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

近年，料理をする際にレシピサイトを利用する人は多く，レシピサイトは，ユーザにとって重要な情報源である．特に，レシピサイトの中でも，クックパッドや楽天レシピなどの，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．また，ユーザによって求めるレシピは異なるため，ユーザはより多くのレシピから，複数の投稿型レシピサイトを利用してレシピを検索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，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，レシピ検索をするユーザにとっては有用ではない．そのため，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．重複レシピのうち一つを残すためには，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．この知見に基づき，本研究では，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て，実装を行った．

重複レシピの判別の手がかりを見つけるために，重複レシピの分析を行った．分析には，先行研究を参考に，レシピカテゴリをバランス良くカバーできるように考慮した，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，クックパッドと楽天レシピから抽出して使用する．まず，この 10 種類の料理に関して，すべてのレシピの調理内容のペアについて，文字 n-gram で Jaccard 係数を用いることで，2 つのレシピ間の類似度を計算する．次に，類似度に基づいてランキングをし，上位 200 件から重複レシピを抽出する．また，抽出した重複レシピの一部を分析して，重複レシピの調理内容における差異を明らかにした．さらに，明らかにした差異のうち，調理内容の一部の代替を考慮することで，重複レシピの判別の精度が向上するか検証した．

実験には，前述の 10 種類の料理のレシピデータを使用して，記号，文末，数字，切り方，材料といった調理内容の一部の代替を考慮する 5 種類の置換辞書を人手で作成する．まず，置換辞書を利用して，調理内容の同一視されうる異なる要素を同じ要素に置換する．次に，レシピペアを類似度に基づきランキングし，上位 100 件において重複レシピがどれだけ含まれるか判定する．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．

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．

（指導教員　関 洋平）