ARG WI2 No.11, 2017
WI2-2017-25

重複する料理レシピを判別するためのコーパスの構築

島田 理紗子[†,a]　小邦 将輝[††,b]　平手 勇宇[†††,c]　関 洋平[††††,d]

† 筑波大学大学院 図書館情報メディア研究科　†† 筑波大学 情報学群 知識情報・図書館学類
††† 楽天株式会社 楽天技術研究所　†††† 筑波大学 図書館情報メディア系

a) *s1621617@u.tsukuba.ac.jp*　b) *s1613123@u.tsukuba.ac.jp*　c) *yu.hirate@rakuten.com*　d) *yohei@slis.tsukuba.ac.jp*

概要　本研究では，重複する料理レシピを判別するために構築したコーパスについて紹介する．重複レシピとは，レシピを構成する要素の一部またはすべてが他のレシピと一致しているレシピを指す．重複レシピには，重複の度合いや，料理の同一性などの組み合わせによって複数のパターンがあるため，本研究では料理の共通性や調理手順の共通性の重複の度合いによってレシピを 7 段階に分類し，コーパスを構築した．さらに，構築したコーパスのレシピのうち，一部が重複しているレシピの変更部分に対して，レシピ用語の種類をタグ付けした．その結果，重複の度合いによって，変更されるレシピ用語の種類の傾向が異なることがわかった．

キーワード　コーパス，重複レシピ，レシピ用語

1　はじめに

投稿型料理レシピサイトには，100 万件を超えるレシピが登録されているものがあり，探したいレシピを見つけるのに充分なレシピ数が登録されている．その一方で，登録されているレシピの中には似た内容のレシピが存在しており，サービスが低下する可能性のある，レシピの内容が重複しているレシピが含まれる．本研究では，重複しているレシピを，レシピの内容の一部，または全てが他のレシピの一部またはすべてと一致しているレシピと定義した．これを重複レシピと呼ぶ．本研究では，重複レシピを判別するために，コーパスの構築を行い，構築したコーパスの有効性について検証する．料理レシピは，レシピタイトル，材料，レシピの写真，調理手順などの要素から構成されている．このうち，調理手順はレシピの書き手ごとに多くの書き方が存在し，同一の料理でも全く異なるレシピとなる．しかし，中には料理も調理手順も類似している重複レシピが存在している．久保ら [1] は，調理目的に着目し，それに沿った調理内容の書き換えが行われているかどうかで重複レシピを分類した．また Oguni et al.[2] では，調理手順と材料が同一なレシピをブラックレシピ，調理手順と材料が類似していてオリジナリティがないレシピ，または，調理手順が一致しているが材料が異なるレシピをグレーレシピとし，どちらも重複レシピとみなし分類した．

本研究では，重複のパターンごとに，重複の度合いを多段階で区別した重複レシピを定義する．その際，料理の種類が共通しているかどうかと，調理手順の重複の度合いを考慮し，重複レシピを 7 段階に定義した．この定義を基に，重複レシピのコーパスを作成しコーパスの有効性について検証する．

本稿の構成は次の通りである．2 章では，料理レシピに関する研究や重複レシピに関する研究について述べ，本研究の位置づけについて述べる．3 章では，本研究における重複レシピの定義についてまとめ，7 段階の分類について説明する．4 章では，定義した重複レシピのコーパスを構築し，コーパスの有効性について検証する．最後に 5 章で本研究についてまとめる．

2　関連研究

2.1　料理レシピの研究

料理レシピのテキストに着目した研究では，阿部ら [3] のレシピの手順からテキストの特徴を抽出し調理時間を予測する研究がある．この研究では，調理手順において材料と動作のつながりが重要であると考えられている．本研究では，この研究を参考に，重複レシピにおいて，調理手順における材料（とそれに伴う動作）の入れ替えが，調理目的の変更を反映していると考えられるものは，重複の度合いが低いと判定する．また，笹田ら [4] は，レシピ用語に食材や道具などのタグ付けを行い，コーパスを用いてレシピ用語のタグの自動認識を実現した．重複レシピを判別する際には，このようなレシピ独自のタグが手がかりとなる可能性がある．本研究では，重複レシピにおいて，元のレシピから変更した用語がどのタグに対応しているかによって，重複の度合いが異なるという仮説を立てる．この仮説の有効性については，4 章で考察する．

2.2　類似レシピの研究

花井ら [5] は，「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」を酷似レシピと定義し，材料名をクエリとして抽出したレシピを対象にクラスタリングを用

いて酷似レシピの抽出を行い人手で判定を行った．この研究では，調理動作を手がかりとしてクラスタリングを行っている．また，苅米ら [6] の研究では，調理手順の類似について，使用する材料が異なっていても，調理動作が一致していれば，調理手順がほとんど同じ料理が存在すると述べている．本研究では，これらの研究を参考に，料理が異なっていた場合でも，調理手順が一致していれば重複と判定する．

2.3 重複レシピの研究

久保ら [1] は，10 種類の料理を対象として，料理レシピサイトを横断した場合に，重複レシピがより多く存在することを明らかにした．この研究では n-gram 類似度を用いて類似しているレシピペアの抽出を行い，人手で 3 段階のアノテーションを行っている．また Oguni et al.[2] は，重複レシピを 3 段階に分け，調理手順の類似度に加え画像の類似度を組み合わせて検出した．本研究では，重複レシピを投稿者ごとに抽出したレシピを用いて，7 段階のアノテーションを行うことで，多様な重複のパターンを区別して重複レシピを判別することを目標とする．

2.4 剽窃についての研究

剽窃されたテキストを判別する研究は多く行われている．上野ら [7] は，レポートに含まれるテキストが Web のどのページから剽窃されたかを調べるシステムの構築を行い，Stamatatos[8] は，剽窃文章コーパスを用いて，剽窃文章を抽出する研究を行っている．本研究では，これらの研究の手法を参考にしつつも，料理レシピの特徴を反映したテキストの重複を判定する方法について考察する．

3 重複レシピの定義

本研究では，重複レシピの定義を料理，材料の共通性や調理手順の重複のパターンによって，完全重複，部分重複 A，部分重複 B，部分重複 C，部分重複 D，非重複 A，非重複 B の 7 段階に定めた．

- 完全重複
  - 料理に共通性がある
  - 材料に共通性がある
  - 調理手順が一致している
- 部分重複 A
  - 料理に共通性がある
  - 材料の追加や変更が，材料リストだけで行われている（調理手順における表記では，変更はない）
  - 調理手順では，記号や “すべて” といった全体や部分を表す語彙による表記の追加や変更はあるが，それ以外は一致している
- 部分重複 B
  - 料理に共通性がある
  - 材料の追加や変更が，材料リストと調理手順で行われている
  - 調理手順の変更箇所は，材料の追加や変更に伴うものであり，材料名のみ，または調理目的を考慮した工夫が見られない調理動作が装飾された材料名（例:茹でた○○，ざく切りにした○○，水洗いした○○，ちぎった○○）に変更されており，オリジナリティがなく，調理手順全体に共通性がある
- 部分重複 C
  - 料理に共通性がある
  - 材料の追加や変更が，材料リストと調理手順で行われている
  - 調理手順の変更箇所に工夫がみられ，追加や変更が行われた材料について，温度や調理時間など具体的な要素や，調理目的を考慮した工夫が見られる調理動作が加わっているもの（例 : 180 度に温めたオーブンで 3 分焼く，強火できつね色になるまで 2 分ほど加熱する）．調理手順全体では共通性がある箇所もあるが，オリジナリティもみられるもの
- 部分重複 D
  - 異なるカテゴリの料理または共通性のない料理
  - 調理手順が一致している，または共通性がある
- 非重複 A
  - 料理に共通性がある
  - 調理手順に共通性がなく，比較先レシピに手順や説明が追加されているなど調理手順が詳細化されており，オリジナリティがある
- 非重複 B
  - いずれのラベルも付与されないレシピ

また，すべてのラベルに共通して，以下の補足説明を設けた．

- 料理の共通性とは，楽天レシピの登録カテゴリ，またはレシピタイトルに同じ料理名が含まれるものを指す
- 漢字，ひらがな，送り仮名の変換違い，句読点や記号，改行数の違い（同一手順内における間隔をあけるための改行など）があった場合でも，重複とする
- 材料リストの中での，材料の順番の入れ替えは，材料の変更としない

- レシピのオリジナリティは，材料が追加･変更されたことによって意味のある手順が追加される，またはレシピに沿った書き方に変更されているとその箇所はオリジナリティがあると判断する
- レシピを比較した際，投稿日が新しいレシピの材料が減少しており，減少した材料に関する調理手順のみ削除された場合は，調理手順の変化としない
- その料理を作成する上で必ず行う基本的な調理動作が一致しているが，表層的な表現が細部まで一致していない場合，調理手順に共通性は認めなくて良い

4 重複レシピを判別するためのコーパスの構築と検証

4.1 重複レシピコーパスに用いるデータ

楽天レシピで公開されているデータを用いて，20 代学生のアノテーター 4 名（筆頭著者 1 名を含む，男性 2 名，女性 2 名）を雇用し定義に基づきレシピのアノテーションを行い，被験者間一致度を計算した．

実験のデータは，投稿数の多いユーザのうち，短時間に 20 件以上投稿を行ったユーザ上位 10 ユーザのレシピ各 10 件を用いた．短時間に投稿を多く行ったユーザのレシピは，時間短縮のため，重複した内容のレシピを投稿している可能性が高いと推測される．本研究では，このレシピ群を重複レシピ候補群と呼ぶ．重複している可能性のあるレシピと比較するためのレシピは，重複レシピ候補群のレシピより投稿日が古く，n-gram 類似度（$n = 3$）が高いレシピをオリジナルレシピ候補として抽出した．

重複レシピ候補とオリジナルレシピ候補のレシピを比較し，レシピのタイトル，カテゴリ名，材料リスト，調理手順をレシピサイト上で確認しながら，該当するラベルを各レシピに付与した．その際，はじめにレシピで作成する料理が共通しているか，異なる料理かを判定し，図 1 のように重複度が高い完全重複から判定を行い，異なる場合次のラベルを判定を行い，最終的にどのラベルにも当てはまらない場合は非重複 B を付与した．

アノテーター 4 名のそれぞれの被験者間一致度を Cohen's kappa[9]，アノテーター 4 名全体の一致度を Fleiss's kappa[10] を用いて計算した．その結果を，表 1 に示す．

すべてのアノテーターの組み合わせにおいて，0.65（Substantial Agreement[11]）以上であり，アノテーターの組み合わせによっては高い一致度（Almost Perfect Agreement[11]）も含まれる結果となった．また，アノテーター 4 名全体の Fleiss's kappa は，0.737（Substantial Agreement）であった．

この結果より，本研究では，アノテーションを行った

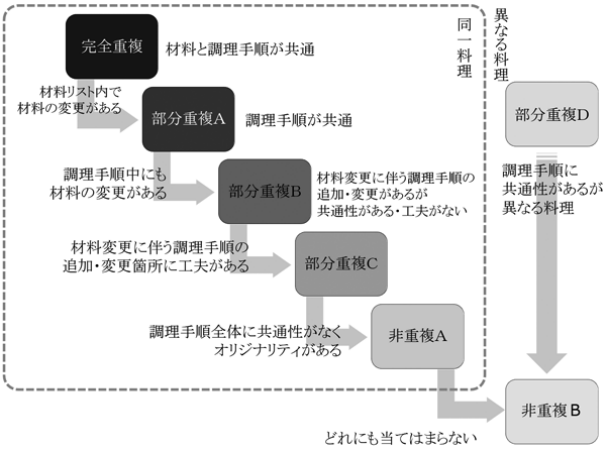

図 1 重複度合いを考慮したレシピのラベル判別

表 1 アノテーションの被験者間一致度 (Cohen's kappa)

アノテーター番号	Ⅰ	Ⅱ	Ⅲ	Ⅳ
Ⅰ	-	0.852	0.706	0.721
Ⅱ	0.852	-	0.675	0.721
Ⅲ	0.706	0.675	-	0.801
Ⅳ	0.721	0.721	0.801	-

レシピを使用し，コーパスの構築を行い，検証を行った．コーパスの構成は表 2 に示す．また，コーパス検証に用いるレシピは，3 名以上のアノテーターのラベルが一致した 88 レシピのうち，部分重複のいずれかを付与した 75 レシピを使用した．各ラベルを付与したレシピ数は，部分重複 A が 8 レシピ，部分重複 B が 51 レシピ，部分重複 C が 9 レシピ，部分重複 D が 7 レシピである．

表 2 重複レシピコーパスの構成内容

ラベル名	内容
ラベル	完全重複，部分重複 A，部分重複 B，部分重複 C，部分重複 D，非重複 A，非重複 B
レシピ ID	楽天レシピのレシピ ID
レシピタイトル	レシピのタイトル
カテゴリ	そのレシピが登録されているレシピのカテゴリ
材料リスト	レシピの材料
調理手順	1 行にまとめた調理手順

4.2 重複レシピコーパスの有効性の検証

4.1 で構築したコーパスから，ラベル別に特徴量の抽出を行い，コーパスの有効性について検証を行った．特徴量の抽出は，アノテーションを行った際に用いたレシピペアを使用し，調理手順の変更箇所を対象に行った．

笹田ら [4] が作成したタグ付与コーパスを参考に，変更箇所の語彙にタグを付与し，タグごとに変更のあったレシピ数を集計した．タグの付与は，アノテーター 2 名（筆頭筆者を含む，男性 1 名，女性 1 名）で行い，被験者間一致度を計算した．

4.3 結果

集計した結果を表 3 に示す．被験者間一致度は，0.72（Cohen's kappa，Substantial Agreement）であった．

表 3 部分重複レシピの変更箇所のタグ分布

タグ名	部分重複							
	A		B		C		D	
食材	2	25.0%	50	98.0%	9	100.0%	6	85.7%
道具	1	12.5%	9	17.6%	8	88.9%	2	28.6%
継続時間	1	12.5%	9	17.6%	6	66.7%	1	14.3%
分量	0	0.0%	0	0.0%	1	11.1%	1	14.3%
調理者の動作	0	0.0%	36	70.6%	8	88.9%	6	85.7%
食材の動作	0	0.0%	13	25.5%	5	55.6%	5	71.4%
食材の様態	0	0.0%	21	41.2%	8	88.9%	5	71.4%
道具の様態	0	0.0%	1	2.0%	3	33.3%	1	14.3%
レシピの総数	8	–	51	–	9	–	7	–

部分重複 A のレシピは，食材を指す記号の種類が変化したレシピや，コップがグラスに変化したレシピがあったが，それ以外のレシピに変更点はなかった．部分重複 B のレシピは，食材とそれに合わせた食材を処理する，調理者の動作が変化したレシピが約 70%あった．部分重複 C のレシピは，すべてのレシピで食材の入れ替えが行われており，調理者の動作，食材の様態も約 90%のレシピで変更が行われていた．また，部分重複 B と異なり，道具や継続時間にも変化があった．部分重複 D のレシピでも，食材と調理者の動作の変更がほとんどのレシピで行われていたが，他の部分重複レシピと比べて，食材の動作の変更が多い結果となった．

4.4 考察

部分重複 A は，材料変更に伴う調理手順に変更がないレシピのため，調理手順の変更がほぼ行われなかった．部分重複 B は，「“材料名” を切って」のように，材料名と調理者の動作の二つを変更するレシピが多かった．これは，材料の変更に伴う調理手順の変更を行う場合に，調理動作のみ変更することで，手軽にレシピを作成しているためと推察される．それに比べ，部分重複 C では，道具や継続時間，切り方などを，変更した材料に合わせて詳細に書くことで，調理手順に工夫が見えた．また，部分重複 D は，部分重複 A，B，C と異なり，食材の動作が変更されたレシピが多かった．これは部分重複 D は異なる料理を対象にしており，「“材料名” が焼けたら」といった調理の流れの一部が変更されているためであると考えられる．これらの結果より，このコーパスを用いることで，部分重複が重複の度合いによって区別できることがわかった．

5 おわりに

本研究では，重複した料理レシピを判別するためのコーパスの構築を行い，コーパスの有効性について検証した．その結果，ラベルごとにレシピの変更箇所が異なり，部分重複レシピにおいて 4 つのパターンに分類できることが分かった．今後の課題として，コーパス作成に用いたレシピ数の分布がラベルごとに偏りがあるため，さらにアノテーションを行い，レシピ数を増やすことでより精度の高いコーパスの構築を目指す．

謝辞

本研究は，楽天株式会社提供の「楽天データセット」を用いて分析を行った．また，本研究の一部は，科学研究費補助金基盤研究 B（課題番号 16H02913）の助成を受けて遂行された．

参考文献

[1] 久保遥，関洋平：投稿型レシピサイトを横断した重複レシピの判別，第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM2016），C8-3, 2016.

[2] Oguni, M., Seki, Y., Shimada, R., and Hirate, Y.: Method for detecting near-duplicate recipe creators based on cooking instructions and food images, Proc. of 9th Workshop on Multimedia for Cooking and Eating Activities (CEA 2017), Melbourne, Austraria, pp. 49-54, 2017.

[3] 阿部卓也，立間淳司，青野雅樹：料理レシピサイトから抽出される特徴に基づいた調理時間予測，情報科学技術フォーラム講演論文集，14(2), pp. 103-104, 2015.

[4] 笹田鉄郎，森信介，山肩洋子，前田浩邦，河原達也：レシピ用語の定義とその自動認識のためのタグ付与コーパスの構築，自然言語処理，22(2), pp. 107-131, 2015.

[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，研究報告システムソフトウェアとオペレーティング・システム（OS），2014-OS-131(26), pp. 1-7, 2014.

[6] 苅米志帆乃，藤井敦：料理どうしの類似と組合せに基づく関連レシピ検索システム，言語処理学会第 14 回年次大会発表論文集， pp. 959-962, 2008.

[7] 上野修司，高橋勇，黒岩丈介，白井治彦，小高知宏，小倉久和：複数の Web ページから剽窃したレポートの発見支援システムの実装，情報処理学会研究報告コンピュータと教育（CE），2006-CE-087(130), pp. 41-46, 2006.

[8] Stamatatos, E.: Intrinsic plagiarism detection using character n-gram profiles, In Proc. of the SEPLN '09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, San Sebastian (Donostia), Spain, pp. 38-46, 2009.

[9] Cohen, J.: A coefficient of agreement for nominal scales, Educational and Psychological Measurement, 20(1), pp. 37-46, 1960.

[10] Fleiss, J. L.: Measuring nominal scale agreement among many raters, Psychological Bulletin, 76(5), pp. 378-382, 1971.

[11] Landis, J. R., and Koch, G. G.: The measurement of observer agreement for categorical data, Biometrics, 33(1), pp. 159-174, 1977.