1E-3

Web 上の表情報の例示検索方式

横川 智浩† 吉田 稔‡ 山田 剛一† 絹川 博之† 中川 裕志‡
†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

1. はじめに

Web には HTML で記述された膨大な数の表情報があり、これらは情報が構造化されているため良質な情報であることが多い。しかし、単語群を検索質問とする通常の全文検索では、個々の表が持つ構造的な情報は無視され、また、検索質問も構造化されていないため、表形式で記述されていることによるメリットが現れないという問題点がある。そこで、検索対象を Web の表情報とし、ユーザの検索意図である情報内容とその整理形式を表形式で例示し検索する、表情報の例示検索方式を提案する。

ユーザは例示検索する際、あらかじめ用意した表形式の検索インタフェース（図 1）に検索条件を入力する。そして、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し、よりユーザの要求を満たすものが上位になるように順序付けて提示する。このとき順序付けには、表の構造の特徴を機械学習した結果を用いる。

2. 例示検索方式

2.1 例示表の入力方式

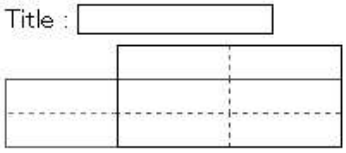

図 1 .表形式の検索インタフェース

ユーザは、図 1 の表形式の検索インタフェースに、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。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。

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。[2]

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。
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。
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。
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.例示表の入力例

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。

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、本来の表としてではなく、レイアウトを目的として使われている表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、表中にさらに表を含むもの、箇条書きを表で実現しているもの、などがあげられる。

それらの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。[3]

SVM の機械学習には、例示表の構造特性を反映した特徴と、Web 上の表の一般的な特徴を使用する。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である。以下に、使用するフィーチャー（特徴）の具体例を示す。

例示表の構造特性を反映した特徴

- 例示表の表頭にある検索単語が、検索された表の表頭にあるか否か
- 例示表の表側にある検索単語が、検索された表の表側にあるか否か
- 例示表の表本体にある検索単語が、検索された表の表本体にあるか否か

Web 上の表の一般的な特徴

- 表の大きさ
- 表に含まれる画像の数
- 1 セルあたりの文字列長

Query by Example of Web Information expressed as Tabular Formulation

Tomohiro YOKOKAWA†, Minoru YOSHIDA‡,
Koichi YAMADA†, Hiroshi KINUKAWA†, Hiroshi NAKAGAWA‡
† Graduate School of Engineering, Tokyo Denki University
‡ Information Technology Center, The University of Tokyo

上記の様なフィーチャーで機械学習を行う。機械学習によって生成されたモデルが定義する『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

異なった 8 つの分野において、それぞれ 1000 個の表を用意し、5-fold クロスバリデーションを用いて、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を行う。

ただし、8 つの分野は重複しないように設定する。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を検索に使用する。さらに、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。

機械学習のためのフィーチャーは 2.4 節に示したものに加え、合計 263 個を使用するものとする。

3.2 実験結果

表 1 に、実験結果を示す。正確度は、正解と不正解が正しく判別された割合である。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. 機械学習によるモデルの検証実験結果

分野	正確度	適合率	再現率	F-measure
旅行	87.9%	6.7%	1.0%	1.7%
乗り物	82.7%	27.9%	62.4%	38.5%
ゲーム	85.6%	29.4%	90.9%	44.4%
企業	90.9%	64.8%	80.0%	71.6%
コンピュータ	76.6%	87.1%	59.6%	70.8%
酒	82.7%	56.2%	65.0%	60.3%
茶	96.5%	80.3%	42.7%	55.8%
植物	88.5%	27.8%	65.5%	39.0%

4. 考察

3 の実験において、正確度が高く F-measure が低いことから、GoogleAPI の検索結果として得られた表には不正解が多いことが読み取れる。これは、表情報の例示検索によって、GoogleAPI の検索結果に含まれる多くの不正解の表を判別することの重要性を示している。また、高い正確度で判別できていることは、例示検索方式の有効性を示している。

さらに、適合率と再現率が分野によって差が激しいことについて考察する。ここでは、表 1 の旅行分野の適合率と再現率が著しく低いことに着目する。ただし、旅行分野で与えた例示表は図 2 のものである。

旅行分野において、例示表の構造特性の影響を調べるために、例示表の構造特性を反映したフィーチャーを排除し、機械学習を行った。以下に、検証実験結果を示す。

表 2. モデルの検証結果（例示表関連フィーチャーを排除）

分野	正確度	適合率	再現率	F-measure
旅行	90.1%	50.7%	73.0%	59.8%

表 1 と表 2 の旅行分野に着目すると、例示表の構造特性を反映したフィーチャーを排除したことによって、適合率・再現率がともに著しく上昇していることがわかる。つまり、例示表の構造特性を反映したフィーチャーを使用することによって、精度を下げるような学習が行われていると考えられる。

旅行分野で例示表の構造特性を反映したフィーチャーが有効でなかった原因を調べるために、旅行分野で機械学習に用いた 1000 個の表のうち、正解の表における例示表の検索単語の出現率と出現位置を調査した。その結果を表 3 に示す。

表 3.検索単語の出現回数と出現位置

検索単語	出現率[%]			
	表頭	表側	表本体	全体
ツアー	19	9	3	23
料金	21	49	16	75
ヨーロッパ	0	1	0	1
アジア	11	29	11	34

上記の結果から、旅行分野において、ヨーロッパとアジアは表頭と表側に出現することが少ないとわかる。つまり、属性を表す単語が表頭や表側に出現する割合が少ないことが、適合率と再現率の著しい低下の原因と考えられる。

また表 3 から、旅行分野の正解の表に、例示表の検索単語を含まないものが多いこともわかる。検索単語の出現率が低いため、機械学習に有効でなかったと考えられる。

さらに、上記の検索単語は同義語や下位語が多く存在することに着目する。以下に具体例を示す。

同義語

- ツアー → 旅行、旅
- 料金 → 価格、代金、￥

下位語

- アジア → アジア各地の地名
- ヨーロッパ → ヨーロッパ各地の地名

このように、使用した検索単語の同義語や下位語が多く存在することが、旅行分野において例示表の構造特性を反映したフィーチャーが有効でなかった原因として考えられる。

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提案した。SVM の機械学習によって生成された分類モデルを用い、検索した表情報を順序付けてユーザに提示することができる。

今後は、各分野の精度に差が出ていることに着目し、原因を調査する。特に、考察に記述した旅行分野の詳しい調査を行い、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] C.J.DATE: An Introduction to Database Systems (Third Edition), ADDISON-WESLEY PUBLISHING COMPANY, February 1982

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

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