同義語辞書を用いた相関ルールによるレシピ検索数変化

2011SE087 伊藤 康佑　2011SE118 加藤 大地　2011SE218 斎藤 恭平
指導教員： 河野 浩之

1　はじめに

料理レシピを検索する際に「ふわふわ」などの擬音語を表すオノマトペを用いることが多い．本研究では，オノマトペと食材の相関ルールを検出することにより，オノマトペが含まれないレシピも検索されるようにする.「レシピ検索システムにおけるオノマトペとレシピ用語集合の関連付け」[2] では,「ふわふわ」と「ふわっ」などの同義語を別々のオノマトペとして扱っていたことにより，オノマトペとレシピ用語の関連性に問題があった．そこで，本研究では同義語辞書を用いることで表記ゆれを無くし，中村らの提案した料理レシピ検索の性能向上を目指す．本論文は全 6 章で構成され，2 章では，オノマトペを用いた料理レシピ検索の先行研究の紹介と比較を説明する．3 章では，同義語辞書を用いた相関ルール検出のアーキテクチャを説明する．4 章では，アーキテクチャに基づき相関ルールの検出を行い，得られた相関ルールからレシピ検索を行う．5 章では，得られた結果から評価を行う．6 章では，これまでに得られた結果のまとめについて説明する．

2　オノマトペを用いたレシピ検索の先行研究

2.1 節では，レシピ検索システムにおけるオノマトペとレシピ用語集合の関連付けについて，2.2 節では，食材の優先度を考慮した料理レシピ検索について，2.3 節では，先行研究の比較を行う．

2.1　オノマトペロリ：オノマトペを利用した料理推薦システム [3]

料理サイトにおいて「とろとろした料理」，「じっくり焼く」など味や感触などの表現にオノマトペが頻繁に利用される．たとえば「ピリッ」とした料理を検索したい場合，レシピ内の文章かタイトルに「ピリッ」という文字が含まれなければ検索することができない．そこで TF-IDF(Term Frequency-Inverse Document Frequency) 法を用いて「ある文章のグループでの出現数が多い」単語のうち，「他の文章のグループでの出現数が少ない」ものをその文章のグループに特徴的なオノマトペとし, 以下の二つ方法を用いてオノマトペでのレシピ検索を実現する．一つ目の方法として, オノマトペを関連の強い用語に置き換えて検索する．二つ目の方法として, レシピに使われている食材名などからレシピとオノマトペの関連度を算出してランキング表示する．

2.2　レシピ検索システムにおけるオノマトペとレシピ用語集合の関連付け [2]

[3] では TF-IDF 法による 1 対 1 の関連度に基づいてレシピを推薦していた．しかし，1 つの用語は別の用語との組み合わせでオノマトペとの関連が異なる．[2] では頻出アイテムの集合に対して「X ⇒ Y」となる関連をアプリオリアルゴリズムを用いて求める．アプリオリアルゴリズムの評価指標として，支持度，確信度，リフトを用いる．[2] では「オノマトペ 1 語と用語 2 語」の相関ルールを求めた結果, 用語の集合オノマトペ 1 語，用語とオノマトペの集合用語 1 語，オノマトペ 1 語用語の集合，用語 1 語オノマトペと用語の集合という 4 パターンで示される．

2.3　先行研究の問題点をふまえた提案

先行研究での問題点は，オノマトペと食材の相関ルールを検出する際に,「ふわふわ」と「ふわっ」などの同義語を別々のオノマトペとして扱っていることや食材の表記ゆれなどがある．これにより正確なオノマトペと食材の相関ルールは検出されていない．これを踏まえた改善点は，オノマトペ辞書や食材辞書を作成するによりオノマトペの同類語をまとめることや食材の表記ゆれをなくす.

よって本研究では，料理レシピ掲載サイトから料理レシピの抽出を行い，次に抽出された料理レシピを python 上で MeCab を利用して形態素解析を行い，オノマトペと食材だけにする．その時にオノマトペ辞書と食材辞書により，オノマトペの同義語をまとめ，食材名の表記ゆれを改善する．

そしてオノマトペと食材の相関ルールを解析ソフトを用いて求める．また，相関ルールの検出にはアプリオリアルゴリズムを利用する．それにより，オノマトペ同義語辞書を使用した時の違いが検出できる．また検出された相関ルールから料理レシピを検索を行うプログラムを実行する．

3　同義語辞書を用いた相関ルールの検出のアーキテクチャ

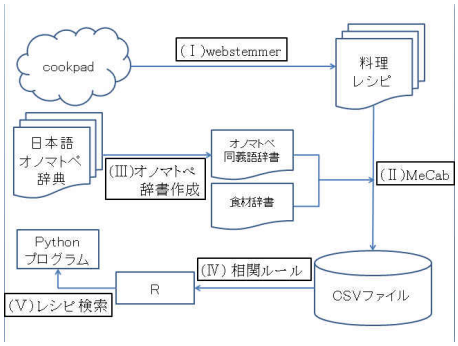

図 1　同義語辞書を用いた相関ルールの検出のアーキテクチャー

本章では，図 1 の相関ルール検出のアーキテクチャについて，3.1 節では，(I)Web ページから料理レシピの抽

出，3.2 節では，(Ⅱ) 抽出を行った料理レシピの形態素解析を行い，3.3 節では，(Ⅳ) オノマトペと食材の相関ルールついて説明する．

3.1 レシピ抽出

はじめに，料理掲載サイトの各料理レシピの抽出を行う．抽出を行うツールを表 1 に表す．本節では，抽出する料理レシピをテキスト化することで，料理レシピ掲載サイトから料理レシピの抽出を行う．そのためのツールは表 1 に示してある．

表 1 本文抽出ツール一覧

抽出ツール	特徴
GETHTML	ページ全体のテキストを抽出可能
NNScanText	標準の描画処理を行った画面なら テキスト抽出可能
WebTextClip	掲示板 · 時刻表 · など 種類を選ばずテキスト抽出が可能
webstemmer	Web クローラによりサイトの レイアウトを学習し，ブログなどの 本文抽出が可能

まず GETHTML は，現在開いている Web ページの内容を簡単に抽出を行うことができるツールである．1 ページだけ抽出を行うことが特徴である．次に NNScanText は，標準の描画処理を行った画面ならテキストに抽出が可能で，Web ページだけでなく，作業画面などでテキストをコピーできないものまで可能である．WebTextClip は，Web ページの HTML データをから抽出を行なうので，ブログや掲示板などの種類を選ばず使用できる．そして webstemmer は，ニュースサイトの記事本文や記事のタイトルなどをプレインテキスト形式で，自動的に抽出を行うソフトウェアである．web サイトのトップページの URL を与えれば全自動に解析が行われるので，人の手はほとんど必要がない．

よって本研究では，料理掲載サイトから複数のページにある料理レシピの抽出を行なうので webstemmer を用いる．

3.2 形態素解析

形態素解析ツールには，表 2 に示す以下のものがある．

表 2 形態素解析ツール一覧

形態素解析ツール	特徴
KAKASI	漢字かな混じり文をひらがな文 やローマ字文への変換が可能
MeCab	KAKASI に比べて高速で 応用性が高い
Ekwords201	日本語/英語の文書データから キーワード (単語，連語) を 抽出 · 集計するソフト

KAKASI とは，漢字かなまじり文をひらがな文やローマ字文に変換することを目的として作成したプログラムと辞書の総称のことである．

また Ekwords201 は，手軽に日本語と英語ん文書を形態素解析を行なえるが，応用性が低いので，今回は KAKASI より解析が高速で，Ekwords201 より応用性の高い MeCab を用いる．

3.3 相関ルール

この節では，オノマトペと食材の相関ルールを求める．相関ルールの検出にはアプリオリアルゴリズムを利用する．相関ルールを求めるのに使用されるツールを表 3 で表す.

表 3 解析ツールの一覧

解析ツール	特徴
統計ソフト R	配列・リスト・テーブルなどの複雑 なデータ構造も構築・管理できる
weka	クラスタリング，統計分類， 相関ルールといった標準的な データマイニングタスクを サポートしている
RapidMiner	日本語非対応であるが一通りの分析 ができ，拡張すれば weka 同様の 分析フローが構築できる

統計ソフト R は，データベースや csv ファイルなどからデータの読み込みやリストや配列などの複雑なデータ構造にも対応し，膨大なデータの処理もおこなえることや，これまでに使ってきたツールとも一連の流れで実験を行なえるので，本研究では，統計ソフト R を使用する．

4 オノマトペ同義語辞書を用いた相関ルールの検証

4.1 節で cookpad について，4.2 節では webstemmer による料理レシピの抽出，4.3 節では形態素解析，4.4 節では同義語辞書作成，4.5 節では統計ソフト R を用いた相関ルールの検出について述べていく．

4.1 cookpad

本節では，実験に用いる料理レシピサイトを表 4 で表す.

表 4 料理レシピサイト一覧

レシピサイト	特徴
cookpad	料理レシピ数 190 万件を超える 日本最大の料理レシピサイト
楽天レシピ	91 万件以上のレシピを気になる 食材や料理から検索できる
カロレピ！	レシピ数 1 万件以上 カロリーや栄養価がわかる
E· レシピ	料理レシピ掲載数 3 万件以上 プロがつくる簡単レシピサイト

cookpad(http://cookpad.com) は，料理レシピの投稿・検索サイトとして 1998 年に誕生した．現在のレシピ数は 190 万件を超えて，利用者は 2000 万人，日々料理をする 20 代や 30 代の女性を中心に利用されている．cookpad は，一般的な料理レシピとオノマトペが含まれた料理レシピが多い．なぜなら cookpad のような投稿型の料理レシピサイトでは，料理の様子や味覚を伝える際にオノマトペが多く利用される傾向があるからである．

よって本研究で実験に用いる料理レシピサイトは，料理レシピ数 190 万件を越え，多くのユーザーによってオノマトペを含む料理レシピが投稿されている cookpad を使用する．

4.2 Webstemmer による抽出

本節では，webstemmer を用いて料理レシピの抽出の方法を説明していく．webstemmer は，ニュースサイトの記事本文や記事のタイトルなどをプレインテキスト形式で，自動的に抽出を行うソフトウェアである．web サイトのトップページの URL を与えれば全自動に解析が行われるので，人の手はほとんど必要がない．本研究では，まず初めに料理レシピを抽出を行うために，webstemmer を用いて cookpad の料理レシピをランダムに抽出を行う．本研究では，cookpad に掲載されている 190 万件から 2 万 7880 件のレシピの抽出を行った．

4.3 形態素解析

本節では，webstemmer によって抽出された cookpad の料理レシピを python 上で MeCab を用いて形態素解析を行う．相関ルールを求めるにあたり，料理レシピからオノマトペと食材の抽出を行う必要がある．

料理レシピからオノマトペと食材を抽出する方法は，まずオノマトペリストと食材リストを用意しておく，そして料理レシピを読み込ませて形態素解析を行う．その時にオノマトペにあたる副詞と食材にあたる名詞のみにしておく．そしてオノマトペリストと食材リストに適合したものを重複なくリストに格納する．

4.4 辞書作成

本節では, 料理レシピから単語を取り出すためのオノマトペリストと食材リスト, 同義語を一つにまとめる同義語辞書の作成について述べる．

食材リストには， 食品サイト Whole Food Catalog(http://wholefoodcatalog.com/) から抽出した食材 1878 品目を登録した． オノマトペリストには，日本語オノマトペ辞典 [1] の意味分類別さくいんに記載されている 2470 語の中から料理レシピに使われると思われる「事物に関するオノマトペ」283 語を登録した. この 283 語をオノマトペ辞典に記載された意味・用法を基に 87 語の同義語辞書として作成した. 例えば, 「張る」,「膨らむ」に分類される「ふわふわ」,「ふんわか」,「ふんわり」は共通の「ふくらんでやわらか」という意味を持つため同義語として登録した.

食材リストと同義語辞書を用いて料理レシピから食材名とオノマトペを抽出する．

4.5 統計ソフト R による相関ルールの検出

本節では，python のプログラムによって作成したオノマトペと食材のリストが格納された csv ファイルを統計ソフト R に読み込み，アプリオリアルゴリズムを用いて相関ルールの検出を行う．統計ソフト R を用いて相関ルールを求める方法は図 2 のようになっている．

```
library(arules)
d.tran=read.transactions(file="Desktop/sample1.csv",sep=
",",format="basket")
d.tran
d.ap=apriori(d.tran)
d.ap=apriori(d.tran,parameter=list(supp=0.035,maxlen=3))
d.ap.sub1=subset(d.ap,subset=(lhs %in% "じっく
り")&(lift>1.0))
inspect(SORT(d.ap.sub1,by="support"),n=30)
inspect(head(sort(d.ap,by="support"),n=10))
```

図 2 統計ソフト R によるコマンド

はじめに 64bit 版の統計ソフト R を起動し，アプリオリアルゴリズムを利用するためのパッケージである arules をインストールする．次に d.tran=read.transactions(file=”Desktop/sample1.csv” ,sep=”,”,format=”basket”) により，csv ファイルからオノマトペと食材を読み込み，バケットデータとして格納する．そして d.ap=apriori(d.tran) により相関ルールを作成し，読み込んだデータからアプリオリアルゴリズムのデータ作成する．本研究では基準となる支持度，確信度とリフト値は support=0.003，confidence=0.005，lift=1.0 とする．支持度は，全トランザクションのうち，条件部と帰結部に登場する全てのアイテムを含むトランザクションの割合である．確信度は，条件部のアイテムを買う人が帰結部アイテムを買う確率である．リフト値は，帰結部または条件部の支持度が高くても，そのアイテム集合はいつも起こるなどの一般性の高すぎるアイテム集合を含む相関ルールを除くことができる．相関ルールの確信度を帰結部の支持度で割った値である．今回の実験では，オノマトペ同義語辞書にあるすべてのオノマトペ含む相関ルールをアプリオリアルゴリズムより検出している．同義語辞書を使用した場合の相関ルールの件数は 6405 件，同義語辞書を使用しなかった時の相関ルールの件数は 6455 件であった．図 3 は，同義語辞書を使用した時に「じっくり」と食材の相関ルールであり，それぞれの支持度，確信度，リフト値が示されている．

```
> d.ap.sub1=subset(d.ap,subset=(rhs %in% "じっくり")&(lift>1.0))
> inspect(head(SORT(d.ap.sub1,by="confidence"),n=500))
   lhs                rhs            support confidence     lift
1  {トマト,
    鶏肉}         => {じっくり} 0.003080785 0.31764706 4.418824
2  {オリーブ,
    コンソメ}     => {じっくり} 0.004678229 0.25465839 3.542581
3  {オリーブ,
    スープ}       => {じっくり} 0.003194888 0.25454545 3.541010
4  {コンソメ,
    トマト}       => {じっくり} 0.003879507 0.23287671 3.239574
5  {オリーブ,
    パセリ}       => {じっくり} 0.003993610 0.23026316 3.203216
6  {トマト,
    ニンニク}     => {じっくり} 0.005020539 0.20853081 2.900895
7  {オリーブ,
    ニンニク}     => {じっくり} 0.006275673 0.20000000 2.782222
8  {オリーブ,
    トマト}       => {じっくり} 0.006960292 0.18318318 2.548282
```

図 3 相関ルールの検出

4.6 相関ルール基づくレシピ検索

本節では，アプリオリアルゴリズムにより検出された確信度高い相関ルールからオノマトペと関連の強い食材名からレシピ検索を行う．

本研究では，python のプログラムを使用することによりレシピ検索を行う．料理レシピの検索方法は，図 3 のように相関ルールが検出された時に確信度が高いルールからオノマトペと関連の強い食材をキーワードとしてレシピ検索を行う．レシピ検索はキーワードがレシピ内に出現するかで判断をしている．例として図 3 で「じっくり」に関連の強い食材は「トマト」と「鶏肉」が検出されていて，それをキーワードとしてレシピ検索を行い，キーワードが含まれるレシピが出力される．

5 評価

今回の実験では，同義語辞書を用いることで表記ゆれを改善し，R によってオノマトペと食材の相関ルールをアプリオリアルゴリズムにより 6222 件のルールが検出され，同義語辞書を使用しなかった時は 6262 件のルールが検出された．これはオノマトペや食材の表記ゆれがなくなりオノマトペや食材の数が減少したことにより検出されるルールが減少した．またオノマトペ 228 語それぞれアプリアリアルゴリズムにより相関ルールを検出を行ったが，料理に関連の強いオノマトペ 13 語を含む相関ルールのみが検出された．

オノマトペ	食材		検索数	辞書あり-辞書なし
しっかり	クリームチーズ	ジャガイモ	18	6
	バニラ	茶	12	
ふんわり	ニンジン	強力粉	15	-20
	クロワッサン	強力粉	35	
さっぱり	ごま油	青じそ	15	-7
	きゅうり	オクラ	22	
しんなり	アユ	モチ	18	-12
	キャベツ	マーガリン	30	
こんがり	チーズ	ロールパン	39	17
	エビ	食パン	22	
じっくり	オリーブ	茶	29	-4
	ニンニク	薄力粉	33	
ゆっくり	コーヒー	ココア	61	-5
	ゼラチン	卵	66	
ふっくら	トマト	強力粉	125	18
	ミルク	強力粉	107	
シャキシャキ	モヤシ	菜	36	0
	もやし	菜	36	
とろとろ	トロロ		106	17
	サラダ	マカロニ	89	
ふかふか	タマゴ	強力粉	498	30
	強力粉	卵	468	
こってり	鶏肉	醤油	374	101
	ミリン	大根	273	
もりもり	サラダ	ニンニク	195	-5
	サラダ	大根	200	
	辞書あり合計		1529	合計
	辞書なし合計		1393	136
	辞書あり平均		117.61538	
	辞書なし平均		107.15385	

図 4 検索結果

図 4 は，同義語辞書を使用した時と使用しなかった時の相関ルールを表している．上が同義語辞書ありの場合で，下が同義語辞書なしの場合である．また検出された相関ルールよりオノマトペと関連の強い食材からレシピ検索を行った結果も示している．例えば「しっかり」というオノマトペに対して同義語辞書を使用した場合は,「クリームチーズ」,「ジャガイモ」というルールが検出されて，それにより検索されたレシピ 18 件，同義語辞書を使用しなかった場合は,「バニラ」,「茶」という相関ルールが検出されていて，それにより検索されたレシピは 12 件であることが示されている．

図 5 は，オノマトペ 13 語を同義語辞書を使用した時と使用しなかった時に支持度を 0.001 から 0.01 まで変化させて，それぞれ検出された相関ルールからオノマトペと関連の強い食材出てきて，その食材レシピ検索を行った時のレシピ検索数の変化を表してある．支持度を 0.001 から 0.01 から変化させて実験を行なったところ，支持度が 0.004 と 0.007 の時以外は，レシピ検索数が向上した．

結果として従来の料理レシピ検索システムでは，検索されなかったレシピが同義語辞書を使用することにより検索されるようになった．同義語辞書を使用した場合と使用しなかった場合では，3.32 %のレシピを多く出力できた．

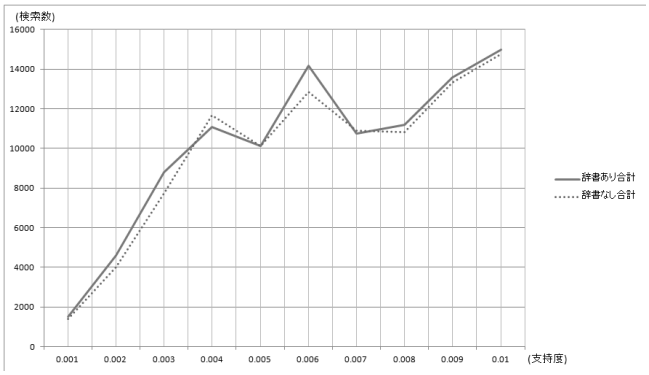

図 5 同義語辞書ありの場合の検索数の変化量

6 まとめ

本研究では，同義語辞書を用いることで表記ゆれを改善し，オノマトペによる料理レシピ検索の検索数変化の検証を行った．オノマトペと食材の同義語辞書を作成し,「ふわふわ」と「ふわっと」,「たまご」と「卵」など同じ意味を持つ単語を同じものとして扱うことで，より多くの検索数を出力することができた．実験では，支持度を 0.001 から 0.01 から変化させて実験を行なったところ，支持度が 0.004 と 0.007 の時以外は，レシピ検索数が向上した．これにより 3.32 %のレシピ検索数が向上した．また，本研究ではレシピ検索に python を用いたため，データ数の上限が少なかったり，様々な条件で検索が行えないことや検索に時間がかかるという欠点がある．MySQL を用いることでこの課題は解決されると考える．

参考文献

[1] 小野正弘,“ 擬音語・擬態語 4500 日本語オノマトペ辞典，”小学館，pp.33-64，2014.

[2] ラートサムルアイパン・カンウィパー，渡辺知恵美,“ レシピ検索システムにおけるオノマトペとレシピ用語集合の関連付け，”電子情報通信学会研究報告，情報処理技術，Vol.150，no.15，pp.1-8，2010.

[3] ラートサムルアイパン・カンウィパー，渡辺知恵美，中村聡史,“ オノマトペロリ：オノマトペを利用した料理推薦システム，”情報処理学会研究報告，情報処理技術，Vol.73，no.6，pp.1-7，2009.