投稿型レシピサイトを横断した重複レシピの判別

久保　遥† 関　洋平††

† 筑波大学 情報学群 知識情報・図書館学類　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
E-mail: †s1211493@u.tsukuba.ac.jp, ††yohei@slis.tsukuba.ac.jp

あらまし　本研究は，ユーザ投稿型レシピサイトを横断した重複レシピの判別について検証する．本研究では，重複レシピを，レシピとそれを模倣しているレシピ（群）のグループと定義する．重複レシピはユーザのレシピ検索に有用ではない．また，本研究の分析の結果，重複レシピは，レシピサイトを横断して投稿されやすい傾向があることがわかった．本研究では，重複レシピ間の調理内容の書かれ方の差異を判別に考慮することが，重複レシピの判別に有効かどうか，10 種類のレシピを対象とした実験を通して検証する．結果として，文末，数字，材料の置換が有効な場合があることを確認した．

キーワード　料理レシピ，N-gram，重複文書，投稿型レシピサイト

1. はじめに

1.1 本研究の目的と意義

近年，料理をする際にレシピサイトを利用する人は多い．「料理レシピに関する調査」[1] によると，料理レシピの入手先はレシピサイトが最も多くなっている．また，クックパッド (注1) による「料理に関するアンケート」[2] でも，料理をする際に最も参考にする情報源はレシピサイトとなっている．このように，レシピサイトは，料理をする人にとって大切な情報源である．

多くのレシピサイトの中でも，クックパッドや楽天レシピ (注2) など，一般ユーザがレシピを Web 上に掲載できる，投稿型レシピサイトが急速に発展している．2015 年 12 月 3 日時点で，クックパッドに掲載されているレシピ数は 223 万品，楽天レシピに掲載されているレシピ数は 105 万品である．クックパッドは代表的なレシピサイトであり，1998 年からサービスを開始している．また，楽天レシピは，2010 年からサービスを開始しているが，急速に掲載レシピ数を伸ばしている．

クックパッドのアンケート [2] によると，レシピサイトを利用する理由としては，「レシピが豊富」が最多であった．これはユーザによって求めるレシピが異なるため，より多くのレシピの中から自分の求めるレシピを見つけたいためと考えられる．また，杉山らの研究 [3] では，レシピ検索において，以下の傾向があることを明らかにした．

（1） レシピ検索では一般的な Web 検索に比べて，上位の検索結果が選ばれない

（2） ユーザは，複数のレシピのタイトル，写真，キャッチコピーではなく，主に本文を見比べて選択する

つまり，ユーザによって求めるレシピが異なるため，ランキング上位の少数のレシピを見るだけでは不十分であり，ユーザは調理手順などのレシピの本文を見ながら，複数のレシピを比較する．そしてユーザが，自分の要望に沿ったレシピを選択するために，より多くのレシピの中からレシピを検索をしようとした時，複数のレシピサイトを利用することが考えられる．しかし，レシピの中には調理手順が全く同一であったり，または調理内容の一部は変更されているが，同一のレシピとみなせるものがある．このような，**レシピとそれを模倣しているレシピ (群) のグループ**を，本研究では**重複レシピ**と呼ぶ．

重複レシピは，レシピ検索をしているユーザにとっては有用ではない．重複レシピのうち一つを残して他のレシピを取り除くことによって，ユーザのレシピ選択を支援できる．模倣しているレシピ (群) を取り除くためには，重複レシピを判別する必要がある．また，重複レシピは，レシピサイトを横断すると，多くなる傾向が見られる．そこで本研究では，レシピサイトを横断した重複レシピを判別する手法を実験を通して検証する．

1.2 重複レシピ

本節では，重複レシピとはどのようなレシピかを述べる．

前節で述べたように，ユーザによって求めるレシピは異なるため，ユーザはレシピの本文を見比べながら自分の求めるレシピを選択する．しかし，レシピの中には重複レシピが存在しており，ユーザーのレシピ検索の妨げとなる．本研究では，重複レシピの判断基準として，以下の 2 つを設ける．

（1） 料理の目的が同じもので，調理内容の一部が変更されているもの

（2） 料理の目的が異なるにも関わらず，目的に応じた調理内容に変化していないもの

料理の目的を比較する際には，レシピのタイトルやレシピの概略を使用する．例えば，クックパッドには「簡単！本格麻婆豆腐　子供の黒（辛くない）」(注3) と「簡単！本格麻婆豆腐　大

(注1)：http://cookpad.com/
(注2)：http://recipe.rakuten.co.jp/
(注3)：http://cookpad.com/recipe/1474912

人の赤（ピリ辛）」(注4) というタイトルのレシピが掲載されているが，この料理の目的は，前者は子供用の料理を作ることであり，後者は大人用の料理を作ることである．したがって，この 2 つのレシピの料理目的は異なっているといえる．なお，例外として，一つだけの手順から構成されるレシピに関しては，調理手順が完全に一致している場合のみ重複レシピとする．これは調理手順が短すぎるため，模倣しようがないためである．

2. 関 連 研 究

本研究の目的は，投稿型レシピサイトを横断した重複レシピの判別である．重複レシピはユーザの検索の妨げとなるため，判別することによってユーザがレシピを選択しやすくなる．杉山らの研究 [3] によると，ユーザにとってレシピを選択する上で，調理内容は重要である．よって，ユーザは，比較しているレシピのタイトルが異なっていても，調理内容が同一のレシピは重複レシピとみなす．同一のレシピを眺めるのは，ユーザにとって手間のかかる作業である．また，レシピの中には，写真が掲載されていないレシピや，異なるレシピに同じ写真が使用されているものもある．以上の知見から，本研究では，重複レシピかどうかを判定する際に，調理内容に着目する．

また，重複レシピの判別のためには，レシピ同士の模倣性を評価する必要がある．この評価には，小高らの研究 [4] を参考にし，文字 n-gram を利用する．さらに本研究では，重複レシピの調理内容の差異を考慮することで，重複レシピを発見しやすくなるか検証を行う．

ユーザにとって有用でないレシピを抽出する研究には，花井らの研究 [5] がある．花井らは，酷似レシピを抽出する手法を提案しているが，酷似レシピと本研究で扱う重複レシピは異なる．2014 年の花井ら [6] の研究では，酷似レシピを，「ユーザが検索を行い，提示された各レシピのタイトルとスニペットを見た状態で，違いを感じないようなレシピ」と定義している．本研究では，ユーザがレシピを選択する際に重視する調理内容に着目して，重複レシピの判別を行う．

3. 重複レシピの分析

本節では，重複レシピがどのような特徴を持つのか明らかにするための分析を行う．まず，重複レシピを抽出した後，重複レシピの調理内容の差異について分析する．

3.1 レシピのデータ

重複レシピを分析するためのレシピデータには，クックパッドのデータセット (注5) と楽天レシピのデータセット (注6) を用いる．2 つのデータセットを用いる理由は，レシピサイトを横断した重複レシピについて検証するためである．本節の分析では，Seki ら [7] の研究を参考にして，10 種類の料理（肉じゃが，親子丼，カルボナーラ，クリームシチュー，エビチリ，フレンチトースト，かぼちゃの煮物，麻婆豆腐，ポテトサラダ，スイートポテト）のデータを重複レシピの分析に使用する．レシピデータの抽出にあたっては，レシピのタイトルに該当料理名が含まれているものを，該当料理のレシピとする．実際に分析に使用した料理のレシピ数を，表 1 に示す．

表 1　分析に使用したレシピデータ

料理名	クックパッド	楽天レシピ
ポテトサラダ	7,330	1,858
肉じゃが	5,184	1,239
フレンチトースト	4,835	966
カルボナーラ	4,577	840
麻婆豆腐	3,223	675
スイートポテト	3,207	486
親子丼	2,172	563
かぼちゃの煮物	1,505	453
クリームシチュー	1,369	637
エビチリ	1,231	280

3.2 重複レシピの抽出

最初に，10 種類の料理に関して，重複レシピを取り出す．その際に，レシピサイトを横断した重複レシピのペア数 (注7) と同一レシピサイト内の重複レシピのペア数 (注8) を比較するために，1 つの料理に関して 3 通りのレシピサイトの組み合わせについて，以下の手順で重複レシピを判定する．

（1） 比較するレシピサイトの該当料理の，各レシピの調理手順を結合して，それぞれ 1 つの文字列にする．文字列に含まれる改行コード，空白，NULL に関しては削除する．

（2） すべての文字列のペアについて python-ngram(注9) を用いて，文字 n-gram で Jaccard 係数を用いて類似度を計算する (注10)．

（3） 類似度が高い順にランキングし，上位 200 件のペアが重複レシピかどうかを，人手で判定する．

重複レシピのペアは，レシピのペアを比較した際に，(1) 料理の目的が同じもので，レシピの調理内容の一部が変更されているもの，または (2) 料理の目的が異なるにも関わらず，目的に応じた調理内容に変化していないものとする．ただし，一つだけの調理手順で構成されているレシピ同士のペアに関しては，調理内容が完全一致しているレシピのみを重複レシピとする．

(注4)：http://cookpad.com/recipe/1474893
(注5)：http://www.nii.ac.jp/dsc/idr/cookpad/cookpad.html
(注6)：http://rit.rakuten.co.jp/opendataj.html

(注7)：レシピサイトを横断した重複レシピとは，類似度を計算するレシピのペアのうち，片方が楽天レシピに，もう片方がクックパッドに掲載されているものである．
(注8)：同一レシピサイト内の重複レシピとは，類似度を計算するレシピのペアが，両方ともクックパッドに掲載されている，もしくは，両方とも楽天レシピに掲載されているものである．
(注9)：https://github.com/gpoulter/python-ngram
(注10)：今回の調査では，日本語の特徴 [4] を考慮し，$n = 3$ とした．

3.3 抽出した重複レシピの分析

抽出した重複レシピのペア数を，表 2 に示す．重複レシピのペア数を比較すると，10 種類全てにおいて，レシピサイトを横断した重複レシピのペア数が，同一レシピサイト内の重複レシピのペア数を上回っている．この結果から，重複レシピは，レシピサイトを横断して投稿されやすい傾向があることが明らかとなった．

重複レシピを見つけるために，レシピのペアを類似度が高い順にランキングし，上位 200 件まで人手で判定した．その際の，重複レシピと判断したレシピのペアの，最下位の順位を表 3 に示す．最下位の順位を見ると，レシピサイトを横断した重複レシピペアと楽天レシピ同士の重複レシピペアに関しては，100 位以下には重複レシピは見られなかった．また，比較するレシピサイトに関わらず，130 位以下には重複レシピは見られなかった．以降の実験では，この分析結果に基づき，200 位以下には重複レシピのペアは無いものとみなす．

表 2　重複レシピのペア数 (件)

	比較するレシピサイト		
	クックパッドと楽天レシピ	クックパッド同士	楽天レシピ同士
ポテトサラダ	**60**	7	8
カルボナーラ	**50**	23	4
肉じゃが	**48**	26	6
フレンチトースト	**41**	8	5
スイートポテト	**32**	9	12
麻婆豆腐	**27**	13	5
親子丼	**25**	3	1
クリームシチュー	**24**	5	2
エビチリ	**19**	2	0
かぼちゃの煮物	**14**	6	2

表 3　重複レシピペアの最下位の順位（位）

	比較するレシピサイト		
	クックパッドと楽天レシピ	クックパッド同士	楽天レシピ同士
ポテトサラダ	79	116	72
カルボナーラ	69	102	75
肉じゃが	69	102	18
フレンチトースト	87	128	64
スイートポテト	34	34	19
麻婆豆腐	27	94	84
親子丼	27	4	4
クリームシチュー	26	7	3
エビチリ	27	8	–
かぼちゃの煮物	14	16	16

3.4 重複レシピにおける調理内容の差異の分析

表 2 から，重複レシピは，レシピサイトを横断して投稿されやすい傾向があることが明らかになった．本節では，レシピサイトを横断した重複レシピペアの調理内容を人手で分析し，重複レシピの判別に利用するために，調理内容の差異を明らかにする．

開発データの作成

重複レシピの差異を分析するための開発データは，3.3 節で抽出した，重複レシピペアを利用して作成する．すなわち，重複レシピ内で出現した回数が一回 (注11) 同士のレシピペアのうち，ランキング最上位から奇数順位のレシピペアを開発データとする．参考までに，表 2 に示した，重複レシピのペアに含まれるレシピが，重複レシピとして出現した回数を，表 4 に示す．

表 4　レシピサイトを横断した重複レシピペアにおける同一レシピの出現回数の分布

	出現回数				
	1 回	2 回	3 回	4 回	5 回
ポテトサラダ	115	1	1	–	–
カルボナーラ	98	1	–	–	–
肉じゃが	82	4	2	–	–
フレンチトースト	80	1	–	–	–
スイートポテト	51	4	–	–	1
麻婆豆腐	49	1	1	–	–
親子丼	42	4	–	–	–
クリームシチュー	43	1	1	–	–
エビチリ	34	2	–	–	–
かぼちゃの煮物	28	–	–	–	–

分析結果

分析の結果，調理内容の書かれ方には，大きく分けて以下の 3 つの差異が見られた．

- 調理内容の一部の代替
- 調理内容の一部の追加
- 文の分割

書かれ方の差異の例を，表 5(注12) に示す．2 つ並んだ調理内容のうち，上はクックパッドに掲載されていたレシピであり，下は楽天レシピに掲載されていたレシピである．また 2 つの調理内容の差分を太字にしている．以下，表 5 を利用して，上記の項目について説明する．例えば，<1>というのは，表 5 の<1>の番号を示す

調理内容の一部の代替

調理内容の一部の代替とは，レシピを比較した際に，調理内容の一部が言い換えられていることである．調理内容の一部の代替には，(1) 記号，(2) 文末，(3) 数字，(4) 切り方，(5) 材料の 5 種類がある．以下，この 5 種類について説明する．

(1)　記号とは，使用されている記号が異なることである．例えば，<1>では,「●」と「★」といった，異なる記号が使用さ

(注11)：重複レシピと判断したペアの中には，複数のレシピに対して重複レシピと判断されているレシピがある．

(注12)：<11> の ID は変更している

れている．

(2) 文末とは，文の末尾の書かれ方が異なることである．例えば，<2>の文末は,「のせる。」と「のせておく。」と，異なる書き方をしている．

(3) 数字とは，調理時間や手順の番号などの，説明に使用する数字が異なることである．例えば，<3>，<4>，<5>で，数字の代替が行われている．<3>では，調理時間を指す数字の「１５分」と「１０分」とが異なり，<4>では，調理手順を指す数字の「7」と「6」とが異なり，<5>では,「①」と「【1】」のように，数字の表記が異なる書き方をしている．

(4) 切り方とは，野菜や肉の切り方が異なることである．例えば，<6>では,「小口切り」と「薄切り」と，異なる切り方を示している．

(5) 材料とは，材料が変更されたり，異なる書き方をしているものである．例えば，<7>では,「じゃが芋」と「新じゃが」と，異なる．

調理内容の一部の追加

調理内容の一部の追加とは，レシピを比較した際に，片方のレシピに調理内容の一部が追加されていることである．追加される調理内容の一部には，(1) 句読点や♪などの記号，(2) 料理とは関係のない文，(3) 料理と関係のある文の 3 種類がある．以下，この 3 種類について説明する．

(1) 記号とは，句読点や♪などの記号が，片方のレシピに追加されていることである．例えば，<8>や<9>が記号の追加にあたる．<8>には，上の文に「♪」の記号が追加されており，<9>は，上のレシピに「。」が追加されている．

(2) 料理とは関係のない文とは，ユーザへのコメントや他のレシピの紹介など，直接的に料理とは関係のない文が追加されていることである．例えば，ユーザへのコメントとは<10>を，他のレシピへのリンクとは<11>のようなものを指す．

(3) 料理と関係のある文とは，材料の分量や料理をする際のコツが追加されていることである．例えば，<12>では，下のレシピに，油をひかずに豚ひき肉を炒めるために,「テフロンのフライパン」を使用すると詳しく書いている．

文の分割

文の分割とは，一文が二文に分割して書かれていたり，逆に二文が一文に繋がって書かれていることである．例えば，<13>のように,「豚ひき肉を色が変わるまで炒めて〇の材料を加えて炒め合わせて☆を加える。１に青菜と豆腐を加えて一煮立ちする。」という文が一文に纏められ,「豚ひき肉を色が変わるまで炒めて〇の材料を加えて炒め合わせ、☆を加えて青菜と豆腐を加えて一煮立ちする。」と書かれているものを指す．

本研究では，重複レシピにおける調理内容の差異のうち，調理内容の一部の代替について検証する．これはユーザがレシピを模倣する際に，調理内容の言葉の代替を他の 2 つと比べて頻繁に行うと考えたためである．よって，調理内容の一部の代替を考慮することにより，重複レシピの判別の精度が向上するか検証する．

表 5 調理内容の書かれ方の差異の例

<1>	片栗粉をまぶしたエビを両面焼き、端っこで ● も炒める。
	片栗粉をまぶしたエビを両面焼き、端っこで★も炒める。
<2>	水にさらした玉ねぎを手でぎゅっと絞り、熱いじゃが芋の上にのせる。
	水にさらした玉ねぎを手でぎゅっと絞り、熱いじゃが芋の上にのせておく。
<3>	鍋に水、かぼちゃ、調味料を入れて１５分くらい煮る。
	鍋に水、かぼちゃ、調味料を入れて１０分くらい煮る。
<4>	7に５のホワイトソースを加え、更に煮込みます。
	6に５のホワイトソースを加え、更に煮込みます。
<5>	お鍋にたっぷりお水を入れ ①、②（ジャガイモは後で…）を入れ煮込む。
	お鍋にたっぷりお水を入れ【1】、【2】（ジャガイモは後で…）を入れ煮込む。
<6>	きゅうりを小口切りにし、塩少々（分量外）をふり 5 分ほどおき、水気を絞る。
	きゅうりを薄切りにし、塩少々（分量外）をふり 5 分ほどおき、水気を絞る。
<7>	じゃが芋は小口切りにし水に１０分さらす。
	新じゃがは小口切りにし水に１０分さらす。
<8>	冷蔵庫に入れ冷ましたらできあがり☆（温かいほうがお好みの型はそのままでも ♪）
	冷蔵庫に入れ冷ましたらできあがり☆（温かいほうがお好みの型はそのままでも）
<9>	焼きあがったら粉糖を茶こしに入れて，降りかける。出来上がり 。
	焼きあがったら粉糖を茶こしに入れて，降りかける。出来上がり
<10>	**2008/1/31、ピックアップレシピに載せていただきました！ありがとうございました**
	(記載なし)
<11>	**「隠し味は練乳☆しっとりポテトサラダ」 ID：xxxxxxx**
	(記載なし)
<12>	油はひかずに豚ひき肉に生姜・豆板醤とテンメンジャンを加えてよく炒める
	油はひかずに（テフロンのフライパン）豚ひき肉に生姜・豆板醤とテンメンジャンを加えてよく炒める。
<13>	豚ひき肉を色が変わるまで炒めて〇の材料を加えて炒め合わせて☆を加える。１に青菜と豆腐を加えて一煮立ちする。
	豚ひき肉を色が変わるまで炒めて〇の材料を加えて炒め合わせ、☆を加えて青菜と豆腐を加えて一煮立ちする。

4. 重複レシピ判別手法の概要

本節では，3 節で述べた，重複レシピの調理内容の一部の代替を考慮した，重複レシピの判別手法を提案する．4.1 節で提案手法の概要を述べた後で，4.2 節で調理内容の一部の代替を考慮するために使用する，調理内容の要素の置換辞書について

説明する．

4.1　提案手法の概要

3 節で述べたように，重複レシピはレシピサイトを横断して投稿されやすい傾向がある．本研究では，重複レシピはユーザのレシピ検索において有用ではないため，レシピサイトを横断した重複レシピの判別手法を提案する．図 1 に提案手法を示す．

本研究での重複レシピ判別手法は，2 つある．

1 つ目は，図 1 の左に示した，レシピペアの類似度に基づき重複レシピの判別に使用する手法である．

2 つ目は，図 1 の右に示した，レシピペアの類似度を計算する前に，調理内容の一部の代替を考慮する手法である．重複レシピには，3.4 節で述べたように，(1) 記号，(2) 文末，(3) 数字，(4) 切り方，(5) 材料といった，調理内容の一部の代替が現れる．この観察に基づき，代替された要素を置き換える辞書を人手で作成する．代替される要素を置き換えることによって，調理内容の一部の代替を考慮できると考える．使用する置換辞書については，4.2 節で述べる．

5 節では，この 2 つの手法を比較し，調理内容の一部の代替を考慮することにより重複レシピの判別精度が向上するか検証する．

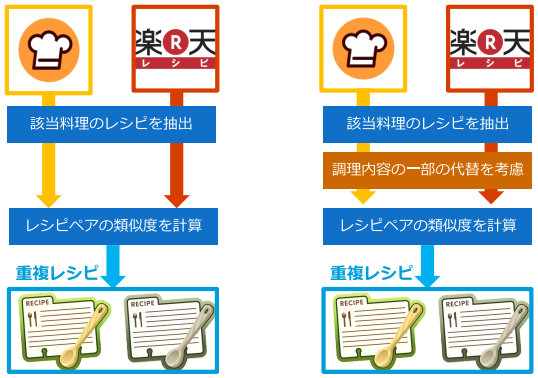

図 1　重複レシピの判別手法

4.2　調理内容の要素の置換辞書の概要

3.4 節において重複レシピにおける調理内容の分析を行った結果，重複レシピの調理内容には，大きく分けて 3 つの差異が現れることがわかった．本研究では，そのうちの一つである調理内容の一部の代替を，重複レシピの判別に利用する．

調理内容の一部の代替には，(1) 記号，(2) 文末，(3) 数字，(4) 切り方，(5) 材料の 5 種類の代替がある．そこで，3.4 節で述べた開発データを参考にして，5 種類の調理内容の一部の代替を考慮する (1) 記号置換辞書，(2) 文末置換辞書，(3) 数字置換辞書，(4) 切り方置換辞書，(5) 材料置換辞書を，人手で作成する．また，それぞれの置換辞書を利用して，代替される要素を同じ要素に置換する．

例えば，表 5 の<7>のように，レシピにおいて「新じゃが」と「じゃが芋」は代替されうる材料と考える．そこで，材料置換辞書に，「新じゃが，#f」と「じゃが芋，#f」というデータを追加する．これは，レシピの調理内容に含まれる「新じゃが」と「じゃが芋」という材料を，「#f」に置換することを表している．置換辞書を使用することによって，レシピにおいて同一視されうる異なる要素を同じ要素とみなすことが出来る．

今回作成した記号置換辞書は，336 種類の記号を置換する．これは，UTF-8 の文字コード表 (注13) を参考に，重複レシピを分析した際に，実際に使用されていた記号の所属する区分の全ての記号を，すべて同じ要素に置換している．

文末置換辞書は，34 種類の文末を置換する．これは，重複レシピを分析した際に，実際に置換されていた文末同士をグループでまとめ，そのグループ内で同じ要素に置換している．

数字置換辞書は，75 種類の数字を置換する．これは，<1>や (1) などの表記ゆれや，時間や温度などの値を，すべて同じ要素に置換している．この際に，連続する要素は，1 つの要素に置換している．

切り方置換辞書は，8 種類の切り方を置換する．これは，食材の切り方 (注14) を参考に，同じ材料で使用する切り方をグループにし，そのグループ内で同じ要素に置換している．

材料置換辞書は，100 種類の材料を置換する．重複レシピを分析した際に，材料が置換されても，料理の目的が変わらない材料やひらがなやカタカナなどの異なる表記をグループでまとめ，そのグループ内で同じ要素に置換する．

5.　重複レシピの判別手法の評価

5.1　目　　的

4 節までに述べたように，調理内容の一部の代替を考慮することによって，重複レシピの判別精度が向上するか検証する．

5.2　方　　法

実験には，3.1 節で述べた，クックパッドと楽天レシピから抽出した，10 種類の料理のレシピデータを用いる．また，重複レシピを判別する際に，レシピペアを類似度でランキングするが，3.4 節で使用した開発データのレシピペアは取り除く．

調理内容の一部の代替を考慮するためには，4.2 節で述べた 5 種類の置換辞書（記号置換辞書，文末置換辞書，数字置換辞書，切り方置換辞書，材料置換辞書）を利用する．実験では，各 10 種類の料理に対して，以下の 6 つの重複レシピ判別手法をした結果の，上位 100 件に含まれる重複レシピ数を比較する．

（1）　調理内容の一部の代替を考慮せずに，3.2 節で述べた重複レシピの判別をした場合

（2）　記号置換辞書を利用した場合

（3）　文末置換辞書を利用した場合

（4）　数字置換辞書を利用した場合

（5）　材料置換辞書を利用した場合

（6）　切り方置換辞書を利用した場合

本実験においては，(1) の結果をベースラインとする．

(注13)：http://www.shurey.com/js/works/unicode.html
(注14)：http://www.kurakon.jp/cooking/kihon/cut.html

5.3 結　　果

重複レシピペア数の結果を，表 6 に示す．ベースラインとは，調理内容の一部の代替を考慮せずに重複レシピの判別を行ったものである．記号，文末，数字，材料，切り方とは，5 種類ある置換辞書のうち，それぞれ一種類のみを利用して重複レシピの判別を行ったものである．ベースラインと比較して，重複レシピペア数に変化があったものは太字にしている．

置換辞書を利用することで重複レシピペア数に変化があった料理は，肉じゃが，フレンチトースト，親子丼，かぼちゃの煮物，エビチリの 5 種類であった．最も重複レシピペア数が増えたのは，材料置換辞書を利用した肉じゃがとフレンチトーストのレシピであった．

表 6　判定した重複レシピペア数の比較

	ベースライン	記号	文末	数字	材料	切り方
ポテトサラダ	32	32	32	32	32	32
肉じゃが	28	28	28	**29**	**31**	28
フレンチトースト	21	21	21	**22**	**24**	21
カルボナーラ	26	26	26	26	26	26
麻婆豆腐	16	16	16	16	16	16
スイートポテト	20	20	20	20	20	20
親子丼	14	14	**15**	14	14	14
かぼちゃの煮物	7	7	**8**	7	**9**	7
クリームシチュー	14	14	14	14	14	14
エビチリ	11	11	**12**	**12**	11	11

5.4 考　　察

文末の代替，数字の代替，材料の代替を考慮することによって，重複レシピペア数にはわずかに増加が見られた．一方，記号の代替，切り方の代替を考慮しても，10 種類の全ての料理に対して変化は見られなかった．

文末，数字，材料の代替でレシピ数の増加が見られたのは，表現がレシピユーザに依存せず，文字列を置き換えることによって，レシピペアの類似度が向上したと考える．

ベースラインと比較して，最もレシピ数が増えたのは，肉じゃがとフレンチトーストの料理に材料の代替を考慮した判別手法で，重複レシピペアは 3 ペア増加した．重複レシピペアと判別できたレシピは,「人参」と「ニンジン」など，異なる表記を同じ要素に置換することで，レシピペアの類似度が向上しているものが多かった．

しかし，重複レシピではないレシピペアのうち，調理手順が 3 手順以内の短いレシピの類似度も，ベースラインと比較して上がっている．つまり，材料置換辞書で置換できる材料を増やすだけでなく，調理手順の長さも考慮にいれることができれば，さらに重複レシピの判別精度が向上すると考える．

また，新しく見つかった重複レシピペアには，3.4 節で述べたような，料理とは関係ない文，例えばユーザへのコメントが片方に多く追加されていることによって，レシピペアの類似度が低くなっているレシピペアがあった．よって，料理とは関係のない文を削除することで，重複レシピの判別精度が向上すると考える．

一方，記号の代替，切り方の代替を考慮しても，10 種類の全ての料理に対して変化は見られなかった．これは，切り方が,「3〜4cm 程度に切る」や「三等分に切る」など，ユーザによって異なるかたちで行われるためと考える．また，記号の代替は，一文字の置き換えだったため，記号の代替を考慮しても，類似度が向上しなかったと考える．

ポテトサラダ，カルボナーラ，麻婆豆腐，スイートポテト，クリームシチューに関しては，調理内容の一部の代替を考慮しても，ベースラインと比較して重複レシピペア数は増えなかった．ベースラインのランキングの上位 100 件に含まれる重複レシピではないレシピペアを見直した結果，調理内容の類似度は高いものの，重複レシピではないレシピペアが，多数含まれていた．よって，重複レシピペアではないレシピペアが重複レシピの判別の妨げとなっていると考える．以上の観点から，これらの料理については，重複レシピではないが，調理内容の類似度が高いレシピペアをランキングから削除することで，重複レシピの判別精度が向上すると考えられる．

6. おわりに

本研究では，レシピとそれを模倣しているレシピ (群) のグループを，重複レシピと呼ぶ．重複レシピは，レシピ検索をするユーザにとっては有用ではない．本研究での分析の結果，重複レシピはレシピサイトを横断して投稿されやすい傾向があることが明らかになった．この見解に基づき，投稿型レシピサイトを横断した重複レシピの判別手法を提案し，実装を行った．

結果として，文末，数字，材料の置換が有効な場合があることを確認した．

今後は，重複レシピの判別に考慮すべきこととして，類似度が高い，重複レシピではないレシピペアをランキングから削除することや，調理内容から，料理とは関係のない文を削除することを検討している．

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」ならびに，楽天株式会社と高度言語情報融合フォーラム (ALAGIN) が提供する「楽天データ」を利用した．ここに深く感謝する．

本研究の一部は，筑波大学研究基盤支援プログラム（B タイプ），科学研究費補助金基盤研究 B（課題番号 25280110），萌芽研究（課題番号 25540159）の助成を受けて遂行された．

文　　献

[1] マルハニチロホールディングス．料理レシピに関する調査．https://www.maruha-nichiro.co.jp/news_center/research/pdf/20130227_recipe_cyousa.pdf (accessed 2015-12-8).

[2] クックパッド．料理に関するアンケート．https://cf.cpcdn.com/info/assets/wp-content/uploads/20140306000000/pr130723-survey.pdf (accessed 2012-12-8).

[3] 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 (DEIM 2013), D3-5, 2013.

[4] 小高知宏, 村田哲也, 高建斌, 諏訪いずみ, 白井治彦, 高橋勇, 黒岩丈介, 小倉久和. n-gram を用いた学生レポート評価手法の提案. 電子情報通信学会論文誌. D, Vol. 86, No. 9, pp. 702–705, 2003.

[5] 花井俊介, 灘本明代, 難波英嗣. スパムレシピ抽出のための酷似レシピクラスタリング手法. 情報処理学会研究報告, Vol. 2014-OS-131, No. 26, 2014.

[6] 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 (DEIM 2014), F8-6, 2014.

[7] Seki Yohei and Kouta Ono. Discriminating Practical Recipes Based on Content Characteristics in Popular Social Recipes. In *Proceedings of the 2014 Association for Computing Machinery International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp 2014)*, pp. 487–496, Seattle, WA, USA, 2014.