# Web上の料理レシピの抽出とその利用

田島 幸恵
東京工業大学大学院総合理工学研究科
tajima@lr.pi.titech.ac.jp

奥村 学
東京工業大学精密工学研究所
oku@pi.titech.ac.jp

## 1 はじめに

インターネットの普及により、Web 上には様々な情報が記述されるようになった。これらの情報は個人から企業、その他様々な組織によって管理されている。多くの場合企業などに管理されている情報は、豊富でありかつ検索等の利用が容易であるように整備されている反面、内容にその企業の販売している商品が使われているなど情報が偏りがちな傾向がある。対して個人などに管理されている情報は、内容に上述のような偏りがない反面、情報が少なく整備もされていない傾向がある。更に、これらの情報は分野ごとにまとまっているわけではなく散在している。

そのため、Web 上の莫大な情報を一つの対象について収集し活用したいと考えても、まず収集が難しい。更に、仮に収集したとしても、多くのページでは著者が異なるため情報の書式も異なりがちであり、必要な部分の情報を抽出することも難しい。このため、存在している情報の極一部しか利用できないのが現状である。

しかし、対象分野によっては情報の書式に一定の特徴を持つ場合がある。それを利用することで、他の対象について記述された部分を省き必要な情報だけを収集できると考えられる。本研究では、一定の特徴を備えている分野として「料理レシピ」を選択する。

[1] では「料理レシピ」の収集をし、多数のレシピを用いた検索を可能にする手法が述べられている。複数のレシピ集の url を人手で入力することでそれらのサイトに掲載されているレシピの検索が可能になる。しかし、この方法ではレシピ集の url を Web 上から人手で発見し入力しなければならず、前述した傾向を持つ企業等に管理されたレシピの検索は可能でも個人が管理しているようなレシピの検索は難しい。

本論では、様々な情報が散在している Web から料理レシピを含むページを収集し、レシピ部分のみを抽出、それを利用する手法について述べる。

従来行われている料理レシピに関する研究 [2, 3, 4, 5] はレシピ部分のみを利用できることが前提になっている。本手法で抽出されるレシピは、これらに対するコーパスに成りうると考えている。

## 2 提案手法

Web からの料理レシピを含むページの収集、対象ページからのレシピ部分の抽出を下記のように行う。

1. Web からの候補ページの取得
2. 候補ページが料理レシピであるかどうかの判別
3. レシピを含むページからレシピ部分の抽出

### 2.1 Web からの候補ページの取得

料理レシピを含むページを収集するために、Google APIs にレシピに含まれやすい語をクエリとして与える。

Web 上の `html` ソースは正確に書かれているとは限らず、ソースをそのまま解析に利用するとエラーの原因となる可能性があるため、`html` の標準化を進めている w3c による `html` 整形ソフトの tidy を利用し、エラーを起こすことなく正確な形に書き換えることが出来たページのみを収集する。

### 2.2 料理レシピであるかどうかの判別

ドメインの異なる 100 ページ分の料理レシピに対して調査を行ったところ、「タイトル」を含むページが 99 ページ、「材料と分量」を含むページが 91 ページ、「作り方」を含むページが 87 ページ、三要素全てを含むページが 80 ページであった。そこで、三要素を持つページが料理レシピであるとする。

「材料と分量」「作り方」と比較して「タイトル」はタグのかかり方や周辺の文字列に特徴がない。「タイトル」以外の二要素を持つページは 79 ページであり、判別時に「タイトル」を要素として用いないことで得る誤抽出は 1 ページに過ぎない。そこで、「材料と分量」「作り方」の要素を持つページをレシピを含むページと判断する。

#### 2.2.1 材料とその分量の判別

材料とその分量の記述のされ方には次のような特徴がある。

- 材料名に使われる文字列には若干の多様性があり外国産の素材の名前などの新語が現れることもある。対して分量を示すために用いられている単位の種類にはそれほどの多様性はない。
- `html` タグのかかり方や材料と分量の間のセパレータ、先頭に付与されるアイテマイズの印などはほぼ繰り返しで同じ物が使われている。

これらの特徴を満たすものを「材料とその分量」であると判断する。

#### 2.2.2 作り方の判別

手順の抽出に関する研究として [6] があり、SVM を用いて <li> の付与された範囲が手順であるか否かの判定をする手法の提案をし、F 値で 0.6 以上の精度を得ている。しかし、レシピにおける作り方の記述の約 75%は <li> を用いずに手書きで番号を記述している。また、対象ページを限定しているため、他の内容を示すリストが出現しにくいと考えられる。

作り方の記述のされ方には次のような特徴がある。

- 調理手順の順序を示すために、多くの場合各手順の先頭に昇順に番号がふられている
- タイトル名などの主に名詞から構成される調理手順以外の記述と異なり、手順を示す場合には記述は文からなり適当数の動詞や助詞を含む

これらの特徴を満たすものを「作り方」であると判断する。

### 2.3 レシピ部分の抽出

前節の手法では、各要素の正確な範囲までは決定していない。以降で要素の記述されている範囲を正確に決定する。

#### 2.3.1 材料と分量の範囲の特定

前節で述べた手法では、一番最初あるいは最後に現れる材料名や単位名が未知であったり間に使われているセパレータが他では使われていないものであったりする場合抽出もれをしてしまう。そこで、範囲と判定した部分の上下数行にわたって範囲と判定した部分にかかっているタグと同様のタグがかかっておりかつ材料名か単位名を含んでいる場合、その文字列も材料と分量の範囲の可能性がある箇所であるとした。

#### 2.3.2 作り方の範囲の特定

前節で述べた手法では、<li> タグが用いられていない場合、作り方の開始位置はわかるが、終了位置はわからない。作り方に記述されている最後の手順の終了位置を決定することで、作り方の範囲は決定される。

以下に述べる 4 手法を用いて、範囲が最も小さくなるものを最終的に選択する。

- 新しく作り方の範囲、材料と分量、あるいは html ソースの終端が出現する場合、その位置を終端とする。
- 各調理手順の内、最後の手順を除く手順に共通なパターンを抽出し、最後の手順において、このパターンを満たさないタグが出現した位置を終端とする。例えば、図 1 では、調理手順「ジャム...」「りんごは...」はどれも <table><tr><td> である。これに対し「秋の味覚...」をマークアップする <p> はこのパターンを満たさないため、「鍋に水...」の直後が終端となる。

```
<table>
<tr>    <td>  <b>(1)</b></td>
        <td>  ジャムを作る。<br/>
              りんごは、皮と....</td></tr>
<tr>    <td>  <b>(2)</b></td>
        <td>  鍋に水と....</td></tr></table>
<p>           秋の味覚...
```

図 1: タグの例

- 各調理手順の終端に出現する空行や <img> や <hr> などの記号列をセパレータとして取得する。取得したセパレータの共通部分と最大部分を判定に利用する。共通部分を満たさず最大部分に含まれない記号列が出現し、かつ以降で調理の手順が出現しない場合、その位置を終端とする。
- 「出来上がり」「完成」などの作り方の終端に現れやすい語をキーワードとし、キーワードが出現し、直後に複数の記号からなる行を持ち、かつ以降で調理の手順が出現しない場合、その位置を終端とする。

#### 2.3.3 二要素の組み合わせの抽出

1 ページに複数のレシピが記述されている場合、定義を満たすレシピと分量や作り方が記述されていないレシピが同時に存在することがある。このような定義から外れる部分を除き、材料と分量、作り方の組み合わせを抽出する必要がある。

組み合わせには、材料から見て作り方が前にある場合と後にある場合の二通りが考えられる。

以下の手法を用いて、どちらの組み合わせであるかを決定する。

(1) 各作り方の開始位置、終了位置、含んでいる文字列を収集する

(2) 材料と分量の範囲と直前直後に存在する作り方の範囲の位置が十分近いと判断された場合、それを組み合わせになりうる候補範囲であるとする

(3) 材料の範囲に含まれている材料名の中で候補範囲に出現するものを数える

(4) （「材料と分量ー作り方」での出現数） < （「作り方ー材料と分量」での出現数）となる場合、現在対象としている「材料と分量」は直前の「作り方」と組になり、それ以外は直後の「作り方」と組になる

レシピでは多くの場合「材料と分量ー作り方」の順で記述をするため、(4) において、出現数が等しい場合は「材料と分量」は直後の「作り方」と組にする。

#### 2.3.4 タイトルの特定

入力された文字列に対して「タイトルに含まれる文字列」の範囲を決定するため、チャンカーとして YamCha[7] を用いてタイトルの特定を行う。

予備実験の結果、「b(タイトルに含まれる最初の文字列)、i(タイトルに含まれる b 以外の文字列)、o(タイトルに含まれない文字列)」タグを判定に用いるタグとし、学習を行う。

### 2.4 抽出結果を利用した分類

料理レシピの総合サイト [8] では、国籍、用途、種類、味などによる分類がされている。中でも、「辛い」「こってり」「すっぱい」「さっぱり」の味のカテゴリは、レシピで用いられている材料と調理動作に大きな影響を受けていると考えられる。そこで、レシピに使用されている材料と調理手順に含まれる動詞を素性として SVM を用いて、4 カテゴリに分類する。

## 3 実験と考察

前章までで提案した手法の有効性を確かめるために、人手で正解を付与した異ドメイン 300 ページ分に対して実験、評価を行う。材料と分量、作り方の抽出実験を行った後、分割されたレシピに対してタイトルの抽出実験を行う。

### 3.1 材料と分量、作り方の抽出実験

Web 上にあるレシピ数は莫大であり、抽出もれを防ぐことよりも正確な抽出をすることが重要と考える。そこで、本手法では「評価値 = (人手で得た正解と手法で抽出したものの一致数) / (手法で抽出した数)」とする。

表 1: 材料と分量、作り方の抽出結果

| | 評価値 | |
|---|---|---|
| 材料と分量 | 836/1114 | (0.750) |
| 作り方 | 962/1114 | (0.864) |
| 積集合 | 804/1114 | (0.722) |

その結果、表 1 が得られた。レシピとして利用するためには、「材料と分量のみ」「作り方のみ」の抽出ではなくこれらが共に抽出されている必要があり、二要素の積集合の結果も示した。

また、これらの値を用いて材料と分量のみ抽出できていたレシピ数、作り方のみ抽出できていたレシピ数などを算出すると表 2 のようになる。なお、表に示す○は抽出が出来ていたことを、×は抽出が出来ていなかったことを意味している。

表 2 から、「(a) 作り方の抽出のみの失敗」より「(b) 材料と分量の抽出のみの失敗」の方が多いことや作り方の抽出が失敗した時、「(a) 材料の抽出の成功」より「(d) 材料の抽出の失敗」の方が多いことが言える。

表 2: 各部分の抽出結果

| 材料 | 作り方 | 抽出数/総数 | |
|---|---|---|---|
| ○ | × | 32/1114 | ....(a) |
| × | ○ | 158/1114 | ....(b) |
| ○ | ○ | 804/1114 | ....(c) |
| × | × | 120/1114 | ....(d) |

これらは、本手法がまず作り方の範囲を決定し、範囲外になった部分から材料と分量の抽出を行っていることに起因すると考えられる。

また、失敗の原因として以下のものがある。

- 材料がテーブルで記述されている
- 材料の記述の仕方が一定ではない
- 材料の記述と作り方の記述の距離が離れている
- 複数のレシピが記述されている中に、作り方の記述がないものや材料の分量がないものが含まれる
- 作り方の順番が数字で記述されていない

### 3.2 タイトルの抽出実験

前項で抽出した 1114 レシピに対して 3 分割交差検定を行った。その結果、正解率 0.841 となった。

抽出の失敗は下記のような原因による。

- 画像によるタイトル表示
- レシピの区切りの失敗
- ChaSen の解析の失敗

### 3.3 分類手法の評価実験

#### 3.3.1 closed データへの手法の適用

[8] から得られる、カテゴリが付与されている 6749 レシピに対して 3 分割交差検定を行った。その結果、正解率 0.637 となった。

手法で用いた素性の有効性を示すために、各カテゴリの精度と再現率と F 値を以下の (A)(B)(C) で示す三種類の素性に対して求め、表 3 に示す結果が得られた。

(A) 材料のみ
(B) 調理手順で用いられている動詞のみ
(C) 材料と調理手順で用いられている動詞

表 3: SVM を用いた時の結果

| カテゴリ名 | (A) | (B) | (C) |
|---|---|---|---|
| 辛い | 0.6795 | 0.4592 | 0.6295 |
| こってり | 0.5245 | 0.3969 | 0.5970 |
| さっぱり | 0.2913 | 0.1955 | 0.7363 |
| すっぱい | 0.4680 | 0.2438 | 0.4793 |

また、全てのレシピをそのカテゴリに分類した時に得られる値を baseline とし、「辛い」に 0.2912、「こってり」に 0.2691、「さっぱり」に 0.2402、「すっぱい」に 0.1996 の値を得た。

#### 3.3.2 closed データに対する実験の考察

表 3 は、(A)(B) よりも、(C) の方が結果が良くなることを示している。これは特に「さっぱり」に対して顕著である。このことから、レシピは材料をいかに調理するかによって分類先が分かれることと、実験に用いた素性が有効であることが言える。

また、(A)(B) の結果を比較すると、(A) の方が正しい分類が行われることがわかる。このことから、料理レシピでは「材料」の方がよりレシピを特徴づけることがわかる。

現在 (B) は、調理動作を示す動詞以外に、「くる」「れる」「食べる」などの具体的な動作が分からない動詞や調理動作を指さない動詞も含んでいる。調理動作を示す動詞からなる辞書を生成し、辞書に出現する表現のみを素性とするなどの処理が必要と考えられる。

#### 3.3.3 open データへの手法の適用

第 2 章で述べた手法を用いて Web 上から収集した、レシピを含むページが約 30000 ページある。これに対して、同様の実験を行った。

#### 3.3.4 open データに対する実験の考察

分類結果には直感的に不適当に感じるものもあった。不適当に感じるものとして、「すっぱい」に分類された「カッテージチーズ豆腐のねぎソースかけ」と「タルト・タタン」を選択し原因を考察する。

(1) 「カッテージチーズ豆腐のねぎソースかけ」は直感的には、豆腐に類するものにねぎソースをかけていることから「さっぱり」、またはソースに辛み材料であるトウバンジャンを利用していることから「辛い」に分類される。しかし、恐らく牛乳からチーズを作る段階で使用されている「酢」が強い影響を与える素性として働いたと考えられる。

(2) 「タルト・タタン」はりんごを使ったタルトである。他にも「りんごのカスタードケーキ」などの、直感的には、材料にバター等を用いることから「こってり」に分類される洋菓子のレシピが「すっぱい」に分類されていた。しかし、同じ洋菓子でも「モンブラン」などのリンゴを用いないケーキは「こってり」に分類される。このことから、「りんご」が「すっぱい」への分類に強い影響を与える素性として働いたと考えられる。

(1) の原因として、「酢」がレシピの味に影響を与えない段階で用いられていることが素性からはわからないことが原因として考えられる。(2) の原因として、学習データに含まれるレシピには「すっぱい」カテゴリに「りんご」を使用したものが多かったことが考えられる。

特に (2) に対しては、学習に用いるデータを増やすことで対応できると考えられる。

## 4 まとめ

本論では、料理分野に特徴的な語をクエリとして Web からページを取得し、その中から料理レシピを含むページを収集し、レシピは「タイトル」「材料と分量」「作り方」の要素から形成されているという考え方を元にこの三要素を一組のレシピとして抽出し、抽出したレシピを用いて検索支援に用いる分類を行った。

提案手法により正解率 72%でレシピの抽出がされた。更なる手法の洗練が求められる。誤決定をしたページの特徴を洗い出すことで抽出の正解率が上げられると考えられる。

分類の正解率は 64%となっており、より正解率を上げる必要がある。

一般的に SVM では学習データのサイズが大きい方がより高精度な分類が行われる。そこで、より大きな学習データを用いることで正解率を上げることができると考えられる。更に、本論では各レシピで使用している材料と調理中に現れる動詞を用いて分類を行ったが、前述の通り、調理中に現れる動詞が必ずしも調理動作を示しているとは限らない。そこで調理動作表現の辞書を生成し、含まれる動詞を材料のみからなる現在の素性に追加することを考えている。

また、現在の分類は [8] で得られるカテゴリを利用して行ったが、味の特徴を示す語は他にも多数ある。レシピの特徴はしばしばタイトルに記述されることから、Web から収集したレシピのタイトルを形態素解析し、得られる特徴語に対して同様の分類が可能と期待している。収集される分類先が付与されたデータが少量の場合には SVM の正解率が低下するため、NaiveBayes を用いることを考えている。

## 参考文献


[1] Tara Calishain and Rael Dornfest. *GOOGLE HACKS*. O'Reilly, 2003.

[2] 安達久博. 類推に基づく料理定義文の自動獲得. 情報処理学会研究報告, NL-112-9, 1996.

[3] 柴田知秀, 黒橋禎夫. 料理教示発話の理解と作業構造の自動抽出. 情報処理学会研究報告, NL-164, 2004.

[4] 唐沢隆, 浜田玲子, 井出一郎, 坂井修一, 田中英彦. 料理教材テキストからの素材と調理法に関する知識の抽出. 第 66 回情報処理学会全国大会, Vol.2, 2T-2, 2004.

[5] 浜田玲子, 井手一郎, 坂井修一, 田中英彦. 料理テキスト九材における調理手順の構造化. 電子情報通信学会論文誌 (D-II), Vol.J85-D-II, No.1, 2002.

[6] 武智峰樹, 徳永健伸, 松本裕治, 田中穂積. WWW ページからの手順に関する箇条書きの抽出. 情報処理学会論文誌, Vol.44, No.8, 2003.

[7] 工藤拓, 松本裕治. Support vector machine を用いた chunk 同定. 自然言語処理, Vol.9, No.5, 2002.

[8] COOKPAD. http://cookpad.com/.