レシピの素性を用いた重複レシピ判別の検証

小邦　将輝† 　島田理紗子†† 　平手　勇宇††† 　杉山　一成†††† 　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
†† 筑波大学大学院 図書館情報メディア研究科　〒 305–8550 茨城県つくば市春日 1-2
††† 楽天株式会社 楽天技術研究所　〒 158–0094 東京都世田谷区玉川 1-14-1 楽天クリムゾンハウス
†††† シンガポール国立大学 計算機科学科　Computing 1, 13 Computing Drive, Singapore 117417
††††† 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: †{s1613123,s1621617}@u.tsukuba.ac.jp, ††yu.hirate@rakuten.com, †††sugiyama@comp.nus.edu.sg,
††††yohei@slis.tsukuba.ac.jp

あらまし　ユーザが投稿したレシピを掲載する「ユーザ投稿型のレシピサイト」には，異なるレシピであるにも関わらず，調理手順が完全に一致しているレシピや，調理手順の一部が変更されている場合でも，同一のレシピとして判断できるレシピ，すなわち，重複レシピが存在する．レシピサイトは多様なユーザが利用することから，検索結果の多様性が要求されるため，重複レシピの存在は有用でない．そこで，本研究ではレシピの素性を用いた重複レシピの判別手法を提案する．提案手法では，調理手順テキスト間類似度と料理画像間類似度の双方を素性に用いて重複レシピを判別する．提案手法の有効性を検証するために，調理手順テキスト間類似度のみを素性に用いて重複レシピを判別する手法，料理画像間類似度のみを素性に用いて重複レシピを判別する手法と比較実験を行い，提案手法は，2 つの比較手法に対して，有意に判別精度が向上することを明らかにした．

キーワード　重複レシピ，機械学習，レシピ検索，剽窃検出

1. はじめに

1.1 背景と目的

料理を作成する際に，以前は料理本や雑誌などの書籍を用いて，レシピを探す機会が多かったのに対して，近年では主婦層や独り暮らしの学生などを中心に，レシピサイトを利用する機会が増加している．マルハニチロホールディングスの調査 (注1) によると，多くの人々が料理をする際にレシピサイトを参考にすることが明らかにされている．また，クックパッドによる調査 (注2) では，料理をする際に最も参考にする情報源として，レシピサイトが挙げられている．

多くのレシピサイトの中でも，ユーザが作成したレシピを Web サイト上に掲載する「投稿型のレシピサイト」が発展している．日本を代表する投稿型のレシピサイトである楽天レシピ (注3) には，140 万件を超えるレシピが投稿されている (2018 年 2 月現在) .

投稿型のレシピサイトには，ユーザによってアレンジされたレシピが幅広いジャンルに渡って掲載されている．また，同一の料理についても，異なった手順や材料を用いて作成されたレシピが存在しており，ユーザは複数のレシピを吟味することによって，自らが調理するレシピを選択することが可能である．例えば，楽天レシピで「肉じゃが」をクエリとして検索した場合 3,925 件 (注4),「かぼちゃの煮物」をクエリとして検索した場合 2,383 件 (注5) のレシピがヒットする．

一方で，投稿型のレシピサイトには，異なるレシピであるにも関わらず，調理手順が完全に一致しているレシピや，調理手順の一部が変更されている場合でも，同一のレシピとして判断できるレシピが存在する．レシピサイトは多様なユーザによって利用されるため，検索結果の多様性 (ダイバーシティ) が要求される [10] が，このようなレシピがサイト上に数多く存在すると，ユーザがレシピの検索を行う際に，検索結果の多様性に影響を及ぼすと考えられる．

本研究では，類似したレシピペアのことを「**重複レシピペア**」と定義する．また，重複レシピペア中で投稿時間が後のレシピを「**重複レシピ**」，投稿時間が前のレシピを「**オリジナルレシピ**」と定義する．すなわち，重複レシピはオリジナルレシピを模倣して作成されたものとみなす．

1.2 重複レシピ

杉山ら [14] は，ユーザ自らが求めるレシピを調べるにあたって，調理手順を比較することを明らかにした．すなわち，ユーザの利便性を高めるためには，レシピ検索結果の多様性が求められる．しかし，レシピサイトには，異なるレシピであるにも関わらず，他のレシピと調理手順が類似した重複レシピが存在しており，これらはレシピ検索結果の多様性を妨げる．

(注1)：https://www.maruha-nichiro.co.jp/news_center/research/pdf/20130227_recipe_cyousa.pdf (参照: 2018 年 2 月 6 日)
(注2)：https://cf.cpcdn.com/info/assets/wp-content/uploads/20140306000000/pr130723-survey.pdf (参照: 2018 年 2 月 6 日)
(注3)：http://recipe.rakuten.co.jp/
(注4)：https://recipe.rakuten.co.jp/search/%E8%82%89%E3%81%98%E3%82%83%E3%81%8C/?s=4&v=0&t=2 (参照: 2018 年 2 月 6 日)
(注5)：https://recipe.rakuten.co.jp/search/%E3%81%8B%E3%81%BC%E3%81%A1%E3%82%83%E3%81%AE%E7%85%AE%E7%89%A9/?s=4&v=0&t=2 (参照: 2018 年 2 月 6 日)

ここで，1.1 節で述べたように，重複レシピの中には，過去に投稿されたレシピとの一致度が極めて高い悪質な重複レシピもあれば，部分的に一致する重複レシピも存在する．前者については，料理レシピサイトの規約に違反していることから，運営者によって，強制的に排除することが可能である．一方，後者についてはレシピサイトの規約に違反しているわけではないため，強制的に排除することはできない．しかし，そのようなレシピがレシピサイト上に存在することは望ましくない．一つの対処方法として，レシピ検索結果において，表示される順位を下げるといった措置を講じることが考えられる．

以上の議論を踏まえて，本研究では，重複の度合いに応じて，完全重複，部分重複，非重複の 3 段階で重複レシピの判定を行う．判定を行う際には，料理画像，材料，調理手順に着目する．なお，本研究では，完全重複，部分重複のレシピを重複レシピとする．以下にそれぞれの判定基準 (注6) を示す．

（1） 完全重複 (重複レシピ)

- 材料，調理手順が一致している

（2） 部分重複 (重複レシピ)

- 材料および分量が類似しており，レシピに新規性がない
- 材料が異なっているが，調理手順が同一であり，レシピに新規性がない

（3） 非重複

- 料理が異なる
- 材料が異なり，レシピに新規性がある
- 調理手順が異なり，レシピに新規性がある

2. 関 連 研 究

杉山ら [14] は，レシピの検索結果において，各レシピが持つ特徴を見比べながら，調理するレシピを調べることは手間のかかる作業であると述べ，レシピ検索では一般的な Web 検索に比べて，上位に提示された検索結果が選ばれない傾向にあること，ユーザは複数のレシピのタイトルや写真，キャッチコピーのみでなく，調理手順を見比べることで作成する料理を選択することを明らかにした．阿部ら [11] は，目安調理時間が書かれていないレシピの場合，限られた時間の中で料理を作成できるかがわからないという点に着目し，教師あり学習による目安調理時間の予測を行った．本研究では，これらの研究を参考に，重複レシピにおける調理手順の類似性に着目する．

小高ら [12]，高橋ら [15] は，文字 n-gram を用いて剽窃したレポートを検出する手法を提案した．これらの研究では，日本語の特徴に基づいて，$n = 3$ として実験が行われた．本研究ではデータセットとして日本語のレシピデータを用いるため，重複レシピペア候補を抽出する際には，これらの研究で有効であった，文字 3-gram を用いる．

Tao ら [9] は，Twitter における類似投稿の発見を目的とした研究を行った．この研究では，ツイートを重複度によって，5 段階に分類し，ツイートから，構文的特徴，意味的特徴，ツイート中に含まれる要素の意味的特徴，文脈的特徴を抽出することで，重複検出に有用な特徴について調査した．この研究と同様に，本研究でも重複レシピについて，多段階で評価を行う．重複レシピに関連する研究として，著者ら [6] の先行研究では，本研究と同様に，レシピを重複度に応じて，3 段階で区別した．本研究でも，重複レシピの評価を行う際には，この判定基準を採用する．

Szegedy ら [8] は，畳み込みニューラルネットワークによる画像認識モデルである Inception-v3 を提案した．Inception-v3 は，ImageNet(注7) が主催する，大規模画像データセットを 1,000 クラスに分類する精度を競うコンテストである ILSVRC(注8) 用に開発された畳み込みニューラルネットワークモデルである．Inception-v3 は，72 層から構成されており，pool_3 層という出力層の 1 層手前の層の出力を用いることで，画像の特徴量を得られる．本研究では，重複レシピの判別にあたり，料理画像の類似性にも着目する．料理画像間類似度を算出する際には，学習済みの Inception-v3 モデルを利用し，重複レシピ候補で用いられている料理画像とオリジナルレシピ候補で用いられている料理画像の特徴量を抽出し，画像特徴量間のコサイン類似度 [2] を計算する．

3. レシピペアの類似性に着目した重複レシピ判別手法

投稿型のレシピサイトの中には，レシピを投稿したユーザに対して，報酬を付与するサイトも存在する．多くの報酬を得るためには，多くのレシピを投稿する必要があるため，ユーザの中には，単に多額の報酬を獲得することを目的として，レシピサイト上に掲載されているレシピを模倣・改変しただけの粗悪なレシピを投稿するユーザが存在すると考えられる．このようなレシピの存在は，レシピ検索結果の多様性を妨げることになるため，こうした悪意のあるレシピを検出し，レシピ検索を行った際のページランキングの順位を下げる等の対策が求められる．

本研究では，重複レシピを投稿するユーザについて，次の 2 つの仮説を立てる．

(a) レシピサイト上に掲載されているレシピ中の調理手順を，完全，もしくは部分的に流用し，レシピを投稿する，

(b) レシピサイト上に掲載されている料理画像を使いまわしてレシピを投稿する，

以上より，本研究では，調理手順テキストと料理画像の両方を用いて重複レシピの判別を行う．また，投稿時間間隔による重複レシピの投稿の傾向の違いについて調査する．

(注6)：本判定基準は，楽天レシピ事業部が定めた基準に沿って作成した．

(注7)：http://image-net.org/

(注8)：http://www.image-net.org/challenges/LSVRC/

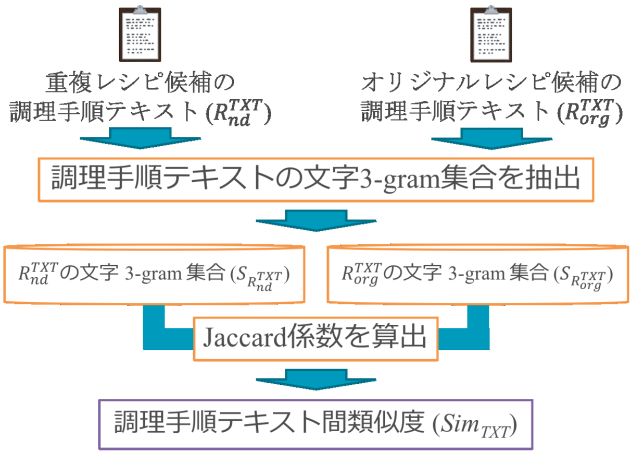

図 1　調理手順テキスト間類似度の算出方法

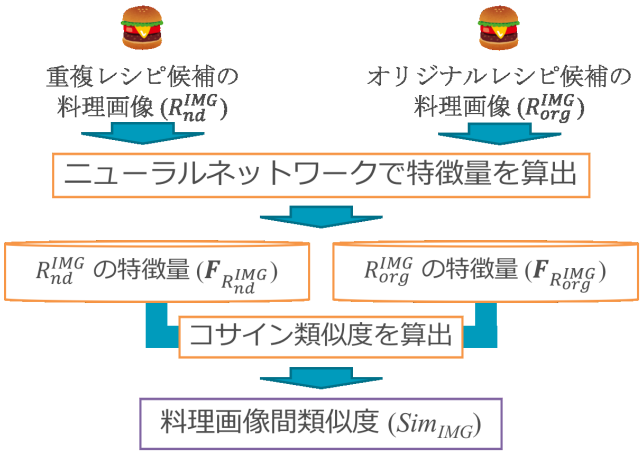

図 2　料理画像間類似度の算出手法

3.1　提案手法の概要

我々の提案手法である，レシピの素性を用いた重複レシピ判別手法では，重複レシピの判別に調理手順テキスト間類似度と料理画像間類似度の双方を素性として用いる．調理手順テキストと料理画像は，全レシピに共通して存在する．異なるレシピ間でこれらが類似していた場合，重複レシピペアである可能性がある．また，双方の素性を用いることによって，片方が投稿者によって変更されていた場合でも，重複レシピを判別することができると考える．

調理手順テキスト間類似度および料理画像間類似度の算出方法については，3.2 節および 3.3 節で述べる．また，重複レシピ候補とオリジナルレシピ候補の抽出方法については，4.1 節で述べる．

3.2　調理手順テキスト間類似度

本研究では，調理手順テキスト間類似度 (Sim_{TXT}) を，2 つのレシピ間の調理手順の類似性を示す指標と定義する．楽天レシピでは，レシピを投稿する際に，調理手順を記述することが必要とされている．各レシピごとに調理手順が記述されているため，異なるレシピ間で調理手順の類似性が高ければ，模倣して作成されたレシピである可能性が高い．

本研究では，著者らの先行研究 [6] と同様に，調理手順テキスト間類似度を文字 n-gram 集合の Jaccard 係数を用いて算出する (図 1)．なお，実験では，日本語で記述されたレシピのデータセットを用いることから，日本語が持つ特徴 [12] をもとに，$n = 3$ とする．重複レシピ候補の調理手順テキスト R_{nd}^{TXT}，オリジナルレシピ候補の調理手順テキスト R_{org}^{TXT} の 3-gram 集合 をそれぞれ $S_{R_{nd}^{TXT}}$, $S_{R_{org}^{TXT}}$ とすると，調理手順テキスト間類似度 Sim_{TXT} は式 (1) で示される．

$$Sim_{TXT} = \frac{|S_{R_{nd}^{TXT}} \cap S_{R_{org}^{TXT}}|}{|S_{R_{nd}^{TXT}} \cup S_{R_{org}^{TXT}}|} \tag{1}$$

3.3　料理画像間類似度

本研究では，料理画像間類似度 (Sim_{IMG}) を，2 つのレシピ間の料理画像の類似性を示す指標と定義する．楽天レシピでは，レシピを投稿する際に，料理の完成画像を添付することが必要とされている．各レシピごとに，料理の完成画像は異なるため，異なるレシピ間で料理画像の類似性が高い場合，画像を使いまわしたり，加工したりするなどして，実際には 1 つのレシピしか作成していないにも関わらず，2 つのレシピを作成したように装い，レシピを投稿している可能性が考えられる．また，料理画像を重複レシピの判別の手がかりとして用いることで，調理手順テキスト中で書き換えを行っている場合であっても，重複レシピの判別精度を高められるものと期待できる．

図 2 に料理画像間類似度の算出手法の概要を示す．料理画像間類似度の算出は，学習済みの Inception-v3 モデルを利用し，重複レシピ候補で用いられている料理画像 (R_{nd}^{IMG})(注9) とオリジナルレシピ候補で用いられている料理画像 (R_{org}^{IMG})(注10) の特徴ベクトルを構築し，これらの特徴ベクトル間のコサイン類似度 [2] を求めることで行う．特徴ベクトル $\boldsymbol{F}_{R_{nd}^{IMG}}$ を持つ画像と特徴ベクトル $\boldsymbol{F}_{R_{org}^{IMG}}$ を持つ画像の間の料理画像間類似度 Sim_{IMG} は，式 (2) で示される．

$$Sim_{IMG} = \frac{\boldsymbol{F}_{R_{nd}^{IMG}} \cdot \boldsymbol{F}_{R_{org}^{IMG}}}{|\boldsymbol{F}_{R_{nd}^{IMG}}||\boldsymbol{F}_{R_{org}^{IMG}}|} \tag{2}$$

コサイン類似度はベクトル間の距離を示しており，1 に近づくほどベクトル間の距離は近い．すなわち料理画像間の類似度が高いことを示している．

4.　実験: レシピペアの類似性に着目した重複レシピ判別手法の評価

本章では，重複レシピを判別する際に，調理手順テキスト間類似度と料理画像間類似度の双方を考慮する我々の提案手法が有効であることを検証するために，評価実験を行った結果を示す．実験では，3.1 節で示した提案手法，および 4.2 節で述べる 2 つの比較手法を重複レシピの判別に適用する．なお，実験では，多層パーセプトロン，サポートベクターマシン，ランダムフォレストを用いて，重複レシピの判別を行う．

以下，実験に用いるデータセット，実験方法，実験結果について述べる．

(注9)：重複を表す “near-duplicate” から “nd” とする
(注10)：オリジナルを表す “original” から “org” とする

4.1 実験データ

本実験では，楽天レシピのデータセットを使用する．本データセットは，2010 年 6 月 30 日から 2016 年 11 月 8 日の間に楽天レシピに投稿された 1,353,406 件の各レシピについて，ユーザ ID，レシピ ID，レシピ投稿時間，調理手順などの情報から構成されている．

レシピサイトの中には 3 章で述べたように，レシピの投稿に対して，報酬を与えるサイトも存在する．そこで，単に報酬の獲得を目的として短時間に多くのレシピを投稿するユーザがいると仮定し，本実験では，短時間に集中的に投稿されたレシピを対象として，重複レシピの判別を行う．1 時間の間にレシピを集中的に投稿したユーザ数について調査を行ったところ，表 1 に示す結果となった．

表 1 より，楽天レシピにおいて，1 時間に 30 件以上のレシピを投稿したユーザが確認された．楽天レシピのシステムにおいて，1 レシピを作成するためには，少なくとも 5 分程度は時間を要すると考えられることから，1 時間に 30 件以上ものレシピを投稿することはきわめて不自然である．

ここで，投稿件数に応じて，ユーザを次の各群に分類する．

- ユーザ群 U_A: 1 時間に 10 件以上 20 件未満のレシピを投稿した 227 ユーザのうち，投稿レシピ件数の下位 40 ユーザ
- ユーザ群 U_B: 1 時間に 20 件以上 30 件未満のレシピを投稿した全 27 ユーザ
- ユーザ群 U_C: 1 時間に 30 件以上のレシピを投稿した全 20 ユーザ

続いて，ユーザ群 U_A, U_B, U_C の各ユーザによって投稿された各レシピと，そのレシピが投稿される時刻以前に投稿されたレシピの間で文字 3-gram の Jaccard 係数を調理手順テキスト間類似度として算出し，各レシピについて最も調理手順テキスト間類似度が高くなったレシピペアを，重複レシピペア候補とする (図 3)．

本実験では，実験の正解データを作成することを目的として，各ユーザ群調理手順テキスト間類似度の高い順に，

- ユーザ群 U_A: 10 件の重複レシピペア候補
- ユーザ群 U_B: 20 件の重複レシピペア候補
- ユーザ群 U_C: 30 件の重複レシピペア候補

についてアノテーションを行う．アノテーションを行う際には，重複レシピ候補とオリジナルレシピ候補の Web ページ上に掲載されている，料理画像，材料，調理手順などから，レシピペアの関係性について評価を行う．アノテーションを行う際には，1.2 節で示した，完全重複，部分重複，非重複の 3 段階のアノテーション基準を用いる．なお，楽天レシピの事業部に，無作為に選出した重複レシピペア候補の 120 件についてアノテーションを依頼し，判定者間の回答の一致度を示す κ 係数 [1] について調査した結果，0.885 (Almost perfect [4]) となった．この結果から，アノテーション基準には一貫性があると判断し，残りの重複レシピペア候補については著者がアノテーションを行った．

アノテーションを行った結果を表 2 に示す．

アノテーションの結果，1 時間に集中投稿されたレシピ数が多い (レシピの投稿時間間隔が短い) ほど，重複レシピの割合が高いことが明らかになった．これは，レシピの集中時間間隔が短い場合は，他のレシピの剽窃を行ってレシピの投稿を行っていることを示唆している．

なお，ユーザ群 U_C に属するユーザ 1 名に関して，ユーザが退会していたことから，評価を行わず，19 ユーザが投稿したレシピに関して評価を行った．また，ユーザ群 U_A に属するユーザが投稿した 2 レシピ，ユーザ群 U_C に属するユーザが投稿した 1 レシピに関しては，オリジナルレシピ候補が削除されていたため，評価の対象としていない．

表 1　集中投稿を行ったユーザの調査結果

1 時間の集中投稿レシピ数 n	ユーザ数
$10 \leqq n < 20$	227
$20 \leqq n < 30$	27
$30 \leqq n$	20

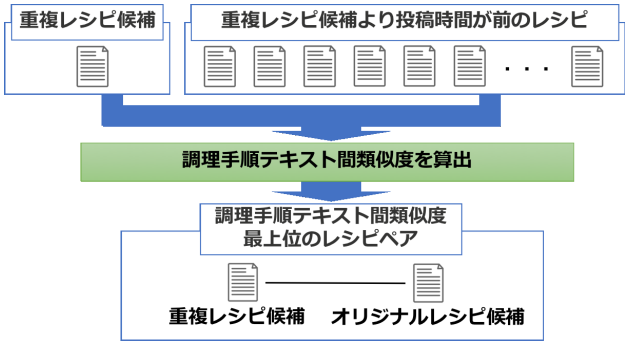

図 3　重複レシピペア候補の抽出手法

表 2　重複レシピペア候補に対するアノテーション結果

ユーザ群	ユーザ数	完全重複	部分重複	非重複	合計	重複レシピペア数
U_A	40	0	106	292	398	106 (26.6%)
U_B	27	6	298	236	540	304 (56.3%)
U_C	19	3	362	204	569	365 (64.1%)

4.2 実験方法

評価実験では，4.1 節でアノテーションを行ったレシピを対象として，提案手法と，次に述べる点を考慮した比較手法 1, 2 の 3 つの手法を適用して，重複レシピの判別を行う．

- 比較手法 1: 調理手順テキスト間類似度のみ
- 比較手法 2: 料理画像間類似度のみ

実験では，多層パーセプトロン，サポートベクターマシン，ランダムフォレストを用いて，重複レシピの判別を行う．分類を行う際には，4.1 節でアノテーションを行ったデータのうち，正例のデータには重複レシピ (完全重複，部分重複) のデータ，負例のデータには非重複レシピのデータを用いる．すなわち，正例のデータとして，アノテーションで完全重複もしくは部分重複として判定された 775 件のデータ，負例のデータとして，アノテーションで非重複と判定された 732 件のデータを用いる．

重複レシピの分類精度の評価尺度には，F 値を採用する．また，評価方法には，10 分割交差検定を採用する．

なお，実装には，scikit-learn(注11)を使用した．また，ハイ

(注11)：http://scikit-learn.org/stable/ (Version: 0.19.1)

表 3　レシピの素性を用いた重複レシピの分類精度 (F 値)

手法名	多層パーセプトロン	サポートベクターマシン	ランダムフォレスト
提案手法 ($Sim_{TXT} + Sim_{IMG}$)*	**0.768**	**0.762**	**0.752**
比較手法 1 (Sim_{TXT})	0.739	0.742	0.719
比較手法 2 (Sim_{IMG})	0.657	0.702	0.628

* 比較手法 1 との間で，t-検定 (両側検定，有意水準 5%，p=0.034) で有意に向上．比較手法 2 との間で，t-検定 (両側検定，有意水準 5%，p=0.011) で有意に向上

表 4　形態素数変化度を素性として用いた重複レシピの分類精度 (F 値)

素性の組み合わせ	多層パーセプトロン	サポートベクターマシン	ランダムフォレスト
D_{mor_c}	0.690	0.709	0.700
Sim_{TXT}	0.739	0.742	0.719
Sim_{IMG}	0.657	0.702	0.628
$Sim_{TXT} + D_{mor_c}$	0.742	0.737	0.731
$Sim_{IMG} + D_{mor_c}$	0.688	0.720	0.688
$Sim_{TXT} + Sim_{IMG}$	**0.768**	**0.762**	**0.752**
$Sim_{TXT} + Sim_{IMG} + D_{mor_c}$	0.761	**0.762**	0.751

パーパラメータは，多層パーセプトロンについてはデフォルトのものを使用し，サポートベクターマシンとランダムフォレストについては，グリッドサーチを用いてハイパーパラメータの調整を行った．

4.3 実験結果

実験の結果を表 3 に示す．

表 3 より，調理手順テキスト間類似度と料理画像間類似度の双方を考慮した提案手法のほうが，いずれか片方のみを考慮する比較手法 1, 2 よりも，重複レシピの判別精度が有意に向上している．このことから，重複レシピの判別において，調理手順テキスト間類似度 (Sim_{TXT}) と料理画像間類似度 (Sim_{IMG}) の双方を考慮することが有効であることが示された．また，比較手法 1, 2 の結果から，調理手順テキスト間類似度 (Sim_{TXT}) のほうが，料理画像間類似度 (Sim_{IMG}) と比較して，重複レシピを判別する際の有効な素性であることがわかった．

5. 考　　察

本章では，本研究の考察として，5.1 節で，形態素数変化度を素性として用いて重複レシピの判別精度を検証した結果を示し，5.2 節で，提案手法の課題について述べる．

5.1 形態素数変化度を素性に用いた重複レシピ判別の検証

4 章で行った実験について，どのようなときに重複レシピを検出できなかったか調査を行った結果，調理手順テキスト間類似度が低いときに，比較手法 1 で検出できない傾向が見られ，料理画像間類似度が低いときに比較手法 2 で検出できない傾向が見られた．一方で提案手法を用いた際には，いずれかの類似度が低い場合でも，もう一方の類似度が高い場合には検出が可能であった．この結果より，素性を組み合わせることの有効性が示された一方で，両方の類似度が低い場合には検出が困難であるといえる．

実際の事例について分析を行ったところ，以下の 2 文のように，調理手順テキスト中で言い換えや書き換えが行われている場合でも形態素数の変化は少ないレシピが存在することが明らかになった．

(a)　サツマイモは 1cm ほどの厚みの輪切りにカットする

(b)　さつまいもは 1cm くらいの厚さに輪切りにしておく

以上のことから，形態素数変化度を素性として用いることで重複レシピの判別精度が向上すると期待できる．本節では形態素変化度を素性として用いて，重複レシピの判別精度を検証した結果を示す．

本研究では，形態素数変化度 (D_{mor_c}) を式 (3) で定義する．なお，MOR_{nd} は，重複レシピ候補の形態素数，MOR_{org} は，オリジナルレシピ候補の形態素数を示す．

$$D_{mor_c} = \frac{MOR_{nd} - MOR_{org}}{MOR_{org}} = \frac{MOR_{nd}}{MOR_{org}} - 1 \quad (3)$$

本実験において，形態素解析を行う際には，MeCab [3] を使用し，辞書には NEologd [13] を用いる．

実験では，4 章における実験と同様に，多層パーセプトロン，サポートベクターマシン，ランダムフォレストを用いて，重複レシピの判別を行う．分類を行う際には，4.1 節でアノテーションを行ったデータのうち，正例のデータには重複レシピ (完全重複，部分重複) のデータ，負例のデータには非重複レシピのデータを用いる．また，重複レシピの分類精度の評価尺度には，F 値を採用し，評価方法には，10 分割交差検定を採用する．

実験結果を，表 4 に示す．表 4 より，形態素数変化度のみを素性として用いた場合 (D_{mor_c})，3 つの分類器すべてにおいて，料理画像間類似度を素性として用いた場合 (Sim_{IMG}) の結果を上回っていることがわかる．その一方で，調理手順テキスト間類似度と形態素数変化度を素性として用いた場合 (Sim_{TXT}+D_{mor_c}) の判別精度は，調理手順テキスト間類似度と料理画像間類似度を素性として用いた場合 ($Sim_{TXT} + Sim_{IMG}$) の分類精度に及ばなかった．調理手順テキスト間類似度と形態素数変化度は，レシピのテキスト部分から算出される素性である．一方，

料理画像間類似度は，レシピの画像部分から算出される素性である．このことから，重複レシピを判別する際には，異なる要素から算出される素性を組み合わせて用いることで，判別精度をより向上させることができると考えられる．ここで，同じく異なる要素から算出される素性の組み合わせである，形態素数変化度と料理画像間類似度を素性として用いた場合の分類精度 ($Sim_{IMG} + D_{mor_c}$) が，有効な分類精度を示せなかった点について，調査を行った．その結果，先述したように言い換えや書き換えが行われている際に形態素数の変化が小さいものもあれば，オリジナルレシピと重複レシピの間で形態素数に大幅な隔たりがある場合も見られた．このため，異なる要素から算出される素性を組み合わせたにもかかわらず，分類精度が向上しなかったと考えられる．

また，本実験において，提案手法を適用した際には，多層パーセプトロンを分類器として用いたときに最良の分類精度が得られた．なお，用いる素性によっては，サポートベクターマシンを分類器として用いたときのほうが，多層パーセプトロンの分類精度を上回る結果となり，ランダムフォレストを分類器として用いた際には，ほぼすべての素性の組み合わせにおいて，3 つの分類器の中で分類精度が低くなった．多層パーセプトロンとサポートベクターマシンを分類器として用いた際に，分類精度が高くなった原因として，サポートベクターマシンは，識別境界面からサポートベクターまでのマージンを最大化するというアルゴリズムで動作しているため，未学習データに対しての高い識別性能を持つこと，多層パーセプトロンは，多層のニューラルネットワークを構築することで，高い表現力を持つことが挙げられる．一方で，ランダムフォレストは高次元のデータの分類に強いという特徴を持つが [16]，最大でも 3 次元の素性しか用いなかった本実験においては有効な分類精度を示せなかったと考える．

5.2 提案手法の課題

提案手法では，重複レシピペア候補を抽出する際に，重複レシピ候補と重複レシピ候補より投稿時間が早いレシピの文字 3-gram 集合の Jaccard 係数を用いた．このことから，短い調理手順からなるレシピにおいて，書き換えが行われた場合に対応することが難しい．例えば，人間は以下の 2 文について，意味が類似していることを認識することができる．

a) 炊飯器でお米を炊く

b) 炊飯器でライスをたく

しかし，これらの 2 つの文の間の文字 3-gram 集合の Jaccard 係数は 0.15 になる．このように，複数の表現がある食材や，表記揺れなどによる書き換えがあると，文字 3-gram 集合の Jaccard 係数は低くなる．この問題に対処するために，食材辞書を構築する方法や，単語の分散表現 [5] を用いて，単語ベクトルが類似している場合には，同一の単語として扱う方法が考える必要がある．

また，本研究では，料理画像間類似度を算出する際に，学習済みの畳み込みニューラルネットワークモデル Inception-v3 を用いた．しかし，Inception-v3 は，1,000 クラス分類を行うことを目的として作成されたモデルであり，学習データに必ずしも料理画像が用いられているわけではない．そこで，料理画像を学習データとして，再度モデルを学習することによって，類似度の算出精度の改善が期待できる．

6. おわりに

「ユーザ投稿型のレシピサイト」には，異なるレシピであるにも関わらず，調理手順が完全に一致しているレシピや，調理手順の一部が変更されている場合でも，同一のレシピとして判断できるレシピ，すなわち重複レシピが存在する．レシピサイトは多様なユーザが利用することから，検索結果の多様性が要求されるため，このようなレシピの存在は有用でない．本研究では，調理手順テキスト間類似度と料理画像間類似度の双方を素性に用いて，重複レシピの判別を行う手法を提案した．

提案手法の評価実験では，多層パーセプトロン，サポートベクターマシン，ランダムフォレストを分類器として用いて，重複レシピの判別を行った．評価実験の結果，調理手順テキスト間類似度と料理画像間類似度の双方を素性として用いる提案手法が，どちらか一方のみを素性として重複レシピの判別を行う手法に対して，有意に判別精度が向上することを明らかにした．

本研究では，調理手順テキスト間類似度の算出に文字 3-gram 集合の Jaccard 係数を用いた．しかし，この手法は書き換えに頑健でない．現在は，単語の分散表現 [5] や，系列変換モデル [7] などを用いて，重複レシピの判別を行う手法を検討している．将来的にはレシピ以外のコンテンツを対象とした，重複文書の判別技術への応用も検討している．

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．

文　献

[1] Cohen, Jacob. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 1960, Vol. 20, No. 1, p. 37-46.

[2] Kaur, Sukhdeep; Aggarwal, Deepak. “Image Content Based Retrieval System using Cosine Similarity for Skin Disease Images”. ACSIJ Advances in Computer Science: an International Journal. 2013, p. 89-95, Vol. 2, No. 5.

[3] Kudo, Taku; Yamamoto, Kaoru; Matsumoto, Yuji. “Applying Conditional Random Fields to Japanese Morphological Analysis”. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004). 2004, p. 230-237.

[4] Landis, Richard J.; Koch, Gary G.. The Measurement of Observer Agreement for Categorical Data. Biometrics, 1977, Vol. 33, No. 1, p. 159-174.

[5] Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey. “Efficient estimation of word representations in vector space”. Proceedings of the workshop at the 1st International Conference on Learning Representations, 2013.

[6] Oguni, Masaki; Seki, Yohei; Shimada, Risako; Hirate, Yu. “Method for Detecting Near-duplicate Recipe Creators

Based on Cooking Instructions and Food Images". Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities (CEA 2017). 2017, p. 49-54.

[7] Sutskever, Ilya; Vinyals, Oriol; Le, Quoc V.. "Sequence to Sequence Learning with Neural Networks". Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS 2014). 2014, p. 3104-3112.

[8] Szegedy, Christian; Vanhoucke, Vincent; Ioffe, Sergey; Shlens, Jon; Wojna, Zbigniew."Rethinking the Inception Architecture for Computer Vision". The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). 2016, p. 2818-2826.

[9] Tao, Ke; Abel, Fabian; Hauff, Claudia; Houben, Geert-Jan; Gadiraju, Ujwal. "Groundhog Day: Near-Duplicate Detection on Twitter". Proceedings of the 22nd International Conference on World Wide Web (WWW 2013). 2013, p. 1273-1283.

[10] Wang, Xiaojie; Dou, Zhicheng; Sakai, Tetsuya; Wen, Ji-Rong. "Evaluating Search Result Diversity Using Intent Hierarchies". Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016). 2016, p. 415-424.

[11] 阿部卓也, 立間淳司, 青野雅樹. "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". 第 14 回情報科学技術フォーラム講演論文集 (FIT 2015). 2015, p. 103-104, Vol. 14, No. 2.

[12] 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. 2003, p. 702-705, Vol. 86, No. 9.

[13] 佐藤敏紀, 橋本泰一, 奥村学. "単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討". 言語処理学会第 23 回年次大会 (NLP 2017). 2017, p. 875-878.

[14] 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). 2013, D3-5.

[15] 高橋勇, 宮川勝年, 小高知宏, 白井治彦, 黒岩丈介, 小倉久和. "Web サイトからの剽窃レポート発見支援システム". 電子情報通信学会論文誌, 2007, p. 2989-2999, Vol. 90, No. 11.

[16] 新妻雅弘, 斎藤博昭. "Random Forest を用いた音楽ジャンル分類". 情報処理学会論文誌, 2009, p. 2910-2922, Vol. 50, No. 12.