D-046

Web 上の表情報を対象とした例示検索 ～表の構造的特徴の利用～
Query by Example Searching for Web Information in Tabular Formulation using Table Structures

前島一弥† Kazuya Maejima
横川智浩† Tomohiro Yokokawa
吉田稔‡ Minoru Yoshida
山田剛一† Koichi Yamada
絹川博之† Hiroshi Kinukawa
中川裕志‡ Hiroshi Nakagawa

1. はじめに

Web 上の表情報は構造化されているため良質な情報であることが多く、情報が構造化されていることを利用することで、従来の単語群を検索質問とする全文検索よりも精度の高い検索を行うことができると考えられる。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][2]

本論文では、フィーチャー（特徴）の最適化を行い、対象分野の偏りのない実験を行ってシステムを評価した。

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。検索インタフェースは、表を属性と値のペアの集合として捉え、表題を入力するセル 1 つ、属性を入力するセル 2 つ、値を入力するセル 2 つで構成している。属性のセルと値のセルは上下でペアとする。属性のセルには表頭や表側に出現すべき単語を入力し、値のセルには属性に対する値が入力されることを想定している。

表題		
属性		
値		

図 1 表形式の検索インタフェース

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を例示表と呼称するものとする。

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 に例示表をそのまま渡すことはできないため、検索インタフェースに入力された例示表を検索条件式に変換し、検索する。以下に、例示表の検索条件式への変換手順を示す。

(1) 左から順に、属性→値の順で検索単語を AND でつなぐ。
(2) 表題と(1)で生成した文字列を AND でつなぐ。

†東京電機大学大学院
‡東京大学情報基盤センタ

図 2 のような例示表の場合、「航空会社」、「JAL」、「料金」の順で、検索単語を AND でつなぐ。次に表題の「ツアー」と、(1)で生成した文字列をつなぐ。つまり、図 2 の例示表は、検索条件式「ツアー AND 航空会社 AND JAL AND 料金」に変換される。

表題	ツアー	
属性	航空会社	料金
値	JAL	

図 2 検索インタフェースの入力例

2.3 検索結果からの表情報の抽出

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すべきでない。具体例としては、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 として、TinySVM を使用している。[4]

SVM の機械学習には、予備実験により有効と判断された 70 のフィーチャー[2]に、6 つのフィーチャーを追加し、76 のフィーチャーを使用する（表 1）。その 6 つのフィーチャーを以下に示す。

- TH タグが使われているか
- TH タグの値が例示表の属性値に含まれているか
- CAPTION タグが使われているか
- CAPTION タグの値が例示表の表題に含まれているか
- SUMMARY 属性が使われているか
- 「レイアウト」の文字列が SUMMARY 属性の値に含まれているか

76 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与える『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。

表 1　フィーチャーの種類および数

カテゴリ	予備実験によって選別されたフィーチャー数
検索単語	8
形状	13
タグ付け	28
文字	27
計	76

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。

3. 評価実験

3.1 対象分野

分野の偏りをなくすため、様々な分野を網羅しているポータルサイトのカテゴリに注目し、4 つのポータルサイト（Yahoo!Japan、infoseek、goo、ライブドア）に共通する 12 のカテゴリを対象分野とする。それら分野に対して検索を行い、表を収集する。

3.2 表の収集

検索質問の構成を統一するため、各分野に対して、3 つの表構成パターンを用意し、表を収集する。3 つの表構成パターンは分野に関連した異なる内容である。表構成パターンについて以下に示す。

- 属性 A と値 A のペア
- 属性 A のみ
- 属性 A と値 A のペアかつ属性 B のみ

12 分野×3 表構成パターン×200= 7,200 表を収集する。

3.3 実験方法

以上収集した 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。

機械学習および検証は 4-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、3 分野で構成された約 1,800 個の表からなる。約 1,800 個の表は 3.2 節の 3 つの表構成パターン（約 600 個）×3 分野で構成されている。フィーチャーは 2.4 節に示したものを使用する。

収集した表は事前に正解と不正解に判別する。判別基準は、表構成パターンごとに異なる。「属性 A と値 A のペア」の場合は属性 A と値 A のペアを含んだ表を正解とする。「属性 A のみ」の場合は属性 A に対して値があった場合正解とする。「属性 A と値 A のペアかつ属性 B のみ」の場合は属性 A と値 A のペアを含む表で、かつ属性 B に対して値があった場合正解とする。また、共通の判別基準として表題がタイトルや前の文章に表れている場合を正解とする。例えば、図2の場合は表題「ツアー」がタイトルや前の文章にあり、かつ属性「航空会社」と値「JAL」のペアを含み、かつ属性「料金」に対して何らかの値があれば、正解となる。

3.4 比較対象

比較対象として Google の検索結果を表単位に変換したものを用いる。これを Google'と呼称する。このとき Google'は、検索結果の上位から表を順番に抽出し、検索単語の出現頻度順にソートしたものとする。検索精度の検証における対象分野は、3.1 節で示した分野である。

3.5 実験結果

精度を求めるため、平均精度を利用する。平均精度の定義を式 (1) に示す。

$$v = \frac{1}{\sum_{i=1}^{N} x_i} \sum_{i=1}^{N} \left[\frac{x_i}{i} \left(1 + \sum_{k=1}^{i-1} x_k \right) \right] \quad (1)$$

ここで、N は表情報の総数、Xi は出力順第 i 位の表情報の正解と不正解の状態を示す変数とする。正解ならば Xi = 1、不正解ならば Xi = 0 とおく。正解の出現数が 1 つまでの平均精度と、すべての正解が出現するまでの平均精度を計算する。

表 2 に、機械学習によって生成された分類モデルの精度を、表 3 に、Google'の精度を示す。

表 2　生成された分類モデルの精度

正解の再現数	精度
正解のうち最上位のもの 1 つ	78.42%
すべての正解	58.45%

表 3　Google'の精度

正解の再現数	精度
正解のうち最上位のもの 1 つ	71.33%
すべての正解	55.87%

4. 考察

3 の評価実験において、機械学習によって生成された分類モデルの精度は Google'の精度より、「正解のうち最上位のもの 1 つ」の場合は 7.09%、「すべての正解」の場合は 2.58%の差が出た。これによって、例示検索方式が Google'に比べ有効であることがわかる。一方、フィーチャーの最適化を図り精度を向上させることが必要である。

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検討した。本論文では、有効と考えられるフィーチャーの追加を行った。

機械学習によって生成された分類モデルの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、3.2 節の表構成パターンごとの検証実験を行い、有効性を示す。また、さらなるフィーチャーの最適化を行い、精度の向上と、検索インタフェースの使い勝手の向上を目指す。

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] 横川智浩「機械学習を用いた Web 上の表情報の例示検索方式」 FIT2006 第 5 回情報科学技術フォーラム, D-049, pp.115～116, September, 2006.
[3] Google: http://www.google.com/intl/ja/
[4] TinySVM: http://chasen.org/~taku/software/TinySVM/