unix** 改行 (LF) 文字でラインを区切ります。これは Unix システムの標準のライン区切り文字です。 **Macintosh** 復帰 (CR) 文字でラインを区切ります。これは Macintosh システムの標準のライン区切り文字です。 **Windows** 復帰改行 (CR+LF) でラインを区切ります。これは Windows システムの標準のライン区切り文字です。

4. 最初の行をヘッダーと見なすかどうかを **[はい]** または **[いいえ]** のスライディング ボタンによって選択します。その選択に応じて**データ プレビュー**が変化します。

5. Save and Continue > アイコンをクリックすると、変更が保存され、次のステージに移動します。
6. 現在のタスクをキャンセルするには、Cancel アイコンをクリックします。

列の選択

このセクションには、サンプル データの列が表形式で表示されます。マッチングを実行したい列を選択する必要があります。

以下の手順は、グループの作成や一致条件の生成のために列を選択する方法を示しています。

1. **[意味型の検出]** ボタンをクリックして、すべての列の意味型を検出します。検出された意味型が表示されます。デフォルトでは、**なし**が表示されます。

   目的の意味型が表示されない場合は、その列の対応するチェックボックスをオンにした後、ドロップダウンから意味型を選択できます。

   > **注 :** 正確な一致条件を生成するために、このステップを実行することをお勧めします。選択した意味型に基づき、関連するアルゴリズムが一致条件の生成のために使用されます。例えば、*発音アルゴリズム*は*名前*の意味型で使用され、*電話番号*および*郵便番号*の意味型では使用されません。

2. グループの作成や一致条件の生成のために選択する列の **[列名]** チェックボックスをオンにします。
3. それぞれの列にある Null 値の扱い方は、**[Null の処理]** ドロップダウンで選択できます。**[Null はマッチ]** と **[Null はアンマッチ]** のどちらかを選択できます。このオプションのデフォルト値は**[Null はマッチ]** です。このオプションを選択した場合、空のフィールドはレコード ペアの対応するフィールドに等しいと見なされます。**[Null はアンマッチ]** を選択した場合、空のフィールドはレコード ペアの対応するフィールドに等しくないと見なされます。ここで行った選択は、**Enterprise Designer** でのマッチ ルールの*欠落データ*オプションに反映されます。**[Null はマッチ]** を選択した場合は **[100 としてカウント]** が事前に選択され、**[Null はアンマッチ]** を選択した場合は **[0 としてカウント]** が事前に選択されます。

   > **注 :** このオプションは、フィールドに対してグローバルに適用され、フィールドのさまざまな条件に対しても変わらず維持されます。

4. Save and Continue > アイコンをクリックすると、変更が保存され、次のステージに移動します。
5. 現在のタスクをキャンセルするには、Cancel アイコンをクリックします。

グループの生成

このセクションでは、類似レコードのグループの生成および確認方法について説明します。これらのグループは、指定した **[グループ強度]** に基づいて生成されます。低い強度を指定すると、関係性の薄い類似レコードも同じグループとして分類されるので、緩い関係の大きなグループが生成されます。強度を高くすると、関係性が強く、精度の高いグループが生成されます。グループ強度を非常に高い値に設定すると、どのレコードも同じグループに分類されず、単一のエントリを持つ複数のグループが生成されます。

グループを生成するには、次の手順を実行します。

1. **[グループの生成]** ページで、スライダー バーを使用するか、または用意されているテキスト ボックスに強度を入力して **[グループ強度]** を指定します。

   注：[グループ強度] には、*0* ～ *1* の値を指定できます。小数点以下第 2 位まで値を入力できます。

2. Generate Groups ボタンをクリックします。
   生成されたグループがテーブル形式で表示されます。**生成されたグループの合計数、平均サイズ、最小サイズ、最大サイズ**など、グループに関する追加情報が表示されます。**[一意のグループ]** の数も表示されます。

   注： 生成されたグループのリストから、単一のレコードを含むグループを非表示にするには、**[一意のグループを非表示にする]** チェック ボックスをオンにします。

3. 生成されたグループを確認します。

   注： 生成されたグループを確認する際には、エンティティのグループ化したい類似のエントリが、ほぼ同じグループに分類されているか確認します。生成されたグループに満足できない場合は、グループ強度を変更して、グループを再生成します。完全に正確なグループを生成する必要はありません。これらのグループは次のステップへの入力として使用されます。次のステップでは、関連するレコードのペアが表示され、タグ付けを行うことができます。

   例: エンティティ **[名]** をグループ化したい場合、生成されたグループのほとんどに似たタイプの名が分類されているか確認します。適切に分類されていない場合は、**[グループ強度]** を変更してグループを再生成します。

4. Save and Continue > アイコンをクリックすると、変更が保存され、次のステージに移動します。

5. 現在のタスクをキャンセルするには、 

アイコンをクリックします。

レコードのタグ付け

このページでは、レコードにタグ付けを行う必要があります。表示されるレコードは、類似性に基づいてペアになっており、あらかじめタグが付けられています。タグを確認し、画面の右側にある **[一致]**、**[アンマッチ]**、または **[不確定]** ボタンをクリックしてタグを変更する必要があります。

> **注：** あらかじめ設定されているタグは参考までに付けられているものなので、十分に確認してください。

簡単にタグ付けを行えるように、以下のオプションを使用できます。

バルク アクションの実行

詳しく確認した後、**[バルク アクション]** オプションを使用して、すべてのページの複数のレコード ペアに対して、**[一致]**、**[アンマッチ]**、または **[不確定]** として一括でタグを付けることができます。このアクションを実行するには、次の手順に従います。

1. すべてのページで、用意されている各チェック ボックスを使用して、タグ付けする複数のレコード ペアを選択します。
2. **[バルク アクション]** ドロップダウンをクリックして利用可能なオプションから選択します。
3. **[適用]** ボタンをクリックします。

> **注：** 引き続き手動でタグ付けを行うには、バルク アクションでチェックを付けたすべてのレコード ペアのチェック ボックスをオフにします。

フィルタの使用

[フィルタ] をクリックして適用できます。以下のフィルタを使用できます。

- **[すべて]** - すべてのレコード ペアを表示します。
- **[一致]** - [一致] のタグが付けられているレコード ペアを表示します。
- **[アンマッチ]** - [アンマッチ] のタグが付けられているレコード ペアを表示します。
- **[不確定]** - [不確定] のタグが付けられているレコード ペアを表示します。

タグ付けされたレコードの保存と非表示化

[タグ付けの保存] ボタンをクリックして指定したタグを保存し、以前のセッションでの作業の続きを行うことができます。また、**[タグ付けされたレコードの非表示化]** チェック ボックスをオンにすると、タグ付け済みのレコードを非表示にすることもできます。

> **注 :** 正確なマッチ ルールを生成するには、すべてのレコード ペアにタグ付けする必要があります。レコードへのタグ付けが適切でないと、生成される検索条件が不正確になる可能性があります。

結果の分析

[結果の分析] ページには、生成済みのネストした Boolean マッチ ルール、および提供された情報から獲得された潜在的なマッチ キー コンポーネントが表示されます。マッチ ルールは、確認後、**Enterprise Designer** の **[マッチ ルール管理]** オプションの *マッチ ルール* リポジトリにエクスポートし、バッチ ジョブで利用することができます。潜在的なマッチ キー コンポーネントは、確認後、**Enterprise Designer** の **[Match Key Generator]** ステージで使用できます。

[マッチ ルール] タブ

このタブは、2 つのセクションに分かれています。

- 左のウィンドウには、マッチ ルールが表示されます。**[ルール]** を展開すると、すべての条件および下位条件を表示できます。
- 画面の右のウィンドウには、これらの条件のプレビューがテーブル形式で表示されます。ここには、次のような詳細情報が表示されます。
  - **[しきい値]**、**[スコアリング方法]**、**[アルゴリズム]**、**[不足しているデータ メソッド]** などの **[属性]**。
  - これらの各属性の **[値]**。

生成されたマッチ ルールを確認した後、 Save Rule ボタンをクリックして、*マッチ ルール リポジトリ*にエクスポートできます。ポップアップ ウィンドウが表示されるので、**[ルール名]** を指定して、**[保存]** をクリックします。

ルールが保存されます。保存されたルールは、**Enterprise Designer** の **[マッチ ルール管理]** オプションで表示できます。

[マッチ キー] タブ

このタブには、潜在的なマッチ キー コンポーネントがテーブル形式で表示されます。また、マッチ キー コンポーネントが検出された **[列]** および使用される **[アルゴリズム]** も表示されます。潜在

的なマッチ キー コンポーネントは、確認後、シナリオに応じて **Enterprise Designer** の **Match Key Generator** ステージで追加することにより利用できます。

注 : 現時点では、*Substring* アルゴリズムのみがサポートされています。

例: この表には、**[電話番号]** 列で検出された潜在的なマッチ キー **[マッチ キー 1]** が表示されます。使用されるアルゴリズムは **SUBSTRING (1, 7)** です。ここで、 *1* は開始インデックス、*7* は終了インデックスで、**Match Key Generator** ステージのオプションに指定します。すべての潜在的なマッチ キー コンポーネントで、開始インデックスは *1* に固定されています。

マッチ キー	列	アルゴリズム
マッチ キー 1	電話番号	SUBSTRING (1, 7)

システムは、実行されたアクション (**[アップロードされたサンプル データに存在する派生形]**、**[マッチングで選択された列]**、**[グループ化のしきい値の設定]**、および **[タグ付けされたレコード]**) に基づき、データに存在するパターンを明らかにし、マッチ ルールと潜在的なマッチ キー コンポーネントを提供します。生成された結果をデータセットに対してテストすることをお勧めします。

著作権に関する通知

USPS® 情報

Pitney Bowes Inc. は、ZIP + 4® データベースを光学および磁気媒体に発行および販売する非独占的ライセンスを所有しています。CASS、CASS 認定、DPV、eLOT、FASTforward、First-Class Mail、Intelligent Mail、LACS^{Link}、NCOA^{Link}、PAVE、PLANET Code、Postal Service、POSTNET、Post Office、RDI、Suite^{Link}、United States Postal Service、Standard Mail、United States Post Office、USPS、ZIP Code、および ZIP + 4 の各商標は United States Postal Service が所有します。United States Postal Service に帰属する商標はこれに限りません。

Pitney Bowes Inc. は、NCOA^{Link}® 処理に対する USPS® の非独占的ライセンスを所有しています。

Pitney Bowes Software の製品、オプション、およびサービスの価格は、USPS® または米国政府によって規定、制御、または承認されるものではありません。RDI™ データを利用して郵便送料を判定する場合に、使用する郵便配送業者の選定に関するビジネス上の意思決定が USPS® または米国政府によって行われることはありません。

データ プロバイダおよび関連情報

このメディアに含まれて、Pitney Bowes Software アプリケーション内で使用されるデータ製品は、各種商標によって、および次の 1 つ以上の著作権によって保護されています。

© Copyright United States Postal Service.All rights reserved.

© 2014 TomTom.All rights reserved.TomTom および TomTom ロゴは TomTom N.V の登録商標です。

© 2016 HERE

Fuente: INEGI (Instituto Nacional de Estadística y Geografía)

電子データに基づいています。© National Land Survey Sweden.

© Copyright United States Census Bureau

© Copyright Nova Marketing Group, Inc.

このプログラムの一部は著作権で保護されています。© Copyright 1993-2007 by Nova Marketing Group Inc. All Rights Reserved

© Copyright Second Decimal, LLC

© Copyright Canada Post Corporation

この CD-ROM には、Canada Post Corporation が著作権を所有している編集物からのデータが収録されています。

© 2007 Claritas, Inc.

Geocode Address World データ セットには、**http://creativecommons.org/licenses/by/3.0/legalcode** に存在するクリエイティブ コモンズ アトリビューション ライセンス (「アトリビューション ライセンス」) の下に提供されている GeoNames Project (**www.geonames.org**) からライセンス供与されたデータが含まれています。お客様による GeoNames データ (Spectrum™ Technology Platform ユーザ マニュアルに記載) の使用は、アトリビューション ライセンスの条件に従う必要があり、お客様と Pitney Bowes Software, Inc. との契約と、アトリビューション ライセンスの間に矛盾が生じる場合は、アトリビューション ライセンスのみに基づいてそれを解決する必要があります。お客様による GeoNames データの使用に関しては、アトリビューション ライセンスが適用されるためです。

3001 Summer Street
Stamford CT 06926-0700
USA

www.pitneybowes.com