Cookpadのつくれぽ数の時間変動に基づく類似レシピ抽出法の提案

桐本　宙輝†　　風間　一洋†

† 和歌山大学 システム工学部　〒640–8510 和歌山県和歌山市栄谷 930
E-mail: †{s171063,kazama}@sys.wakayama-u.ac.jp

あらまし　人間の生活は 1 日，1 週間，1ヶ月，1 年などの何らかの周期性を持つことが多く，それに食事を調理して食べるという日常的な行動も大きく影響される．本稿では，Cookpad で公開されているレシピのつくれぽ数の頻度変化を各周期に合わせて畳み込んだ特徴ベクトルを用いて，レシピの周期性の有無と，レシピ間の時間的特性の類似性を抽出する手法と，レシピのカテゴリ木の階層構造を利用してクラスタの内容を要約する手法を提案する．実際に，それらの特徴ベクトルを k-means 法でクラスタリングすることで，バレンタインデーのようなイベントや，夏のような季節，平日や休日などの生活パターンに関連したクラスタが抽出されることを示す．

キーワード　Cookpad, つくれぽ, 代表カテゴリ, クラスタリング, 周期性

1. はじめに

インターネットや SNS の普及によって，Cookpad(注1) や楽天レシピ (注2) のようなユーザによるレシピ投稿サイトが活発に利用されるようになった．中でも Cookpad は現在ユーザ数は 5,000 万人，レシピ数は 224 万品を超えており，日々膨大な数のレシピが投稿されている．

Cookpad には，レシピを実際に作った感想を写真付きで投稿することで，レシピの作者への感謝の気持ちを示す「つくれぽ」という機能がある．レシピ数が 200 万品を突破した 2015 年 3 月 31 日時点のつくれぽの総数は 1,191 万件を超えており，非常に多くのユーザがつくれぽの機能を活用している．つくれぽの投稿は，数あるレシピの中から最も適切なレシピを選び，そのレシピに基づいて実際に調理し，出来上がった料理に満足して作者にわざわざ感謝の意を示すというコストの高い行動を，ユーザが意図的に起こしたことを意味する．そこで，レシピのつくれぽ数は人気度を表す指標として利用でき，Cookpad は 2015 年 12 月時点でつくれぽ数が 1,000 件を超えたレシピ 933 品を殿堂入りレシピとして公開している．

投稿されたレシピには，年間を通して食卓に出る日常的なレシピもあれば，ある時期に集中して作られるレシピもある．例えば，夏ならアイス，冬なら鍋のような季節的なレシピや，バレンタインデーのチョコレート，お正月のお節料理のようなイベント特有のレシピもある．提供する情報が一時的にしか参照されないようなニュースサイトのニュース記事と異なり，Cookpad のレシピは何度も繰り返し参照されることから，そのような特定の時間的状況に依存しているレシピのつくれぽは，毎年特定の季節や特定の日の直前に活発に投稿される傾向がある．本稿では，食生活に存在する 1 週間，1ヶ月，1 年のような周期性が，レシピのつくれぽ数の増加の度合いに大きく影響を与えていると仮定して，レシピのつくれぽ数の時間的変化から，ある一定の周期におけるレシピの時間的特性を表す特徴ベクトルを作成し，レシピ間の周期性の有無と時間的特性の類似性を抽出する方法を提案する．

また，時間的特性に基づいて抽出したクラスタのレシピは多岐に渡るために，クラスタリング結果の妥当性を判定することが難しい．そこで，レシピのカテゴリ木の階層構造を利用して部分木として集約し，その部分木の根の直感的な名前を持つ代表カテゴリを用いることで，クラスタの内容をわかりやすく要約する手法を提案する．

(注1)：http://cookpad.com/
(注2)：http://recipe.rakuten.co.jp/

2. 関連研究

類似レシピの抽出と，データの時系列解析という 2 つの観点から，それぞれ関連する研究を紹介する．

2.1 類似レシピの抽出

類似レシピの抽出方法に関しては様々な研究が存在する．

花井らは，レシピに出現する調理法，料理名，食材，調味料に基づいてクラスタリングすることで，類似レシピを抽出した [1]．また，杉山らは，構造化された調理手順から，類似レシピが共有する調理手順と各レシピ間の差異を発見することで，ある料理における典型的な調理手順と，各レシピの特徴をユーザに提示するシステムを提案した [2]．

これらの研究は，食材や調理手順のようなレシピの内容の類似性に基づくアプローチである．それに対して，本研究はつくれぽ数の頻度変化のような時間的特性に基づいて，作られる理由や状況，時期が類似するレシピを分類，推薦できる点や，ユーザが特に価値があると認めたレシピや，類似レシピの中でも特に代表的なレシピを扱うことができる点が異なる．

2.2 データの時系列解析

Twitter データの時系列解析では，風間らは Twitter における震災に関連した話題に用いられた単語を抽出し，それらの単語と他の単語との時系列的変化の類似性を Earth Mover's Distance で判定して，Twitter ユーザの話題や関心の変化を分析する手法を提案した [3]．大久保らは Web サーチエンジンの検索ログの各時間区間における検索語の個人内関連度と使用頻

度系列の相関係数を用いて同一情報に対して用いた検索語をグループ化することで，検索のトレンドを抽出する手法を提案した [4]．これらの研究では，東日本大震災という特別なイベントからの時間経過に伴う変化や検索語の一時的な流行などの現象に着目しているが，本研究では，食生活という日常的な周期性を持つ行動パターンに注目する点が異なる．

また，松林らは，アイテム購買履歴から作成した (ユーザ)×(時間)×(商品) という 3 次元のテンソルを，非負値テンソル因子分解 (NTF) することで，ユーザ単位の購買傾向と時系列特徴，商品購買傾向を合わせた分析を行うと共に，商品のランク比率を用いたグラフ可視化によりアイテム購買の季節性を確認した [5]．ただし，NTF により商品の購買データの潜在的な特徴をパターン抽出することはできても，抽出されたパターンの解釈は必ずしも容易ではない．これに対して，日常的に繰り返す調理行動は 1 日，1 週間，1ヶ月，1 年のような定まった周期性を持つことから，その周期性を活かして分析することで抽出結果の解釈が容易になるだけでなく，日常生活に密接したユーザ支援が容易になると考えられる．

3. レシピの時間的特性の類似性の分析

3.1 レシピの周期性

Cookpad には平日や週末，給料日前や給料日後，春夏秋冬，節句やクリスマスなどの特定の時期に作られるレシピが存在する．さらに，これらのレシピを様々なユーザが繰り返し作ることで周期性が生じる．このような時間的特性はレシピが利用されることで生み出されることから，レシピの内容からは必ずしも解るとは限らないので，レシピがいつどんな状況で作られたかを示すつくれぽ数の時間的変化を用いて推定を試みる．

一般に，周期性を判定するためにはフーリエ変換やウェーブレット変換が用いられることが多いが，つくれぽ数の頻度変化が比較的疎であるためにうまく処理できなかったり，任意の周波数に適用できる代わりに周波数の特定が難しいなどの問題が存在する．ただし，レシピの周期性は，そもそも人間の生活における 1 日，1 週間，1ヶ月，1 年単位の繰り返しから生じることを考慮すれば，つくれぽ数の時間的変化に存在する周期の種類は既知であり，逆にこれらの周期の合成として生まれていると考えることができる．

そこで，任意の期間のつくれぽ数の時間的変化を，1 週間，1ヶ月，1 年という周期に合わせて畳み込んでから正規化した特徴ベクトルを作成し，その距離に基づいて類似レシピを推薦したり，クラスタリングする手法を提案する．

3.2 特徴ベクトルの作成

まず，レシピの周期性を表す特徴ベクトルを作成する．本稿では使用するつくれぽの作成時間は 1 日単位なので，周期を 1 週間，1ヶ月，1 年の 3 種類とする．

レシピ r_i が期間 T_{r_i} のつくれぽの情報を持つ場合に，時刻 $t(0 \leqq t \leqq T_{r_i}-1)$ のつくれぽ数を $f_{r_i}^{(t)}$ として，次元数 T_{r_i} のつくれぽ数の頻度ベクトル $\boldsymbol{f}_{r_i}$ を次の式で表す．

$$\boldsymbol{f}_{r_i} = (f_{r_i}^{(0)}, f_{r_i}^{(1)}, \ldots, f_{r_i}^{(T_{r_i}-1)}) \tag{1}$$

次に，この頻度ベクトル $\boldsymbol{f}_{r_i}$ を，次元数 C に畳み込んだベクトル $\boldsymbol{v}_{r_i,C}$ を作成する．次元数 C は，1 週間の場合は $C=7$，1ヶ月の場合は $C=31$，1 年の場合は $C=366$ とする．月や年の場合には日数が異なるので，次元数は大きい方に合わせる．さらに，$\boldsymbol{f}_{r_i}$ の時刻 t を $\boldsymbol{v}_{r_i}$ の $0 \sim C-1$ の範囲のインデックス値に畳み込む関数を $fold(t,C)$ とする．この関数は，$C=7$ の場合は月曜日～日曜日をそれぞれ $0 \sim 6$ に，$C=31$ の場合はその月で何日目かを示す値 $1 \sim 31$ をそれぞれ $0 \sim 30$ に，$C=366$ の場合はその年で何日目かを示す値 $1 \sim 366$ をそれぞれ $0 \sim 365$ に変換する．ベクトル $\boldsymbol{v}_{r_i,C}$ は以下のように計算する．

$$\begin{cases} \boldsymbol{v}_{r_i,C} &= (v_{r_i,C}^{(0)}, v_{r_i,C}^{(1)}, \ldots, v_{r_i,C}^{(C-1)}) \\ v_{r_i,C}^{(j)} &= \frac{sum(\{f_{r_i}^{(t)} | fold(t,C)=j\})}{|\{f_{r_i}^{(t)} | fold(t,C)=j\}|} \end{cases} \tag{2}$$

ここで，$sum(S)$ は集合 S の要素の値の総和を求める関数である．

最後に，つくれぽ数のスケールの影響を受けずに時間的特性だけで類似性を判定できるように，$\boldsymbol{v}_{r_i,C}$ をすべての要素の和が 1 になるように正規化して，レシピ r_i に対する周期 C の特徴ベクトル $\boldsymbol{p}_{r_i,C}$ を求める．

$$\begin{cases} \boldsymbol{p}_{r_i,C} &= (p_{r_i,C}^{(0)}, p_{r_i,C}^{(1)}, \ldots, p_{r_i,C}^{(C-1)}) \\ p_{r_i,C}^{(j)} &= \frac{v_{r_i,C}^{(j)}}{\sum_{k=0}^{C-1} v_{r_i,C}^{(k)}} \end{cases} \tag{3}$$

この結果，$\boldsymbol{p}_{r_i,C}$ は，つくれぽの周期 C の各要素に相当する時点における生起確率を表すことになる．なお，以降では $C=7$ の場合の特徴ベクトルを週ベクトル，$C=31$ の場合を月ベクトル，$C=366$ の場合を年ベクトルと呼ぶ．

3.3 特徴ベクトルの特徴と制限

レシピのつくれぽ総数が比較的多くても，つくれぽ数の頻度ベクトルは疎であることが多いが，畳み込みにより作成された特徴ベクトルは密になると同時に，時間的な特徴が強調されるという利点がある．これは頻度ベクトルを移動平均などの手法でスムージングして扱うよりも適切であると考えられる．

ただし，この特徴ベクトルを用いる場合には，収集期間が周期より長くなければいけない．周期よりも短い場合は，データの不足のために特徴ベクトルの要素の多くが 0 となり，時間的特性の判定を正確に行うことができないので，複数の周期を用いて分析することを想定した場合には，最長の周期以上でなければいけない．つまり，本稿で用いる最長の周期は 1 年であることから，366 日以上の収集期間のレシピだけを対象とする．

3.4 レシピ間の時間的特性の距離の計算

周期 C を想定した場合の二つのレシピ r_i と r_j の時間的特性の距離 $d_C(r_i,r_j)$ は，それぞれの特徴ベクトル $\boldsymbol{p}_{r_i,C}, \boldsymbol{p}_{r_j,C}$ の間のユークリッド距離を用いて次のように計算する．

$$d_C(r_i,r_j) = \sqrt{\sum_{k=0}^{C-1} (p_{r_i,C}^{(k)} - p_{r_j,C}^{(k)})^2} \tag{4}$$

この距離が小さいほど，二つのレシピの周期性が類似していると見なす．

3.5 レシピのクラスタリング

本稿では，周期 C を想定した場合の二つのレシピ r_i と r_j の距離 $d_C(r_i, r_j)$ を用いて，時間的特性が類似しているレシピ群のクラスタリングを試みる．クラスタリング手法としては，一般的な k-means 法 [6] を用いる．k-means 法は単純なことから，レシピ間の類似度を時間的特性で分離できるかの確認に適していると言える．

各クラスタの時間的特性は，そのクラスタに含まれるレシピの代表的な特徴を表す中心座標の特徴ベクトルをグラフとして可視化することで表す．このグラフを見れば，クラスタに含まれているレシピはどのような季節や日によく作られるのか，あるいは年間を通して日常的に作られるのか，などを容易に把握できる．

4. クラスタ内容の要約

4.1 レシピのカテゴリ階層の利用

提案した特徴ベクトルの類似性は，レシピに含まれる単語や食材名ではなく時間的特性に基づくことから，抽出されたクラスタのレシピは多岐に渡り，クラスタの妥当性を判定することは必ずしも容易ではない．

そこで，レシピのカテゴリに注目する．Cookpad のレシピのカテゴリは約 1,100 個あり，投稿者を含むユーザによって推薦されたカテゴリ候補を，クックパッド編集部が審査した上で登録される．カテゴリは木構造であるが，図書分類法のような明確な基準がないことから，カテゴリ名の重複や異なる分類方針の混在などの問題点が存在する反面,「おもてなし料理」や「夏のさっぱりおかず」などの主観的で特徴がわかりやすい名称が付けられており，これらの表現を用いればクラスタの内容をより直感的に伝えることができる．

ただし，このような表現はカテゴリ階層の比較的上位で使われているのに対して，大部分のレシピは食材名や料理名などの下位のカテゴリ名が付与されている．これはレシピの閲覧という観点では特に問題はないが，クラスタ内容を単純にカテゴリ名で表現すると，カテゴリ数が膨大になる上に，食材名や料理名が大量に現れてクラスタの性質がよくわからないという問題が生じる．

そこで，本稿では，Cookpad のカテゴリ木の階層構造を利用して，複数のカテゴリをまとめて階層の上位の直感的なカテゴリ名を持つ代表カテゴリで表すことで，クラスタの内容をわかりやすく要約する手法を提案する．

4.2 代表カテゴリ選出法

代表カテゴリはカテゴリ木の階層を根の方向に登って抽出するが，その停止判定には注目しているカテゴリを頂点とする部分木にレシピが存在する確率を用いる．

あるクラスタ中のカテゴリ i を頂点とする部分木に存在するカテゴリが付与されたレシピ数を C_i，カテゴリが付与されたレシピ総数を C_{all} とすると，部分木のレシピ存在確率 RP_i は次の式で求められる．

$$RP_i = \frac{C_i}{C_{all}} \tag{5}$$

さらに，カテゴリ i を頂点とする部分グラフに登録されているレシピ数を S_i，カテゴリが付与されたレシピ総数を S_{all} として，RP_i とレシピにカテゴリがランダムに付与されたと仮定した場合の予測確率 EP_i との比率 R_i を求める．

$$EP_i = \frac{S_i}{S_{all}} \tag{6}$$

$$R_i = \frac{RP_i}{EP_i} \tag{7}$$

一般にレシピはカテゴリ木中に局在するが，カテゴリ i の R_i が 1 より大きいほど，その部分木にレシピが集中しているとみなす．さらに，レシピの存在するカテゴリから根の方向に移動した場合に，代表カテゴリとして妥当でないカテゴリに到達した場合には R_i が有意に減少するので，その前のカテゴリで停止する．

抽出したクラスタの代表カテゴリリストを選出する手順を以下に示す．なお，T_1, T_2 は定数である．

（1） クラスタに含まれるレシピのカテゴリ群からカテゴリ i を取り出す．$R_i < T_1$ の場合は次のカテゴリ i を取り出す．カテゴリが存在しない場合は終了する．

（2） カテゴリ i から上の階層に移動しながら，$R_j \geqq T_1 \wedge R_j/R_i \geqq T_2$ を満たすカテゴリ j を根まで探索する．

（3） 見つかった場合は根に最も近いカテゴリ j を，そうでなければカテゴリ i を代表カテゴリリストに追加する．

（4）（1）に戻る．

なお，集約した代表カテゴリの数が多くなる可能性があるだけでなく，代表カテゴリを根とする部分木に登録されているレシピ数が少ない場合には，クラスタ内容を表すためには適切でないことが考えられるので，代表カテゴリのレシピ数で降順に並び替えて，その上位の代表カテゴリでクラスタの内容を表現する．

5. 評　　価

5.1 Cookpad データセット

Cookpad から提供された 1998 年 4 月 23 日から 2014 年 9 月 30 日までの 1,715,595 件のレシピから，つくれぽ数が 100 件以上あり，つくれぽ投稿期間が 366 日以上の 11,507 件のレシピを抽出して，評価に用いた．なお，レシピのレシピ ID，レシピ名，カテゴリ，つくれぽ投稿日と投稿数だけを使用する．

5.2 クラスタの時間的特性の分析

本稿では，つくれぽ数の時間的変化を周期に合わせて畳み込んでから正規化した特徴ベクトルを使用する．このベクトルの各要素は周期の開始時点からの経過時刻におけるつくれぽの生起確率を表すことから，年ベクトルの場合は 1 月 1 日からの経過日数，月ベクトルの場合はその月の日，週ベクトルの場合は曜日と特徴ベクトルの生起確率分布を照合することで，各レシピの時間的特性を容易に分析できる．

同様に，各クラスタの特徴はクラスタの中心が最もよく表すと考えて，クラスタの中心座標の特徴ベクトルの生起確率分布を用いれば，各クラスタの時間的特性を分析することができる．なお，周期 C を想定した場合のクラスタ数 k_C には，値を変え

表 1　イベント性を示すクラスタのピーク日

ピーク日	関連イベント
1 月 1 日	正月
1 月 7 日	人日の節句
2 月 3 日	節分
2 月 14 日	バレンタインデー
3 月 3 日	ひな祭り
5 月 5 日	こどもの日
10 月 31 日	ハロウィン
12 月 25 日	クリスマス

表 2　季節性を示すクラスタのピーク日

盛り上がり時期	関連季節
4 月上旬〜6 月中旬	春
5 月下旬〜8 月中旬	春から夏
7 月上旬〜9 月下旬	夏
9 月中旬〜11 月下旬	秋
11 月上旬〜3 月中旬	冬
1 月下旬〜5 月上旬	冬から春

て何度かクラスタリングを行い，特徴ベクトルや内容が類似しているクラスタが複数出現しない，つまり過分割されない範囲の中の最大値を用いた．

5.2.1　年ベクトルを用いた場合

クラスタ数 k_{366} を 15 に設定して抽出されたクラスタは，特定の数日に集中する局所的なピークを示すようなイベント性を示すクラスタ，数ヶ月に渡る緩やかな盛り上がりを示すような季節性を示すクラスタ，明確な周期性がないクラスタの 3 種類に分類できた．各種類ごとに代表的なクラスタの中心座標の年ベクトルの生起確率分布を図 1 に示す．なお，グラフの横軸は 1 月 1 日からの経過日数，縦軸はその日のつくれぽの生起確率である．

イベント性を示すクラスタは 8 個抽出された．これらのピークの生起確率は他の日と比べて顕著であることから，そのクラスタはピーク日の前後のイベントのためにだけ作られる料理のレシピ群ではないかと推測できる．クラスタの生起確率がピークになる日と，そのピークに関連していると推測できるイベントを表 1 に示す．例えば，図 1(a) は，2 月 14 日に最大のピークを示すことから，バレンタインデーと関係があるクラスタであると思われる．なお，周期の終わりの時点にも小さなピークが見られるが，これは 12 月 25 日であることから，クリスマスにも弱い関連があると思われる．

季節性を示すクラスタは 6 個抽出された．この盛り上がりにおける生起確率の変動は緩やかであることから，食材の入手時期や気候に関連した料理のレシピ群ではないかと推測できる．クラスタの生起確率が盛り上がる時期と，その盛り上がりに関連していると推測できる季節を表 2 に示す．例えば，図 1(b) は，7 月上旬〜9 月下旬に盛り上がることから，夏に関係があるクラスタであると推測できる．ただし，1 月下旬〜5 月上旬に掛けては，他のクラスタと比較して盛り上がりは小さかった．

5.2.2　月ベクトルを用いた場合

クラスタ数 k_{31} を 10 に設定して抽出したクラスタは，年ベクトルの場合と同様に特定の数日に集中する局所ピークや緩やかな盛り上がりが見られたものの，より複雑であった．代表的なクラスタの中心座標の月ベクトルの生起確率分布を図 2 に示す．なお，グラフの横軸は月の 1 日からの経過日数，縦軸はその日のつくれぽの生起確率である．

局所ピークを示すクラスタは，図 2(a) のようなピーク前後の増減がある 5 個のクラスタと，図 2(b) のように月末のピークを除くと生起確率の変動がほとんどないクラスタに分類できた．前者のピーク日を調べると，表 1 に示したような年ベクトルの場合のピーク日と合致していることから，年ベクトルの生起確率のピーク値が高すぎるために，月ベクトルに畳み込んだ場合でも影響を受けているのでないかと推測できる．後者の場合は，ピークは存在するものの，それ以外の日の生起確率がほぼ同じであるという特徴を持っていた．

緩やかな盛り上がりを示すクラスタは，図 2(c) の上旬，図 2(d) の中旬，図 2(e) の下旬の 3 個あった．これらのクラスタは，一ヶ月の周期性を持っていると推測できる．

5.2.3　週ベクトルを用いた場合

クラスタ数 k_7 を 3 に設定して抽出されたクラスタは，どれも緩やかな盛り上がりを示した．各クラスタの中心座標の週ベクトルの生起確率分布を図 3 に示す．なお，グラフの横軸は月曜日からの経過日数，縦軸はその日のつくれぽの生起確率である．

盛り上がりを見ると，まず週末である土曜日と日曜日の生起確率が高いクラスタと低いクラスタに分けることができる．これから，時間的余裕のある休日と，忙しい平日とでは，異なる料理を作っているのではないかと推測できる．さらに後者は平日である月曜日から金曜日までの期間の前半と後半に盛り上がりがあるものに分けることができる．つまり，同じ平日でも，前半と後半では作る料理に違いがあるのでないかと推測できる．

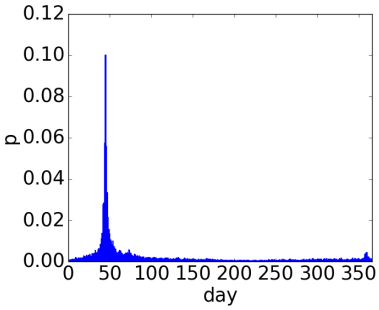
5.3　代表カテゴリによる要約の分析

次に，代表カテゴリによるクラスタの内容の要約が妥当であるかについて分析する．なお，$T_1 = 3.0, T_2 = 0.3$ とした．これは，全体と比較してあるクラスタで有意に出現し，なおかつ代表カテゴリを決める際に過剰にカテゴリ木を登りすぎない数値として設定した．

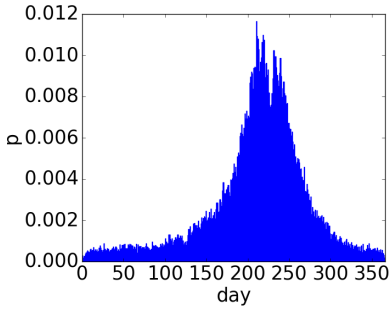
年ベクトルでのクラスタリングにおいて抽出された，冬 (11 月上旬〜3 月中旬) の季節性を示したクラスタに対して，単純にカテゴリを用いた場合と，代表カテゴリで集約した場合を比較する．それぞれのカテゴリに属するレシピ数で降順に並び替えて，その上位 10 件を表 3 に示す．なお，代表カテゴリにおけるカテゴリ数とは，いくつのカテゴリを集約しているかを示す数である．

単純にカテゴリを用いた場合は 5 位の「お餅活用レシピ」や 9 位の「冬！おもてなし料理」でかろうじて冬に関するクラスタであると予測できるが，それ以外は食材が多く，どういった性質を持ったクラスタなのかよくわからない．一方で，代表カテゴリに集約した場合は野菜に関するカテゴリが「旬野菜（秋・冬）」にまとめられ，どの季節に関連しているのか一目瞭然である．

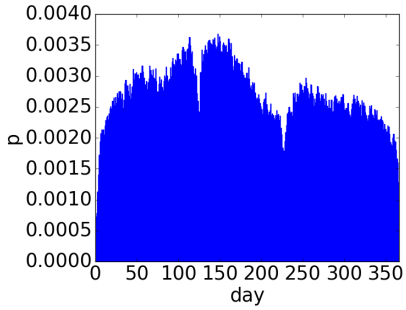
また，鍋に関するカテゴリは細分化されているので，単純に



(a) イベント性を示すクラスタ

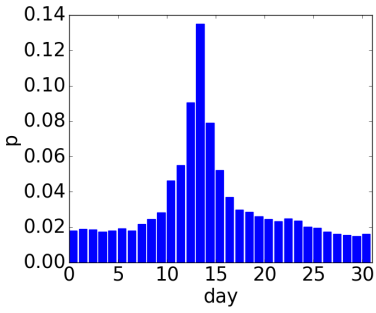


(b) 季節性を示すクラスタ

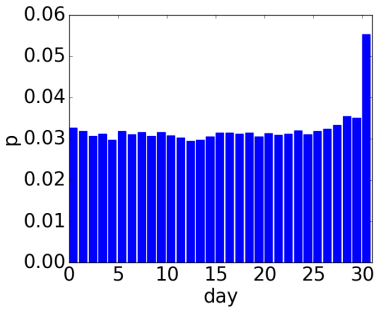


(c) 明確な周期性がないクラスタ

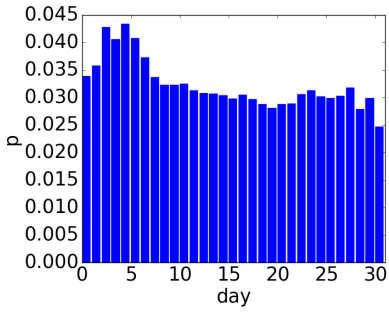
図 1　クラスタの年ベクトルの生起確率分布



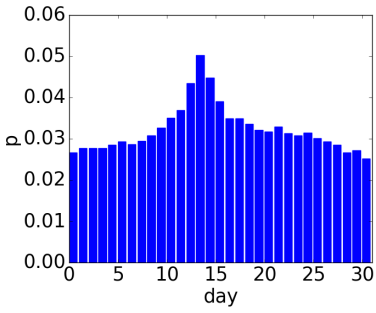
(a) 局所ピークがあるクラスタ (1)



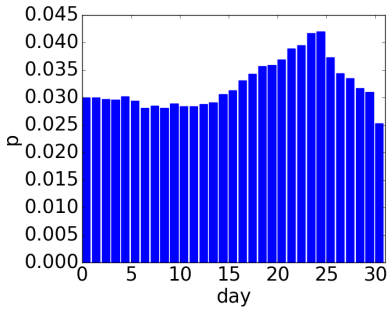
(b) 局所ピークがあるクラスタ (2)



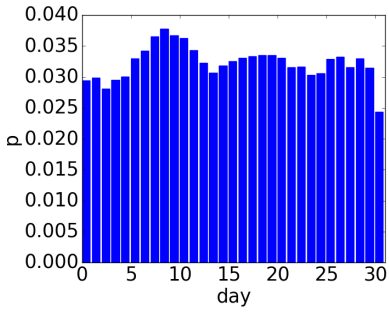
(c) 盛り上がりがあるクラスタ (1)



(d) 盛り上がりがあるクラスタ (2)

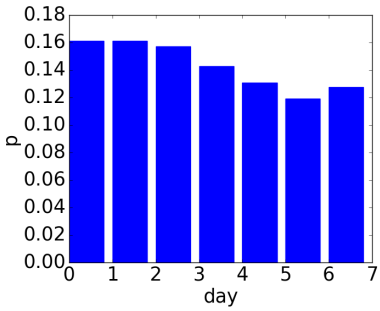


(e) 盛り上がりがあるクラスタ (3)

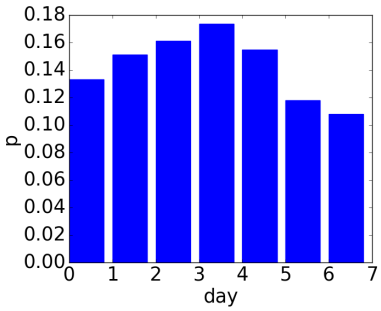


(f) 明確な周期性がないクラスタ

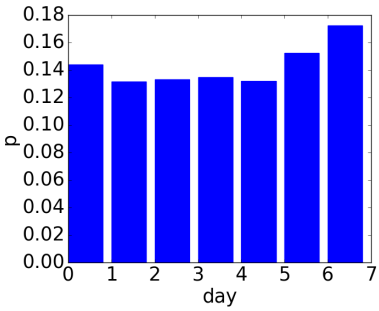
図 2　クラスタの月ベクトルの生起確率分布



(a) 盛り上がりがあるクラスタ (1)



(b) 盛り上がりがあるクラスタ (2)



(c) 盛り上がりがあるクラスタ (3)

図 3　クラスタの週ベクトルの生起確率分布

カテゴリを用いた場合は各カテゴリの頻度が低くなり，上位に現れない．しかし，代表カテゴリとして「鍋もの」にまとめられ頻度が高くなることで，このクラスタに鍋料理が多いという特徴がわかるようになる．

ただし，カテゴリが適切であれば，そのまま代表カテゴリとして抽出される．例えば,「冬！おもてなし料理」がある．このカテゴリは子カテゴリを持っておらず，親カテゴリは「おもてなし料理」で季節性を持たないことから，上位のカテゴリにま

表 3　単純にカテゴリを用いた場合と代表カテゴリに集約した場合の比較

	カテゴリ		代表カテゴリ		
順位	カテゴリ名	レシピ数	代表カテゴリ名	レシピ数	カテゴリ数
1	白菜	64	旬野菜（秋・冬）	214	9
2	大根	64	鍋もの	38	17
3	れんこん	29	その他のお菓子	36	5
4	フルーツのお菓子	23	人気のキャベツレシピ	23	34
5	お餅活用レシピ	21	お正月の料理	23	5
6	里芋	17	パイ	21	4
7	ホットケーキミックスで作れるケーキ	16	ホットケーキミックスで作れるケーキ	16	1
8	かぶ	16	冬！おもてなし料理	15	1
9	冬！おもてなし料理	15	スポンジケーキ	15	3
10	アップルパイ	15	ぶり	12	1

とめられない．このような場合は，単一カテゴリでも代表カテゴリになる．

食材が単一で出現する例として,「ぶり」がある．これは，ぶりの旬は 12 月から 1 月で冬の季節性があるのに対して，その親カテゴリの「魚介のおかず」は様々な魚介類の親カテゴリになっているため，明確な季節性が存在しないからである．

これらの結果から，カテゴリ木を登る際の停止判定は効果的に機能しており，妥当なカテゴリを代表カテゴリとして選出できていることが確認できた．また，代表カテゴリによるクラスタの要約によって，クラスタの性質が容易に把握できるようになった．

5.4　クラスタの内容の分析

次に，抽出されたクラスタに対して代表カテゴリを部分グラフに含まれるレシピ数で降順に並び替えて，その上位 5 件の代表カテゴリ名からクラスタの内容を推測する．さらに実際にクラスタに含まれるレシピを調べることで，その推測が正しいかどうかを確認する．

5.4.1　年ベクトルの場合

図 1 に示した各クラスタから抽出した代表カテゴリをレシピ数の降順に並び替えた上位 5 件を，表 4 に示す．

まず，イベント性を示すクラスタの代表カテゴリを図 1(a) に示す．どの代表カテゴリもチョコレートやお菓子に関係しているが，表 1 のピーク日と各レシピの内容から，バレンタインデーに関連したクラスタであることがわかる．また，レシピ数は 169，カテゴリ数は 37，代表カテゴリ数は 6 であり，少数の代表カテゴリに集約されていることがわかる．4 位や 5 位に出現した代表カテゴリは属するレシピの数が非常に少なく，代表カテゴリとして適していないと判断できる．これらの特徴は，他のイベント性を示すクラスタも同様の傾向が見られた．

次に，季節性を示すクラスタの代表カテゴリを図 1(b) に示す．代表カテゴリは夏に関係しているものと冷たい麺料理に関係しているものが存在するが，表 2 の関連季節が 7 月上旬から 9 月下旬であることと各レシピの内容から，夏に関連したクラスタであることがわかる．また，レシピ数は 438，カテゴリ数は 128，代表カテゴリ数は 47 であり，イベント性を示すクラスタよりも多くの代表カテゴリに分散していることがわかる．他の季節性を示すクラスタでも，同様の傾向が見られた．

次に，明確な周期性がないクラスタの代表カテゴリを図 1(c) に示す．お弁当に関する代表カテゴリが 2 つあるが，他はキャベツ料理，再現料理，子供向け料理と，ほとんどが日常的に作られるレシピであった．唯一時間的特性を示すと思われる「運動会のお弁当」カテゴリに含まれるレシピを調べたが，特に運動会に限定されるようなレシピはなく,「お弁当のメインおかず」よりも凝ったレシピが含まれており，これも日常的に調理されている可能性が高かった．

5.4.2　月ベクトルの場合

図 2 に示した各クラスタから抽出した代表カテゴリをレシピ数の降順に並び替えた上位 5 件を，表 5 に示す．

まず，局所ピークがあるクラスタについて分析する．図 2(a) に示した 14 日付近にピークが現れるレシピで構成されたクラスタは「チョコレートのお菓子」という代表カテゴリに属するレシピが最も多く，それに他のお菓子が続いた．14 日にピークが出現していることを合わせて考慮すると，バレンタインデーの影響を受けていることは明らかである．他の明確なピークが現れたクラスタの代表カテゴリも，年ベクトルで抽出されたイベント性のあるクラスタの代表カテゴリと酷似していた．

5.2.2 で月末のピークを除くと生起確率の変動がほとんどないと述べた図 2(b) のクラスタは,「野菜を使ったのお菓子」というカテゴリが上位に入っていた．これはハロウィン (10 月 31 日) にかぼちゃを使ったお菓子が作られており，その影響で 31 日にピークが現れたと考えられる．一方で，平坦な部分を構成しているのはクラスタに含まれている大部分の日常的なレシピであると思われる．

次に緩やかな盛り上がりがあるクラスタについて分析する．図 2(c) のクラスタは「お餅」という代表カテゴリが出現していた．このクラスタのピークが上旬にあることからも，1 月の上旬あたりにお餅を食べる習慣を反映していると言える．

図 2(d) のクラスタはお菓子類の代表カテゴリが多く，2 番目に多い代表カテゴリが「チョコレートのお菓子」であった．このクラスタのピークが中旬にあることからも，バレンタインデーの影響を受けている可能性が高い．しかし，バレンタインデーに関しては前述した図 2(a) のクラスタが抽出されている．代表カテゴリの内容でこれらを比較すると，図 2(d) の方は「チョコレートのお菓子」よりも「クッキー」の方が多いという点で異なっている．このことから，普段から作られるが，バレンタインデーにやや盛り上がりをみせるようなお菓子のレシピがこの

表 4　年ベクトルの場合の代表カテゴリ

	図 1(a)			図 1(b)			図 1(c)		
順位	代表カテゴリ	レシピ数	カテゴリ数	代表カテゴリ	レシピ数	カテゴリ数	代表カテゴリ	レシピ数	カテゴリ数
1	チョコレートのお菓子	103	11	旬野菜（夏）	149	15	人気のキャベツレシピ	151	48
2	もっと料理を楽しむ	36	23	夏に食べたい料理	21	13	お弁当のメインおかず	52	1
3	おもてなしデザート	9	4	そうめん	17	3	再現レシピ	38	22
4	持ち寄り・プレゼント	2	2	夏！おもてなし料理	15	1	子どもが喜ぶ♪かわいいおかず	37	1
5	シリコンスチーマーでお菓子	1	1	冷製・アイデアパスタ	10	3	運動会のお弁当	34	8

表 5　月ベクトルの場合の代表カテゴリ

	図 2(a)			図 2(b)			図 2(c)		
順位	代表カテゴリ	レシピ数	カテゴリ数	代表カテゴリ	レシピ数	カテゴリ数	代表カテゴリ	レシピ数	カテゴリ数
1	チョコレートのお菓子	65	10	人気のキャベツレシピ	34	43	人気のキャベツレシピ	19	38
2	もっと料理を楽しむ	19	17	野菜を使ったお菓子	25	2	長ねぎ	10	1
3	おもてなしデザート	5	3	夏に食べたい料理	22	12	お餅	9	2
4	カップケーキ	4	1	オクラ	14	1	おから	9	1
5	ザッハトルテ	3	1	ヨーグルトのお菓子	9	1	子どものパーティ	8	1
	図 2(d)			図 2(e)			図 2(f)		
1	クッキー	79	5	人気のキャベツレシピ	38	44	人気のキャベツレシピ	106	48
2	チョコレートのお菓子	59	8	ピザ	16	1	再現レシピ	26	21
3	おもてなしデザート	35	4	おもてなしの冷副菜	16	1	運動会のお弁当	23	8
4	ホットケーキミックスを使ったお菓子	34	4	スポンジケーキ	15	2	食材ひとつでおかず	18	1
5	人気のキャベツレシピ	26	39	秋におすすめ料理	14	12	中華料理	15	5

クラスタに分類されていると考えられる．このようなレシピは，時間的特性でいえば図 2(a) のクラスタよりもこちらに近いので分類結果は妥当だが，1ヶ月内で中旬にピークがくるような周期性を持っているとは言えないので影響を取り除きたい．

図 2(e) のクラスタは「スポンジケーキ」という代表カテゴリが出現している．このクラスタのピークが下旬にあることからも，クリスマスに関連している可能性が高いが，バレンタインデーの場合と同様にクリスマスもピークが顕著なクラスタが抽出されていた．そのため，この「スポンジケーキ」に属するレシピも普段から作られるが，クリスマスにやや盛り上がりをみせるレシピであると思われる．

明確な周期性がない図 2(f) のクラスタは，年ベクトルにおいて明確な周期性がない図 1(c) の代表カテゴリと類似している．よって，これらは 1ヶ月内での時期に関係なく作られるカテゴリであると考えられる．

上旬，中旬，下旬，そして明確な周期性がみられないクラスタに「人気のキャベツレシピ」が代表カテゴリとして現れている．キャベツ料理は日常的に食卓に出るレシピが多いと 5.4.1 で述べたが，月ベクトルにおいても頻出したことから，やはり日常的なレシピが多く属するカテゴリであると推測できる．

5.4.3　週ベクトルの場合

図 3 に示した各クラスタから抽出した代表カテゴリをレシピ数の降順に並び替えた上位 5 件を，表 6 に示す．

まず，平日は一般的な料理やお弁当に関係する代表カテゴリが，週末はお菓子や再現レシピに関係する代表カテゴリが抽出されていることがわかる．これは仕事や学校がなく時間的に余裕のある週末に，お菓子を作ったり，手間が掛かる再現レシピに挑戦していると推測できる．

さらに，平日前半に盛り上がりがあるクラスタと，後半に盛り上がりがあるクラスタの代表カテゴリを比較すると，どちらも普段の食事やお弁当のレシピであったが，前半に「造り置き・冷凍できるおかず」という代表カテゴリがあった．これは，週の初めにおかずをまとめて作り置きをする食習慣がある家庭が多いことを表していると思われる．また，同じお弁当のレシピでも，平日前半は「運動会のお弁当」，後半は「お弁当」という違いがあった．これは平日の前半は凝ったお弁当を作るが，後半は簡単なお弁当にする人が多いか，また凝ったお弁当のおかずを平日の初めに作り置きしているのではないかと思われる．

6. 考　　察

時間的特性を用いたクラスタリングと，代表カテゴリを用いたクラスタ要約によって，人間がレシピとそのカテゴリ名を見ただけでは容易にわからないような特徴や関係を見出すことが可能になった．例えば，同じ時期に作られる異なる種類の料理の発見や，食材や調理法ではなく時期や料理される状況という観点における妥当なカテゴリ分類はもちろん，「運動会のお弁当」と「お弁当」のような，「運動会」という分類意図とは異なる使い分けがなされている可能性を示唆している場合を発見することが可能になった．

今回，図 1(a) に示したようなイベント性のあるクラスタのピーク日は対応するイベント日と一致したが，常に一致するとは限らないと思われる．この理由は，例えばバレンタインデーのチョコレートは事前に作成して 2 月 14 日に渡すような前倒し要因があるからである．ただし，料理を作成した後日につくれぽを書くという遅延要因もあり，今回は偶然相殺されたと推測している．

「人気のキャベツレシピ」は，多くのクラスタの代表カテゴリとして登場していた．この代表カテゴリの階層は自身を含め

表 6　週ベクトルの場合の代表カテゴリ

	図 3(a)			図 3(b)			図 3(c)		
順位	代表カテゴリ	レシピ数	カテゴリ数	代表カテゴリ	レシピ数	カテゴリ数	代表カテゴリ	レシピ数	カテゴリ数
1	人気のキャベツレシピ	98	47	お弁当	152	14	クッキー	137	4
2	運動会のお弁当	30	11	鶏むね肉	78	1	ホットケーキミックスを使ったお菓子	92	3
3	作り置き・冷凍できるおかず	23	1	人気のキャベツレシピ	73	47	おもてなしデザート	71	3
4	ヘルシーおかず	18	6	アレンジ豆腐料理	60	1	人気のキャベツレシピ	54	43
5	中華料理	18	6	豚こま切れ肉	31	1	再現レシピ	46	29

て 4 段で，合計 49 個のカテゴリがあり,「お肉」,「野菜」,「魚介」,「ごはんもの」などの食材カテゴリや,「主食」,「おかず」,「副菜」といった食卓での役割のカテゴリ,「定番メニュー」という日常的なレシピのカテゴリが存在し，カテゴリ木の他の部分と重複している名前が多かった．キャベツは一年中入手できる一般的な野菜であることを考えると，そのレシピの多様性と調理頻度の高さから，多くのクラスタに現れるのではないかと思われる．

なお，月ベクトルを用いた場合に，年ベクトルの影響も強く現れるという問題が生じていた．提案手法ではつくれぽ数の時間的変化が複数の周期の合成によって表されていると仮定したが，畳み込みだけでは個々の周期性を明確に分離することはできない．そこで，特に 1ヶ月と 1 年のように周期が合致している場合に，畳み込みで時間的特性がより強く強調されやすい周期が短い方の特徴ベクトルで問題が顕在化したと思われる．ただし，1 週間と 1 年のように周期が明らかにずれている場合は，特にこのような問題は見られなかった．

7. おわりに

本研究では，1 年，1ヶ月，1 週間といった一定周期のつくれぽ数の頻度変化からレシピの周期性の有無と時間的特性の類似性を抽出することが可能であり，人々の食生活はこのような周期に影響を受けていることを示した．

つくれぽの頻度変化を分析したい周期に合わせて畳み込んだ特徴ベクトルを用い，生起確率分布を可視化することは，時間的特性の類似性を判定するだけでなく，レシピやクラスタの時間的特性を直感的に理解するために有用であることがわかった．例えば，生起確率分布を適切にデフォルメ・彩色してユーザに提示できれば，レシピやクラスタの理解を支援できる．

なお，今回は特徴ベクトルの類似性に基づいてクラスタリングしたが，あるレシピに対して時間的特性の類似したレシピを推薦することも可能であり,「日常生活の同じ時間的状況で料理を作る」という事実に基づいてレシピを推薦する，新しい種類の協調フィルタリングが実現できると考えている．ただし，従来のテキストや材料などの類似性に基づく場合より広範囲の推薦が可能であるために，一旦代表カテゴリとしてまとめてから，ユーザの目的に合致したものを優先的に表示するなどの工夫が必要であると思われる．

しかし，課題も残っている．月ベクトルにおけるクラスタリングでは，1 年周期でのイベントの影響を強く受けていた．この影響を取り除き，1ヶ月内での特徴を抽出する方法を考える必要がある．

クラスタリング手法についても改善の余地がある．現在は k-means 法を繰り返し行い最適なクラスタ数を発見的に決めているので，x-means 法 [7] のような最適なクラスタ数を自動的に算出する方法を用いる必要がある．

また，今回用いなかったデータに対して本手法を適用することで，より多くの知見が得られると考えられる．例えば，今回はつくれぽ数の時間的変化からレシピの時間的特性を表す特徴ベクトルを作成したが，つくれぽ数の代わりにレシピページに対するアクセス数を用いるといった方法がある．ページにアクセスするという行為はつくれぽを投稿するのと比較してユーザにかかるコストが低いため，つくれぽほど明確にユーザがレシピの価値を認めたとは言えないが，データ数は格段に多く，特徴ベクトルが疎になりにくいメリットがある．

また，代表カテゴリによるクラスタの要約は，他の木構造を持つ分類にも適用でき，特にソーシャルメディアのようにユーザ自身がカテゴリ作成やカテゴリ付与を行う場合，分類の厳密性や妥当性，客観性はあまり期待できなくても，内容を直感的に伝えられる分類名が多く生み出されるような場合に効果的であると思われる．

謝　辞

本研究では，クックパッド株式会社と国立情報学研究所が提供する「クックパッドデータ」を利用した．

文　献

[1] 花井俊介, 灘本明代. 酷似レシピ抽出のためのクラスタリング手法の提案. 第 6 回データ工学と情報マネジメントに関するフォーラム 2014(DEIM2014), 2014.

[2] 杉山祐一, 山肩洋子, 田中克己. 手順情報としてのレシピデータに対する類似レシピの要約と微小で重要な差異の発見. 第 5 回データ工学と情報マネジメントに関するフォーラム 2013(DEIM2013), 2013.

[3] 風間一洋, 鳥海不二夫, 榊剛史, 篠田孝祐, 栗原聡, 野田五十樹. 東日本大震災時の Twitter データを用いた単語間の関係の時系列変化の分析. 人工知能学会全国大会論文集, 2012.

[4] 大久保雅且, 杉崎正之, 井上孝史, 田中一男. WWW 検索ログに基づくトレンド情報の抽出について. 情報処理学会研究報告デジタルドキュメント (DD), 1997.

[5] 松林達史, 幸島匡宏, 林亜紀, 澤田宏. 非負値テンソル因子分解を用いた購買行動におけるブランド選択分析. 人工知能学会論文誌, Vol. 30, No. 6, pp. 713–720, 2015.

[6] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 281–297, Berkeley, Calif., 1967. University of California Press.

[7] Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, Vol. 1, 2000.