DBSJ Japanese Journal, Vol.13-J, No.1
October 2014

誕生・使用事由によるレシピ検索 ～生い立ちレシピサーチ～

Recipe Search by Reasons of Births and Usages

門脇 拓也♡　山肩 洋子♢
森 信介♠　田中 克己♣

Takuya KADOWAKI♡　Yoko YAMAKATA♢
Shinsuke MORI♠　Katsumi TANAKA♣

その日何を作るかといった献立を決めることは，多くの主婦がストレスに感じる事柄の一つである **[1]**．料理を作る漠然とした事由はあっても，その事由に合った具体的な料理を思いつくのは難しい．そこで本研究は，このような事由からレシピを検索できるシステムを提案する．**Web** 上には，レシピと併せて，なぜそのレシピが誕生したのか，なぜそのレシピを使用したのかといった，そのレシピの **“生い立ち”** に相当する情報が書かれている『ブログ型レシピ』が数多く存在している．そこで，ブログ型レシピからレシピの誕生・使用事由を，『材料・料理がある』『季節・気候に合う』などのカテゴリごとに抽出する．また，ユーザにその日の気分や出来事などの日記を入力させ，そこからそのユーザが料理を決める事由となるようなキーワードを抽出する．さらに，これらを対応付けることにより，その日のユーザに適したレシピをユーザに提示する．

Many homemakers feel it stressful to decide the menus of the day [1]. Even if they have vague reasons for dishes, it is difficult for them to come up with specific dishes suitable for the reasons. Therefore, this research proposes a system which enables users to search for recipes based on such reasons. In the Web, there are many “blog-type recipes” that describe not only a recipe but also the reasons why the recipe was born or used. This research introduces new approach to search for recipes suitable for users by extracting the reasons why the recipe was born or used from blog-type recipes and by categorizing them.

1. はじめに

一般的な情報検索サービスの多くは，ユーザに検索対象であるオブジェクトの性質を表す単語集合をクエリとして入力させ，クエリに合致するオブジェクトを検索するものである．しかしながら，ユーザは漠然とした検索意図はあるが，その検索意図に合致するオブジェクトが具体的にどのようなものなのかわからないことも多い．一方で，検索対象の Web ページには，オブジェクトの情報に加え，なぜそのオブジェクトが誕生したか，あるいはそれを使用したかといった誕生・使用事由が書かれていることがあり，これが今ユーザが抱えている検索意図に対応する情報であると考えた．そこで本研究は，ユーザのそのときの思いや状況といったコンテキストを twitter[1] や Facebook[2] に書き込むような文章の形でクエリとして入力させ，そのコンテキストに合致した誕生・使用事由を持つオブジェクトを検索することで，ユーザの意図にあったオブジェクトを発見することを目標にする．本研究では具体的な検索タスクとして，ユーザ数が多く，ニーズも高い料理レシピを対象にする．

レシピは，潜在的な検索要求はあるものの，オブジェクトの具体的なイメージを持つことが難しい検索タスクである．「肉じゃが」といった料理名や，「ゴーヤ」といった材料名が決まっていれば，現在 Web で提供されているサービスで十分解決可能である．しかし，「残業で疲れたときに元気が出るレシピを探したい」というように，レシピに対する漠然とした要求はあっても，具体的な料理を思いつくのは難しいことも多い．よって，そのときの状況をクエリとして，ユーザの要求にあったレシピを発見できる検索サービスは有用であると考える．

本研究では，このような検索サービスを実現するため，ブログ型レシピを検索対象とする．レシピブログ [2] に代表されるブログ型レシピには，レシピと併せてその日の生活の記録を書いたブログ型レシピが数多く存在しており，そのレシピが誕生した事由あるいはそれを使用した事由が記載されていることが多い．しかしながら，レシピに直接的には無関係な事柄も数多く記載されているため，どれがレシピの誕生・使用事由に当たる文かを特定する必要がある．そこでまず，誕生・使用事由がどのようなものかを調査するため，レシピブログからランダムに 1,000 ブログを抽出し，そこからレシピの誕生・使用事由に相当する 1,808 文を抽出して手動で分類した結果，計 18 カテゴリに分類された．同じカテゴリに分類された文集合について，単語の出現傾向を分析したところ，ブログ記事の中から各カテゴリに合致する事由文を検出するためには，構成する単語の種類を見ただけでは不十分であることが分かってきた．

そこでまず，ブログから日記部分だけを取り出すことを考えた．ブログ型レシピにおけるレシピの記述形式は，COOKPAD などの一般的なレシピポータルサイトにおける表記とほぼ一致しており，特徴的である．そこで，COOKPAD のレシピ集合で学習した言語モデル [3,4] と COOKPAD の材料リストから構築した辞書を用いて，ブログ記事からレシピを除外する．

次に，事由文かそうでないか，また各カテゴリの事由文かを判断するため，文を構成する単語の種類だけでなく 4 種類の前処理を考案し，それらを組み合わせた学習データをもとに SVM による識別器を構築して各カテゴリの事由抽出精度を評価する．この結果をもとに，各カテゴリの事由を，ブロガーがブログにどのような形で表記するのかといった表現モデルについて議論を行う．

最後に，被検者実験により，本手法で提案するアプリケーションの有用性を評価する．事由文の誤抽出による影響を避けるため，事由文抽出は手動で行った結果を利用する．システムを利用するユーザのシナリオを 3 種類設定し，被験者に対して種類ごとにそのユーザの状況を説明する．さらに，レシピに対する AND 検索による結果と，提案手法によるレシピ検索結果を比較し，その満足度を 5 件法により評価する．

以下，第 2 章で関連研究について紹介し，第 3 章で提案システムのコンセプトおよびユーザシナリオについて述べる．第 4 章でレシピブログにおける誕生・使用事由の分析について述べ，第 5 章でその分析結果を用いた手法を提案する．第 6 章で提案手法

♡ 学生会員　京都大学大学院情報学研究科
kadowaki@dl.kuis.kyoto-u.ac.jp
♢ 非会員　京都大学大学院情報学研究科
yamakata@dl.kuis.kyoto-u.ac.jp
♠ 非会員　京都大学学術情報メディアセンター
forest@i.kyoto-u.ac.jp
♣ 正会員　京都大学大学院情報学研究科
tanaka@dl.kuis.kyoto-u.ac.jp

[1] https://twitter.com/
[2] https://www.facebook.com/

を評価するための実験計画について説明し，その結果を考察する．最後に，第 7 章でまとめと今後の課題について述べる．

2. 関連研究

2.1 レシピ検索

Web 上でレシピを検索できるサービスは数多く存在する．ユーザ投稿型レシピを取り扱った代表的なサービスとして，COOKPAD [5] や楽天レシピ [6] が挙げられる．このようなレシピ検索サイトでは，食材名や料理名のような，レシピに記載されている情報を使ってレシピを検索するのが一般的である．COOKPAD に登録されているレシピには,「このレシピの生い立ち」という，そのレシピを考案するに至った理由を記入する項目があるが，そこに書かれているのは「鱈を簡単に美味しく食べたかったので」といった極めてシンプルな理由に限られ，そのレシピ制作者がどのような環境で何を思って作ったのかといったリアルな状況を窺い知ることは難しい．

2.2 レシピ検索・推薦

食材名や料理名のようなレシピに記載されている情報からではなく，レシピ検索者の嗜好や気分，食材の残り具合からレシピ検索を可能とする研究も数多くある．上田ら [7] は，食材に対する好き嫌いが食事に対する嗜好の一部を構成していると考え，ユーザのレシピ閲覧および調理履歴から，好きな食材・嫌いな食材を推定し，それらを基にレシピに対して得点付けを行うことで，レシピを推薦する手法を提案している．森下ら [8] は，生活者の気分に合わせて献立を提案し購入すべき食材の決定を支援するシステム「気分による献立検索システム」を研究開発し,「時間」「味」といった 6 つの気分検索軸の重要度を評価している．また，一回の料理で使い切れない材料を効率よく活用するため，木原ら [9] は,「一週間」など一定期間内に食材を使い切ることができるようなレシピ群を推薦する手法を提案している．しかしながら，これらのようなレシピ検索・推薦システムは，献立を決める際の嗜好・気分・食材の残り具合という事由それぞれについては解決できるが，献立を決める事由はこれ以外にもあり，かつ，その事由は一つではなく複合的である．

2.3 オブジェクトの周辺情報による検索

本研究は，なぜそのレシピが誕生したのか，なぜそのレシピを使用したのかといった，レシピにまつわる周辺情報を使ったレシピ検索を可能とすることを目的とするが，このように検索対象のオブジェクトにまつわる周辺情報によりオブジェクトを検索する手法は他の分野で様々なものが提案されている．荘司ら [10] は,「面白い」や「泣ける」といった印象語に適した Web ページに，必ずしもそれらの印象語が書かれているわけではないため，それらの印象語を使って検索することができないという問題を，ウェブコミュニケーションデータに含まれるリアクションを用いることによって解決している．また，食べログ [11] では,「口コミ」や「評価」といったオブジェクトの周辺情報も検索の対象とすることで，ユーザのニッチなニーズに適したレストランを検索できるサービスを提供している．このようなレストラン検索サイトでの「口コミ」や「評価」は，一般ユーザによる評価のため，必ずしも信頼できるとは限らないが，大島ら [12] は，東京の寿司屋に対する一般ユーザによるオンラインレビュー情報に対し，HITS アルゴリズムを基にしたアルゴリズムを適用することによって，専門家による評価情報と似たような寿司屋のランキングを行う手法を提案している．

3. 誕生・使用事由によるレシピ検索

本章では，ブログ型レシピを用いた誕生・使用事由によるレシピ検索のコンセプトおよびユーザシナリオについて述べる．

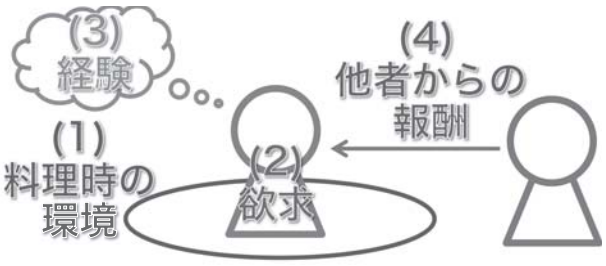


図 1: レシピを選ぼうとするときの 4 つの要因

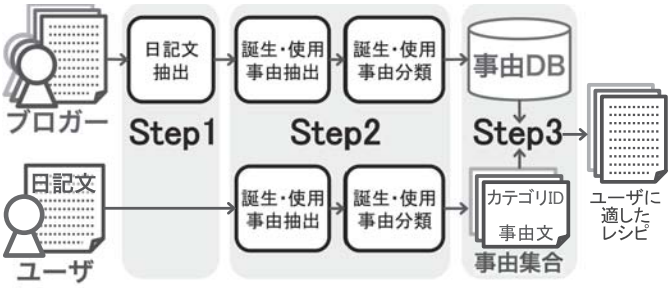


図 2: 誕生・使用事由によるレシピ検索 全体の処理

3.1 誕生・使用事由によるレシピ検索のコンセプト

調理者がレシピを選ぶ事由には，図 1 で示した 4 つの要因があると考える．

1. 料理時の環境
   食材の有無，調理器具の有無，時間的な余裕の有無など
2. 調理者の内的欲求
   生理的な欲求（体調の良し悪し），嗜好など
3. 経験
   他者から得た情報（うわさ，ニュース），自分の経験など
4. 他者からの報酬
   料理を食べる人からの報酬（子供が喜ぶ），第三者からの報酬（レシピコンテスト）など

これらの要因は，ブログ型レシピから誕生・使用事由を抽出する際の判断基準となるものである．

ブログ型レシピのブロガーは，ブログにレシピを載せる際，それと併せて図 1 に示すような要因を日記部分に記していることが多い．一方でレシピを検索したいユーザも，図 1 のような要因を持っており，よってユーザにコンテキスト情報を入力させれば，その中に要因が記されることが期待される．そこで図 2 に示す提案手法により，ブログの日記部分と，レシピ検索のユーザに入力させたコンテキスト情報から，それぞれ要因を表していると思われる文を自動抽出し，カテゴリごとに対応付けることにより，その日のユーザが持つ漠然とした要因に適したレシピを検索できる手法を提案する．

3.2 ユーザシナリオ

以下のようなユーザがいるとする．

> 就職して半年，初めて一人暮らしをしたユーザは，料理や洗濯，掃除などをほとんどしたことがなく，まだ慣れない．接客業をしているユーザは，クリスマス前のため忙しく，ここ最近休みなく働き通しである．22 時頃帰宅したユーザは晩ごはんをまだ食べていない．普段から出費を抑えているユーザは，疲れていても簡単に料理でき，かつ元気が出るような料理を知りたい．

このような状況を説明したコンテキスト情報を提案手法によるレシピ検索システムに入力すると，システムは「疲れている」「普段から出費を抑えている」「簡単に作れる」「元気が出る」といったレシピ使用の事由となる情報を自動抽出する．システムは，ユーザのレシピ使用事由と同様の事由を持つレシピを，そのレシピの誕生・使用事由文とともにユーザに返す．ユーザはレシピだけでなく，そのレシピの誕生・使用事由文も見ることができ，その事由は今のユーザの状況と同様のものであるため，そのレシピに対して親近感がわき，「今の自分と同様の状況であった人が作ったあるいは使ったレシピなら自分も作ってみようかな」とユーザは思うと考える．近年では，twitter や Facebook に代表される SNS で自分のいまの状況を発信するユーザが数多くいる．本研究では，このように SNS で発信された文章が，本レシピ検索システムへの入力となることを想定する．

4. レシピブログにおける誕生・使用事由の分析

本章では，ブログ型レシピにおける誕生・使用事由の出現傾向を分析する．

4.1 データセット

日々の料理を主題としたブログが数多く登録されているポータルサイト「レシピブログ」を対象とする．ブログ内でレシピを紹介しているページの内，2012 年 11 月 1 日から 2013 年 10 月 31 日までの 1 年間に投稿されたアメーバブログ [3] 記事 52,369 件をデータセットとする．

4.2 分析方法

無作為に選んだ 20 件のブログ記事からレシピの誕生・使用事由を含む文を抜き出し，それらを図 1 の 4 つの要因に従って，できるだけ詳細に手動でクラスタリングした結果，表 1 で示す 18 種類のカテゴリが得られた．

そこで，別途無作為に選んだ 2,074 件のブログ記事それぞれに対して，9 名の評価者に以下の要領で誕生・使用事由の抽出とカテゴリ分類をしてもらった．

1. 食材リストおよび調理手順から構成される「レシピ」部分とそれ以外の「日記」部分に各文を分ける
2. 「日記」部分から「レシピ」の誕生・使用事由を含む文を見つける（複数選択可）
3. 各事由が表 1 の中のどのカテゴリに属するか選択する（複数選択可）

4.3 分析結果

評価の結果，誕生・使用事由が含まれるブログ記事が 1,000 件見つかった．各事由カテゴリの出現率は図 3 のとおりである．最も多いのは「材料・料理がある」であり，全体のおよそ 2 割を占める．また，上位 4 つのカテゴリが全体の 6 割以上を占めており，これらのカテゴリが特に重要であると考えられる．

また，1 つのブログ中にいくつの事由が含まれるか調べた結果を図 4 に示す．この図によると，全体の半分以上はレシピの誕生・使用事由を 1 つも含まないことがわかる．これは，各文が誕生・使用事由かどうかを考慮せずに，単にレシピに併記された日記に対して検索を行うだけでは，全く無関係なレシピが見つかってしまう可能性が高いことを示唆している．

さらに，各文にいくつの事由カテゴリが当てはまったか調べたところ，1 つ 92.8%，2 つ 6.86%，3 つ 0.387%，4 つ以上なしという結果であった．9 割以上は文によって排他的にカテゴリが対応付けられたことから，このようなカテゴリ分類は妥当であると言える．

[3] http://ameblo.jp/

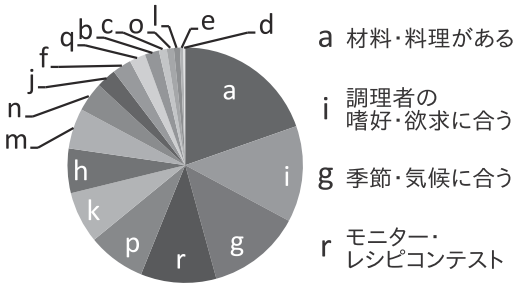


図 3: 各事由カテゴリの出現率（図中の英字は表 1 の ID に対応）

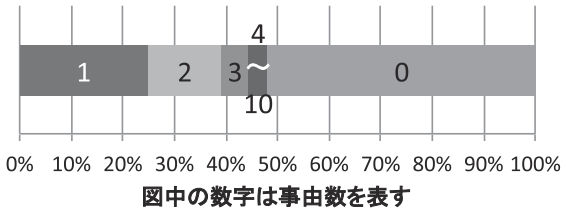


図 4: 1 つのブログ中に含まれる事由数別のブログの割合

5. 誕生・使用事由によるレシピ検索手法

本章では，提案システムを実現するための手法を述べる．

5.1 処理の概要

本節では，誕生・使用事由によるレシピ検索を実現するための処理の概要を述べる．全体の処理の流れは図 2 に示したとおりである．

Step1: ブログ記事からの日記文抽出
ブログ文書に対し，COOKPAD [5] から収集したレシピの手順説明文から構築した言語モデル [3, 4]，および材料情報から構築した辞書を用いて，日記部分のテキストの抽出を行う．

Step2: 誕生・使用事由の抽出とカテゴリ分類
識別器を用いて，Step1 で得た日記文から事由カテゴリごとに事由文を自動抽出する．また，ユーザが入力したコンテキストに対しても同様にして事由文を抽出する．

Step3: ユーザのコンテキストに適合する事由を持つレシピの検索
Step2 で得た，各ブログから抽出した事由文集合と，ユーザの日記から抽出した事由文集合に対し，カテゴリが同じで，かつ事由文が類似しているものを見つける．その数が多いブログほど，ユーザのコンテキストに合致したレシピであると考える（本研究ではこれを適合度と呼ぶことにする）．適合度が高いもののうち上位 K 件を検索結果として出力する．

5.2 ブログ記事からの日記文抽出

ブログ型レシピの本文は，以下のように分類できる．

- レシピ部分 (*recipe*)
  - 材料部分 (*ingr*) … レシピで使用する材料のリストが書かれた部分
  - 手順部分 (*proc*) … レシピの調理手順が書かれた部分
- 日記部分 (*diary*) … レシピ部分以外

レシピの誕生・使用事由が書かれているのは日記部分であり，レシピ部分に現れることはない．よって，もし高い精度で日記部分とレシピ部分を切り分けることができるなら，レシピ部分から事由文を誤抽出することがない．そこでブログ型レシピから事由

表 1: 誕生・使用事由のカテゴリ一覧

要因	ID	カテゴリ名	事由の実例文
料理時の環境	a	材料・料理がある	「我が家にカボチャがゴロゴロとたくさん」
	b	材料・料理がない	「あいにく、いつもストックしてある、デミグラスソースがなかった」
	c	調理器具がある	「新しくキッチングッズを買って、早く試してみたくて」
	d	調理器具がない	「うち、それ用のケーキスタンドがないんです！(汗)」
	e	時間的な余裕がある	「子どもたちがいないので作れるのかも（笑)」
	f	時間的な余裕がない	「そんなわけで、バタバタしていたため、」
	g	季節・気候に合う	「朝は寒くて布団から出るのがつらい。」
	h	イベントがある	「これは先日友達が来てくれた時にちゃちゃっと作った」
調理者の内的欲求	i	調理者の嗜好・欲求に合う	「お餅が食べたくなっちゃったワン♪」
	j	調理者の気分・体調に合う	「なんとな〜くやる気が出ない時」
経験	k	調理者の体験・経験から	「母がよく蒸しパンを作ってくれました」
	l	食べる人の体験・経験から	「「レタスでサラダを包むと美味しいね〜」と。。。」
	m	材料と材料の相性が良い	「やっぱり、トマトソースに、なすはあいます」
	n	材料と料理の相性が良い	「新タマ、サラダにバッチリ！」
	o	料理と料理の相性が良い	「そんなとん平焼きを今回はトーストの上に乗っけちゃいました。」
他者からの報酬	p	食べる人の嗜好・欲求に合う	「娘に夏からずっとせがまれていた」
	q	食べる人の気分・体調に合う	「主人も娘も風邪気味です」
	r	モニター・レシピコンテスト	「すき家の牛丼の具をモニターとして頂きました。」

文を抽出するための前処理として，日記部分とレシピ部分の切り分けを行う．

まず，日記部分の抽出を行う前準備として，レシピ部分の特徴を表すモデルを以下のようにして構築した．

- 手順部分の文特徴を表すモデル：**3-gram** 言語モデル
  COOKPAD [5] に掲載されている 100,000 件のレシピから抽出した 497,536 文の調理手順文に対し，形態素解析器 kytea [13] を使って単語分割したうえで構築された 3-gram 言語モデル [3, 4] を用いた．以降，この言語モデルを LM で表す．この LM により，文の相対エントロピーを算出すると，COOKPAD と似た記述形式である手順部分は言語モデルによく適合するためエントロピーが低くなり，エントロピーの大きさによって，その文が手順文らしいかどうかを判定することが可能であると考えられる．
- 材料部分の文特徴を表すモデル：材料名辞書
  COOKPAD [5] に掲載されている 422,150 件のレシピから抽出した 390,427 種類の材料名から材料名辞書を構築する．この辞書では，例えば「にんじん」「ニンジン」「人参」「☆人参」などを全て区別している．以降，この材料名辞書を *ingr_dict* と表す．もし文を構成する単語のほとんどがこの辞書に登録されたものであるならば，その文は材料部分である可能性が高いと考えられる．
- 材料部分の文特徴を表すモデル：名詞・補助記号・全角スペースの現れやすさ
  材料部分には,「玉ねぎ」「人参」のような材料名だけでなく,「☆砂糖，☆醤油」というように材料の集合を補助記号で参照したり,「玉ねぎ　1 玉」というように全角スペースを間に挟んだりすることが多いという特徴がある.『名詞・補助記号・全角スペース』の 3 種類の品詞を *ingr_class*，それ以外の品詞を *not_ingr_class* と表す．

以上の 3 種類のモデルを用いて，以下で述べる 2 段階の処理により日記部分を抽出する．

第 **1** 段階：文単位の判定

ブログ記事中の各文を，レシピの手順部分，レシピの材料部分，日記部分のいずれかに分類し，レシピの手順部分とレシピの材料部分にはラベル *recipe* を，日記部分にはラベル *diary* をそれぞれ付与する．

日記部分にレシピの説明を書く際に材料名を書くこともあるため，文中に出現する単語が材料名辞書に存在することがわかっただけでは，その文がレシピの材料部分であるとは言い切れない．そこで，材料部分の文特徴に合致するかどうかも判定することで，この問題を解決する．

1. 各文 s に LM を適用してエントロピー $H_{LM}(s)$ を計算する．
2. $H_{LM}(s)$ が閾値 *proc_threshold* 以下の文にラベル *recipe* を付与する．
3. $H_{LM}(s) >$ *proc_threshold* であった文 s を形態素解析し，単語列 $w_1, ..., w_n, ..., w_N$，およびそれに対応する品詞列を得る．ここで，N は文 s を形態素解析した結果得られた単語数を表す．以降で行う繰り返し処理のため，$test = w_1$，$i = 1$ と初期化する．
4. 材料名辞書 *ingr_dict* 中に *test* が存在するか調べる．
5. - 材料名辞書に存在する場合
     文 s について，*ingr_class* に含まれる品詞の数を *ingr_class_num*，*not_ingr_class* に含まれる品詞の数を *not_ingr_class_num* とする．*ingr_class_num*/*not_ingr_class_num* の値が閾値 *ingr_threshold* 以上ならば，文 s にラベル *recipe* を付与する．
   - 材料名辞書に存在しない場合
     もし i が N に等しければ，6. へ進む．そうでなければ i に 1 を加えて *test* の末尾に w_i を追加し，4. に戻る．
6. ラベルが *recipe* でない文全てに対してラベル *diary* を付与する．

第 **2** 段階：連続文に対する判定

レシピ部分は，その間に日記文を挟むことなく，ひとつづきにつなげて記述するのが一般的である．よって，周辺の文のラベルが *recipe* なら，その文のラベルも *recipe* である可能性が高い．そこで，第 1 段階で文単位で得た判定結果をもとに，文のつながりやすさを考慮して日記部分を抽出する．

以降，ブログ記事 B に書かれている各文を $s_i(1 \le i \le M)$ とし，第 1 段階で得た文 s_i のラベルを $label_i$ とする．ここで，M はブログ記事 B に含まれる文の数である．

関数 *is_recipe*(), *is_diary*() を次のように定義する．

$$is\_recipe(label_i) = \begin{cases} 1 & (label_i \text{が } recipe \text{ ならば}) \\ 0 & (\text{上記以外}) \end{cases}$$

$$is\_diary(label_i) = \begin{cases} 1 & (label_i \text{が } diary \text{ ならば}) \\ 0 & (\text{上記以外}) \end{cases}$$

1. 各文 s_i に対して，その直前直後の BA_NUM 文を含む文集合について，*recipe* とラベル付けされた文の数 $recipe\_num_i$，*diary* とラベル付けされた文の数 $diary\_num_i$ をそれぞれ次のように算出する．

$$recipe\_num_i = \sum_{j=-BA\_NUM}^{BA\_NUM} is\_recipe(label_{i+j})$$

$$diary\_num_i = \sum_{j=-BA\_NUM}^{BA\_NUM} is\_diary(label_{i+j})$$

2. もし，$diary\_num_i > recipe\_num_i$ ならば，文 s_i はレシピ部分よりも日記部分に多く囲まれていることから，日記部分として抽出する．

5.3 誕生・使用事由の抽出とカテゴリ分類

本節では，前節で日記部分として抽出されたブログ記事中の文と本検索システムのユーザが入力したコンテキストのそれぞれに対して，レシピの誕生・使用事由に相当する文をカテゴリ別に抽出する処理について述べる．なお，第 4.3 節で行った調査によれば，事由文の 92.8%はカテゴリを 1 つしか持たないことから，ここでは事由文はいずれか 1 つのカテゴリに排他的に分類可能であると想定する．

事由文は，そのカテゴリにかかわらず，そうでない文とでは言語的な特徴が異なるならば，すべてのカテゴリの事由文をまとめて識別器を構築した方が，多くの学習データを用意でき，より高い精度で事由文を抽出できると考えた．そこで，第 1 段階で事由かどうかを判別する識別器を用いて事由文を抽出したあと，第 2 段階で抽出された事由文がどの事由カテゴリに属するか分類する識別器を適用する．

- 事由かどうかを判別する識別器
  第 4 章で行った分析により得た事由文を正例，それ以外の日記文を負例とした 2 値分類器を構築する．正例は 1,808 文であるのに対し，負例の文数は，正例の数よりも非常に多いため，これらすべてを使って識別器を構築するとデータ不均衡による偏りが生じる．よって負例のデータ数が 1,808 文になるようアンダーサンプリングした．識別器は SVM(Support Vector Machine [14]) を用いた．これを *is_reason_svm* とする．
- 事由カテゴリに分類する識別器
  学習データが少ないと，識別器の識別性能は大きく低下する．そこで，本研究では，表 1 の 18 カテゴリの内，第 4 章で行った分析により得たデータ数が 90 文以上あったカテゴリ a,g,h,i,k,m,p,r の 8 カテゴリと，それ以外の全カテゴリ (データ数が 90 文未満のカテゴリ b,c,d,e,f,j,l,n,o,q) をまとめたもの，の計 9 種類のいずれかに分類する 9 値分類器を構築する．データ数が 90 文以上あったカテゴリの中で最も文数が少なかったカテゴリの文数は 99 文であったことから，各分類のデータ数はそれぞれ 99 文になるようアンダーサンプリングする．9 値分類器には SVM を用いた．これを *categorize_svm* と表す．
- アルゴリズム
  1. 前節で日記部分と判別された各文を識別器 *is_reason_svm* に入力し，レシピの誕生・使用事由に相当する文かどうか判別する．
  2. 1. で事由文と判別された文を 9 値分類器 *categorize_svm* に入力し，上で述べた 9 種類の事由カテゴリのうちどれに属するかを分類する．

5.4 ユーザのコンテキストに適合する事由を持つレシピの検索

本節では，ブログ記事と，ユーザのコンテキスト情報に対して，前節で述べた手法によりカテゴリごとの事由文を抽出したあと，各ブログ記事が，ユーザのコンテキスト情報から抽出した事由文にどの程度合致するかを表す適合度を算出することにより，ユーザのそのときのコンテキストに合ったレシピを検索する手法を述べる．

前節で述べた手法により，ユーザの日記から X 個の事由文が抽出されたとし，その各事由文を r_x，そのカテゴリを c_x とする（ただし $1 \le x \le X$）．また，同じく前節の手法によりブログ記事 B の日記部分から Y 個の事由文が抽出されたとし，その各事由文を R_y^B，そのカテゴリを C_y^B と表す（ただし $1 \le y \le Y$）．

各事由文 r_x および R_y^B に対して，その文を構成する単語とその頻度による行列から文書ベクトルを生成する．これをそれぞれ $\boldsymbol{r}_x$，$\boldsymbol{R}_y^B$ とする．

アルゴリズム

1. 一致度 $score_{x,y}^B$ を以下のようにして計算する．

$$score_{x,y}^B = \begin{cases} \dfrac{\boldsymbol{r}_x \cdot \boldsymbol{R}_y^B}{|\boldsymbol{r}_x||\boldsymbol{R}_y^B|} & (c_x \text{と } C_y^B \text{が等しいとき}) \\ 0 & (\text{上記以外}) \end{cases}$$

   ただし，$|\cdot|$ はベクトルの大きさを表す．

2. ブログ記事 B のユーザのコンテキストに対する適合度スコア $BS\,core^B$ を以下のようにして計算する．

$$BS\,core^B = \frac{\sum_{x=1}^{X}\sum_{y=1}^{Y} score_{x,y}^B}{score\_non\_zero\_num^B}$$

   ただし，$score\_non\_zero\_num^B$ は，$1 \le x \le X$，$1 \le y \le Y$ について，$score_{x,y}^B$ の値が 0 でないものの数を表す．

3. $BS\,core^N$ が高い順から上位 K 件のブログ記事と，そのブログから抽出された事由文およびそのカテゴリを，検索結果としてユーザに提示する．

6. 実験と考察

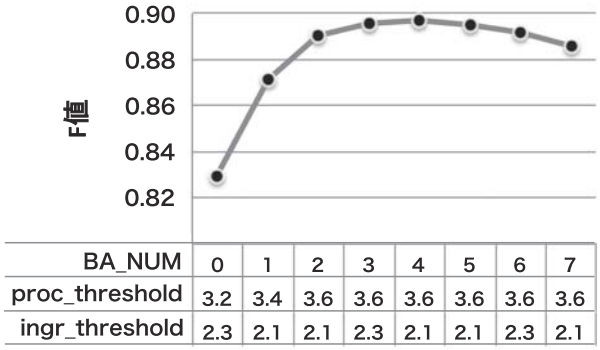
本章では，第 5 章で述べたアルゴリズムの精度評価，および第 3 章で述べたレシピ検索システムの評価を行う．データセットは第 4 章で行った分析で収集したものを用いる．

6.1 ブログ記事からの日記文抽出処理の精度評価

第 4 章で分析に用いた，事由文を一文以上含むブログ記事 1,000 件，計 50,298 文の各文に対し，第 5.2 節で述べたアルゴリズムを用いて日記部分かレシピ部分かを自動判別した結果を，評価者によって手動で判別された結果を正解データとして評価した．同時に，第

表 2: TP, TN, FP, FN の定義

		正解データ	
		日記	レシピ
自動判別結果	日記	TP	FP
	レシピ	FN	TN



BA_NUM	0	1	2	3	4	5	6	7
proc_threshold	3.2	3.4	3.6	3.6	3.6	3.6	3.6	3.6
ingr_threshold	2.3	2.1	2.1	2.3	2.1	2.1	2.3	2.1

図 5: 日記抽出結果の F 値の変化

5.2 節で述べた 3 つのパラメータ (*proc_threshold*, *ingr_threshold*, *BA_NUM*) についても，最適な値を探索した．この処理の目的は，ブログから日記部分を抽出することである．そこで，自動判別結果と正解データの組み合わせに対して，True Positive(TP), True Negative(TN), False Positive(FP), False Negative(FN) を表 2 のように定義し，そこから日記抽出についての適合率 (*Precision*), 再現率 (*Recall*), F 値 (*F−measure*), および全体の正解率 (*Accuracy*) を以下に示す式に従って計算した．

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F\text{-}measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

評価は以下のように行った．*BA_NUM* を 0 から 7 まで 1 刻みで変え，そのそれぞれの *BA_NUM* に対して日記抽出結果の F 値が最も高くなるときの *proc_threshold*, *ingr_threshold* を 0 から 5 まで 0.1 刻みで調べた．その評価結果を図 5 に示す．この図によると，*BA_NUM* = 4, *proc_threshold* = 3.6, *ingr_threshold* = 2.1 のとき，F 値が最も高く 0.897 であった．また，このとき適合率は 0.908，再現率は 0.887，正解率は 0.876 であった．*BA_NUM* の値が 0 のとき，注目している文だけでレシピ部分か日記部分かを判別することになるが，その結果よりも，1 以上のときの方が F 値が高いことから，ブログ記事中においてレシピはまとまって記述されることが多く，文の連続性を考慮した方が精度が向上することがわかる．

次に，誤判定された文にはどのようなものがあるか分析した．文例 A「（3）刻んだ甘納豆を混ぜて出来上がり。」はレシピの手順部分と判断されるべき文である．しかしながら，言語モデルの構築に用いた COOKPAD の手順文は，「1）」「（1）」「1.」のような手順番号がレシピの整形フォームにより与えられていたため，学習に用いた文集合には付いていなかった．そのため，文例 A のように手順番号が文頭に付いた文は誤判定される傾向があったと考える．また，文例 B「生クリームがなくても、卵の良さを感じるカルボナーラです。」はブログ記事中に載せるレシピの特徴を述べるために書かれた文で，日記部分と判断されるべき文である．しかしながら，このような記述は COOKPAD のようなユーザ投稿型レシピでは，手順部分に載せられることもあり，そのような文集合で学習した言語モデルが誤ってレシピの手順部分であるとみなしたと考えられる．文例 A のような問題に対しては，文頭に付いた手順番号を予め除去するといった手段により改善可能であると考えている．また，文例 B のような問題に対しては，第 5.2 節で述べた連続文に対する判定を行うアルゴリズムを改良することにより改善可能であると考えており，これらは今後の課題である．

6.2 誕生・使用事由抽出処理とカテゴリ分類処理の精度評価

本節では，第 5.3 節で述べたように，入力文が事由文かどうかを判別する識別器 (*is_reason_svm*)，事由カテゴリに分類する識別器 (*categorize_svm*) についてそれぞれ 10 分割交差検定を行い，F 値，適合率，再現率，正解率を算出することで精度を評価する．識別器は，LIBSVM [15] による SVM(Support Vector Machine) を用い，カーネルには線形カーネルを，パラメータには LIBSVM のデフォルト値（C-SVC のパラメータ C の値は 1）を用いる．

事由抽出や事由カテゴリ分類を高い精度で行うためには，文中の単語からそのまま文書ベクトルを生成することだけでは不十分であると考えた．そこで，次の 4 つの前処理を行ってから，各文の文書ベクトルを生成することにした．

前処理 **1**：各単語を原形に戻し，原形が同じ単語を同一と扱う

- 例）「使って」→「使う」

例えば，「使わ (ない)」「使い (ます)」「使っ(て)」「使う」「使え (ば)」「使おう」などの活用形の違いは，いずれも本研究で扱うカテゴリ分類における事由文の意味を表現する上で大きな差はないと考えた．これらは原形に変形すると全て「使う」になり，同一に扱えるようになる．よって形態素解析器 JUMAN [16] を使って単語を原形に戻し，その種類別出現数からなる文書ベクトルを素性とする．なお，この前処理 1 を用いない場合は，単語をもとの表記のままで生成した文書ベクトルを素性とする．

前処理 **2**：同じ単語でも品詞が異なるものは異なる単語として扱う

- 例）「楽しみ」→「楽しみ名詞」

「楽しみ」という単語は，名詞の場合は『たのしいと感じること．また，たのしむ物事．趣味や娯楽.』という意味だが，形容動詞の場合は『たのしいこととして期待すること．また，そのさま.』というように品詞が変わると意味も変わる．このように，単語が表す意味は異なるにもかかわらず，表記は同じであるといったことがある．事由カテゴリを判別する上では，意味の違いは重要であるため，形態素解析器 JUMAN [16] を使って単語の品詞を特定し，単語だけでなく，その品詞との組み合わせの種類別出現数からなる文書ベクトルを素性とする．なお，この前処理 2 を用いない場合は，単語の種類別出現数からなる文書ベクトルを素性とする．

前処理 **3**：レシピの固有表現の出現数

- 例）「大根/F と 鶏肉/F を 土鍋/T に 入れ る」
  → F の出現数：2，T の出現数：1

表 1 の事由カテゴリにおける ab「材料・料理がある/ない」，cd「調理器具がある/ない」，および m「材料と材料の相性が良い」は，文中に食材名や料理名，調理器具・器材名が現れるかどうかが有効な特徴であると考えた．そこで，[17] で提案されたレシピに関する固有表現認識器を用いて，単語に食材 (F) や道具 (T) といったタグを付与し，入力文中に出現する F と T の数を素性とする．

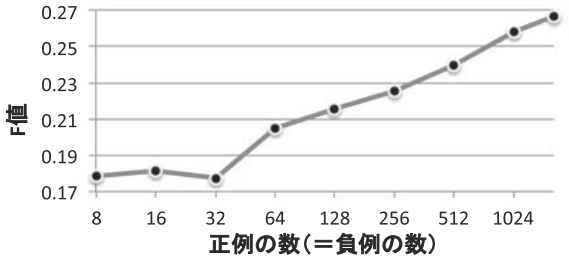


図 6: 事由抽出処理における学習用データ数とその F 値

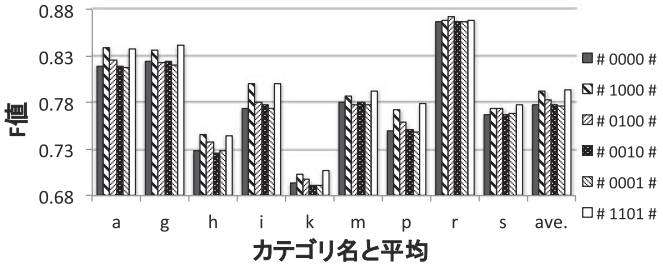


図 7: 事由カテゴリ分類処理における F 値

前処理 **4**：理由を表す語の出現有無情報を付加する

- 例）「今日は寒かった ので，お鍋にした.」
  → 理由語の有無：1

A の内容が B の内容の原因・理由であることを表すとき，A は B の理由節であるという．主節と理由節を結びつける用語には,「から」「ので」「ため」「おかげ」「せい」がよく使われる．このような用語は，同じく理由を示す事由文にはよく出現するが，事由でない文にはあまり出現しないと考え，これらの用語の出現の有無を素性とする．

識別器にとって，どの前処理が有効かを調べるため，前処理 1 から前処理 4 のすべての組み合わせについて，識別器の精度を評価した．

6.3 誕生・使用事由抽出処理の精度評価

誕生・使用事由抽出処理の精度が最も高かったのは前節で述べた前処理 1 と 4 を用いたときであった．学習用データが増加することによって F 値がどのように変化するかを調べるため，2^n の間隔で学習用データ数を変えながら F 値を計算した．その結果を図 6 に示す．図 6 からわかるように，学習データが増加するほど F 値は上昇することを確認した．学習用データに含まれる正例 1,628 文とそれと同数の負例を用いたとき，誕生・使用事由抽出処理の F 値は最も高くなり，その値は 0.266 であった．またこのとき，適合率は 0.163，再現率は 0.733，正解率は 0.756 であった．

6.4 誕生・使用事由カテゴリ分類処理の精度評価

各カテゴリの分類精度の F 値を図 7 に示す．グラフ右側に記されている “# 1101 #” のようなラベルは，どの前処理を用いた結果かを示している．左から n 番目の値は，前処理 n を用いたかどうかを表わす．“1” はそれに対応する前処理を用いたことを，“0” はそれに対応する前処理を用いなかったことをそれぞれ表わす．例えば，“# 1101 #” は，前処理 1,2,4 を用いたことを表している．ベースラインとして，どの前処理も用いなかった結果（“# 0000 #”），および 4 つの前処理のうち 1 つだけを用いた結果（“# 1000 #” など）を示す．4 つの前処理のすべての組み合わせの結果のうち，“# 1101 #” の結果が，最も平均 F 値が高く，0.794 であった．カテゴリごとに適した前処理の組み合わせを用いれば，精度はより改善すると考えられる．

表 3: 誕生・使用事由によるレシピ検索とレシピに対する AND 検索の満足度平均値と最大値（最大値は括弧内に記載）

シナリオの種類	誕生・使用事由による レシピ検索	レシピに対する AND 検索
【状況的 1】	2.48 (4.40)	2.46 (4.00)
【状況的 2】 *	3.48 (5.00)	1.70 (2.80)
【手続き的】 **	1.50 (5.00)	2.00 (5.00)
全体の平均 **	2.49 (4.80)	2.05 (3.93)

* 各被験者における検索結果 10 件の満足度の平均値が，検索システム間で 0.1%水準で有意差が見られた．

** 各被験者における検索結果 10 件の満足度の平均値が，検索システム間で 1%水準で有意差が見られた．

6.5 誕生・使用事由によるレシピ検索の有用性評価

本節では，提案手法が実現された場合に，日記文によるレシピ検索が従来のキーワードによるレシピ検索に比べてどの程度有用であるかを評価する．被検者実験により，誕生・使用事由によるレシピ検索の有用性を評価する．事由文の誤抽出による影響を避けるため，第 4.1 節で述べたデータセットを用いて事由文抽出は手動で行った結果を利用する．

第 3.2 節で示したようなシステム利用者のシナリオを 3 種類用意した．そのうちの 2 種類は，レシピに対する漠然とした要求を表したもので,「今日はとても忙しかった」のようなことしか説明できない利用者を想定している．以下，これら 2 種類のシナリオをそれぞれ「状況的 1」,「状況的 2」と呼ぶ．また，残り 1 種類は「安売りされていたさんまを買った」のように具体的な要求を持つ利用者を想定している．以下，このシナリオを「手続き的」と呼ぶ．5 名の被験者にこれらのシナリオを見せ，自分が今そのような状況にいると考えてもらう．次に，シナリオをもとに作成した日記クエリを第 5.4 節で述べたレシピ検索アルゴリズムに入力し，検索結果がどの程度満足するものであったか評価してもらう．さらにレシピに対する AND 検索による結果も同様に評価してもらい，誕生・使用事由によるレシピ検索結果と比較する．1 つのクエリにつき検索結果上位 10 件のそれぞれに対して，5 件法で評価してもらい，平均値と最大値を算出した．シナリオ [状況的 2] とそれをもとに作成した日記クエリと AND 検索に入力したクエリをそれぞれ次に示す．

- シナリオ

  私は 6 歳の女の子と 8 歳の男の子をもつ母親である．今は 12 月下旬である．明日，家族でクリスマスパーティーをするので，今日はクリスマスの装飾品やプレゼントを買いに行き，家の中の飾り付けを終えた．一日中動きっぱなしだったのでとても疲れた．まだ，パーティーでどのような料理を作るか決まっていない．

- 日記クエリ

  明日はクリスマスパーティー！
  飾り付けやプレゼント購入は終わりました。
  でもまだパーティーで作る料理が決まっていません。

- AND 検索に入力したクエリ

  クリスマス，パーティー，料理

評価結果を表 3 に示す．

シナリオ【状況的 2】に対する評価では，誕生・使用事由によるレシピ検索の方が，レシピに対する AND 検索よりも高い満足度が得られた．シナリオ【状況的 2】は「クリスマスパーティーの料理を探している人」を想定したものだが，誕生・使用事由によるレシピ検索で満足度が高かった検索結果は，クリスマスパーティーをしたときに作った料理を紹介したものであった．一方，シナリオ【手続き的】は「さんまを使った料理を探している人」を想定したものだが，レシピに対する AND 検索の方が満足度平均値は高かった．このことから，「クリスマス」「パーティー」のようにレシピ中には直接書かれないような語をクエリとする検索では，誕生・使用事由によるレシピ検索結果の方が高い満足度を得られるが，「さんま」のような食材名はレシピ中に現れることがあり，レシピに対する AND 検索でも十分な結果を得ることができると考える．

しかしながら，全体の平均をとると，満足度の平均値，最大値ともに誕生・使用事由によるレシピ検索の方が高いという結果を得た．特に，満足度の最大値の平均値は 4.80 と高いことから，検索結果上位 10 件のうち少なくとも 1 件は，ほとんどのユーザが満足するような結果であると言える．

今回の検索対象はブログ記事 1,000 件と少なかったが，対象記事数を増やすことでユーザビリティは向上すると考える．

7. まとめと今後の課題

本研究では，なぜそのレシピが誕生したのか，なぜそのレシピを使用したのかといった情報に着目し，それをもとに検索できる手法を提案した．ブログ型レシピのポータルサイトであるレシピブログの記事を分析した結果，レシピの誕生・使用事由を 18 のカテゴリに分類することができた．

また，ブログ記事から日記部分を抽出し，それとユーザが入力したコンテキスト情報から，識別器を用いることで事由文の抽出およびカテゴリ分類を行い，事由文およびそれが属するカテゴリをもとに，今のユーザの状況に合ったレシピを検索する手法を提案した．

第 6.5 節で行った評価によって，誕生・使用事由によるレシピ検索は，【状況的】シナリオに向いているという傾向が見られた．今後，さらに実験を行うことによって，誕生・使用事由によるレシピ検索は，どのようなユーザにとってより有用なものとなるのか明らかにしていきたい．

[謝辞]

本研究の一部は，文科省科研費基盤 (A)「ウエブ検索の意図検出と多元的検索意図指標にもとづく検索方式の研究」(24240013, 研究代表者：田中克己)，および文科省科研費基盤 (B)「消費者生産型レシピコンテンツの手順・記述から見た多様性の解析手法の提案」(26280039, 研究代表者：山肩洋子)，ならびに文科省科研費基盤 (B)「作業実施映像からの手順文書の自動生成」(26280084, 研究代表者：森信介）によるものです．ここに記して謝意を表します．

[文献]

[1] ヨシケイ開発株式会社． 夕食に関する意識・実態調査. http://www.yoshikei-dvlp.co.jp/archives/001/201311/2013.11.07 夫婦共働き家庭の夕食に関する意識調査.pdf.

[2] アイランド株式会社． レシピブログ． http://www.recipe-blog.jp/.

[3] 森信介, 山地治. 日本語の情報量の上限の推定. 情報処理学会論文誌, Vol. 38, No. 11, pp. 2191–2199, 1997.

[4] Shinsuke Mori, Tetsuro Sasada, Yoko Yamakata, and Koichiro Yoshino. A machine learning approach to recipe flow construction. In *Cooking with Computers*, 2012.

[5] クックパッド株式会社. クックパッド. http://cookpad.com/.

[6] 楽天株式会社. 楽天レシピ. http://recipe.rakuten.co.jp/.

[7] 上田真由美, 高畑麻理, 中島伸介. レシピ閲覧・摂食履歴を用いた嗜好の抽出. Web とデータベースに関するフォーラム (WebDB Forum 2011), 情報処理学会シンポジウムシリーズ, 3G-1-2, 2011.

[8] 森下幸俊, 中村富予. 気分による献立検索システムの検索軸の評価とレシピを活用した食品販売機能の市場ニーズの評価 (データ工学). 電子情報通信学会技術研究報告: 信学技報, Vol. 112, No. 75, pp. 79–84, 2012.

[9] 木原ひかり, 上田真由美, 宮脇佑介, 中島伸介. 食材の無駄を減らす料理レシピ群の推薦 (科学データストレージと応用処理システム, e-science and big data, 一般). 電子情報通信学会技術研究報告. DE, データ工学, Vol. 111, No. 361, pp. 37–42, 2011.

[10] 荘司慶行, 田中克己. 印象語をクエリとするリアクションに基づくウェブ情報検索. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3515–3526, 2011.

[11] 株式会社カカクコム. 食べログ. http://tabelog.com/.

[12] 大島裕明, 田中克己. 東京の寿司屋に対するオンラインレビュー情報の分析 (データ工学). 電子情報通信学会技術研究報告: 信学技報, Vol. 112, No. 75, pp. 49–54, 2012.

[13] Graham Neubig, 中田陽介, 森信介. 点推定と能動学習を用いた自動単語分割器の分野適応. 言語処理学会第 16 回年次大会 (NLP2010), 東京, Vol. 3, , 2010.

[14] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, Vol. 20, No. 3, pp. 273–297, 1995.

[15] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 2, No. 3, p. 27, 2011.

[16] 黒橋・河原研究室. 日本語形態素解析システム juman. http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN.

[17] 森信介, 山肩洋子, 笹田鉄郎, 前田浩邦. レシピテキストのためのフローグラフの定義. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2013, No. 13, pp. 1–7, nov 2013.

門脇 拓也 Takuya KADOWAKI

2014 年京都大学工学部情報学科卒業．京都大学大学院情報学研究科修士課程在学中．情報検索の研究に従事．日本データベース学会学生会員．

山肩 洋子 Yoko YAMAKATA

京都大学大学院情報学研究科 特定准教授．2005 年京都大学大学院情報学研究科博士後期課程単位認定退学．博士（情報学）．マルチメディア情報処理のなかでも，特に食メディア研究に従事．電子情報通信学会，人工知能学会各会員．

森 信介 Shinsuke MORI

1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了. 同年日本アイ・ビー・エム (株) 入社. 2007 年より京都大学学術情報メディアセンター准教授. 京都大学博士 (工学). 音声言語処理および自然言語処理に関する研究に従事. 1997 年情報処理学会山下記念研究賞受賞. 2010 年,2013 年情報処理学会論文賞受賞. 2010 年第 58 回電気科学技術奨励賞. 情報処理学会, 言語処理学会,ACL 各会員.

田中 克己 Katsumi TANAKA

京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．博士（工学）．主に，ウエブ情報検索，情報分析，データベース，マルチメディアコンテンツ処理の研究に従事．IEEE ComputerSociety，ACM，情報処理学会，人工知能学会，日本ソフトウェア科学会，日本データベース学会等各会員．