D_049

機械学習を用いたWeb上の表情報の例示検索方式

Query by Example of Web Information expressed Tabular Formulation

横川 智浩† 吉田 稔‡ 山田 剛一† 絹川 博之† 中川 裕志‡
Tomohiro Yokokawa Minoru Yoshida Koichi Yamada Hiroshi Kinukawa Hiroshi Nakagawa

1. はじめに

Web上には表の形式で構造化されている情報が多くある。従来の検索エンジンではユーザの質問は単語列であるが、これを表の形式で構造化することにより、Web上の表形式に構造化された情報に適した検索を行うことができる。そこで、検索対象をWeb上の表情報とし、ユーザの検索意図である情報内容を表形式で例示し検索する、表情報の例示検索方式を検討している。[1]

特定の分野のみを対象とした機械学習では、分野外の表情報の分類精度が高くならない。そこで、機械学習に使用する表を複数の分野で収集・統合し、複数の分野における表情報の例示検索方式の有効性を検証する。

このとき、機械学習に使用するフィーチャー（特徴）は、予備実験によって分類精度が高いものを選別した。

2. 例示検索方式

2.1 例示表の入力方式

ユーザは、あらかじめ用意した表形式の検索インタフェース（図1）に、検索を所望する表の例を入力することで、検索条件を与える。ユーザから与えられた単語と、その単語が入力された位置情報を条件として、Web上から表情報を検索する。このとき、ユーザが入力した表の例を例示表と呼称するものとする。

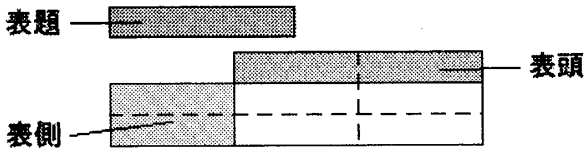

図1.表形式の検索インタフェース

検索インタフェースは、表の構造を表現できる最低限の構成として、表題を1つ、表頭（表の1行目）と表側（表の1列目）を2つずつ、表頭と表側に対応できるように表本体（2行目以降かつ2列目以降）を4つ設けた。表頭と表側には主に属性が、表本体には主に値が入力されることを想定している。

2.2 検索の実行

検索にはGoogleAPIを使用する。GoogleAPIに検索条件を与え、検索されたWebページのURLを取得する。[3]

GoogleAPIで例示表をそのまま使用することはできないため、検索インタフェースの構造を反映した検索条件式を例示表から生成し、それを用いて検索する。以下に、検索条件式の生成手順を示す。

1. 表題と表頭の検索単語をANDでつなぐ。これを(1)とする。
2. 表頭でない部分（表側と表本体）について、横方向に並ぶ検索単語をANDでつなぐ。これを(2)とする。
3. 表頭でない部分について、縦方向に検索単語が並ぶ場合、(2)をORでつなぐ。これを(3)とする。
4. (1)と(3)をANDでつなぐ。

ユーザの要求を『ヨーロッパかアジアのツアーの料金が知りたい』に設定した場合の入力例と、その検索条件式の生成手順を以下に示す。

図2のような入力があった場合、表題と表頭の検索単語（ツアー、料金）をANDでつなぐ。次に、ヨーロッパとアジアは縦方向に並ぶ検索単語なのでORでつなぐ。結果として、検索条件式は、『ツアー AND 料金 AND（ヨーロッパ OR アジア）』となる。

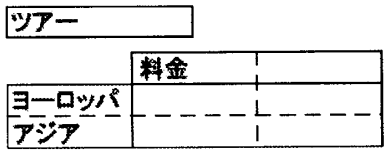

図2.検索インタフェースの入力例

2.3 検索結果から表情報の抽出

GoogleAPIによって検索されたWebページから表を抽出する。

Web上には、レイアウトを目的として使われている表タグが数多く存在するが、これらは検索の対象としない。具体例としては、別の表を包含しているもの、箇条書きを表で実現しているもの、などがあげられる。

このようなレイアウトを目的とした表を除去するために、入れ子になっている一番内側の表のみを検索の対象とする。さらに、箇条書きのレイアウトを実現するために使われている表を除去するために、1行（あるいは1列）の表を検索の対象から外す。

2.4 表情報の順序付け

検索された表情報を、よりユーザの要求を満たすもの

† 東京電機大学大学院工学研究科
‡ 東京大学情報基盤センター

が上位になるよう順序付ける。この順序付けには、SVM (Support Vector Machine)の機械学習によって生成された分類モデルを使用する。SVM は TinySVM を使用する。[4]

SVM の機械学習には、予備実験により有効と判断されたフィーチャーを使用する。これらのフィーチャーをカテゴリ別にグループ化したものを以下に示す。括弧内の数字はフィーチャーの数を表す。

- 例示表の構造特性(8)
- 画像(3)
- 段落(3)
- 書式(29)
- 表の形(6)
- リンク(3)
- 改行(12)
- 文字種(18)
- 文字数(12)
- フォーム(4)

特に例示表の構造特性とは、検索単語が入力された表セルの位置情報や、複数の検索単語間の相対的な位置情報などを利用したフィーチャーである。

上記の98個のフィーチャーを使用し機械学習を行う。機械学習によって生成された分類モデルが定義する『2クラス間の境界面』からの距離を用いて、表が検索意図に合致している度合いに沿う順序付けを行う。

2.5 結果の提示

順序付けた表情報をユーザに提示する。それぞれの表には付属情報としてタイトルと URL を表示する。

提示の際、画像やリンクの URL を相対 URL から絶対 URL に変換する。これは、画像の表示を正常に行うことや、リンクを利用できるようにするためである。

3. 評価実験

3.1 実験方法

Web 上の表を用いて、機械学習によって生成された分類モデルの正解・不正解の分類精度の検証を行う。このとき、例示表の構造特性の有効性を示すために、例示表の構造特性フィーチャーを排除しない場合と排除した場合で機械学習を行い、分類精度を比較する。

機械学習および検証に使用する表は、重複しない異なる8つの分野について、それぞれ1000個ずつの表を収集する。機械学習および検証は5-foldクロスバリデーションで行う。このとき、クロスバリデーションの1群は、各分野を5つに分割した200個の表を8つの分野で統合した1600個の表からなる。フィーチャーは2.4節に示したものを使用する。

表を収集する際に使用する例示表は、分野ごとにユーザの要求を設定し、その要求をもとに作成したものとする。GoogleAPI によって検索される上位のページから1000個を収集し、事前に正解と不正解を判別する。判別基準は、あらかじめ設定したユーザの要求を満たすものを正解とし、それ以外を不正解とする。例えば図2の例示表が与えられた場合、『ヨーロッパかアジアのツアーの料金が提示されているもの』を正解とする。

3.2 実験結果

表1に、機械学習によって生成された分類モデルの精度を示す。

適合率は、分類モデルに正解と判別された表のうち、正解が占める割合である。再現率は、正解と判別されるべき表が、実際に正解と判別された割合である。F-measure の計算式を以下に示す。

F-measure = 2×適合率×再現率÷(適合率 + 再現率)

表1. 生成された分類モデルの精度

例示表の構造特性フィーチャー	適合率	再現率	F-measure
あり	83.9%	82.6%	83.2%
なし	83.8%	69.0%	73.8%

4. 考察

3の評価実験において、例示表の構造特性フィーチャーを使用した場合、適合率と再現率とも82%を超えた。これは、正しく機械学習が行われていることを示している。

さらに、例示表の構造特性フィーチャーを使用した場合と、使用しなかった場合を比較すると、適合率はほぼ同じ値で、再現率において13.6%、F-measure において9.4%の差が出ていることが読み取れる。このことから、正解に分類されるべき表が不正解に分類されていることがわかり、例示表の構造特性を利用することの有効性を示すことができた。

5. おわりに

入力した例示表の情報から、Web 上の表を検索する方式を検討した。機械学習によって生成された分類モデルの分類精度の検証実験を行い、表情報の例示検索方式において、例示表の構造特性を利用することの有効性を示すことができた。

今後は、生成された分類モデルを使用し、例示検索精度の検証を行い、よりユーザの要求を満たす例示検索方式を目指す。

参考文献

[1] 横川智浩「Web 上の表情報の例示検索方式」 情報処理学会第68回全国大会, 1E-3, pp.3-109～3-110, March, 2003.

[2] C.J.DATE: An Introduction to Database Systems (Third Edition), ADDISON-WESLEY PUBLISHING COMPANY, February 1982

[3] Google: http://www.google.com/intl/ja/

[4] TinySVM: http://chasen.org/~taku/software/TinySVM/