D-006

食材集合を用いたレシピ分類
A Recipe Classification Method Using Sets of Main Ingredients

邑中 ひな† 加藤 未響‡ 小早川 倫広‡ 岩田 満‡
Hina Muranaka Miki Kato Michihiro Kobayakawa Mitsuru Iwata

1. はじめに

レシピ提供サービスが増加したことにより、膨大なレシピデータが提供されるようになった。それらを効率的に活用するために、レシピを対象とした推薦や分類の研究が行われ始めた。

Nuno らは、レシピ内から各食材の名前、数量、調理方法などを自動的に抽出、レシピの栄養情報を推定し、レシピ検索における推薦のサポートを行うシステムを開発した[1]。Han らは、6,006 のレシピデータに対して食材から 5,073 次元の特徴量ベクトルを算出し、5,073 次元空間に対して SVM による分類を行った[2]。

食材をデータ表現する方法は、食材をベクトル表現する方法と、集合表現する方法の 2 種類がある。前者の方法では、2 つのレシピの類似性はベクトル間距離を利用可能であり、一般的なクラスタリング手法が適用可能である。一方後者の方法では、集合と集合の間の集合間類似度のみが算出可能であり、距離を用いた一般的なクラスタリング手法を適用できない。食材をベクトル化し空間上で分類を行う研究はあるものの、食材を集合とみなし空間上で分類を行う研究は見受けられなかった。

本研究では、レシピの食材集合の集合間類似度に基づくレシピ分類手法を提案することを目的とする。

2. 提案するレシピ分類の枠組み

本研究におけるレシピ分類の枠組みを示す。本研究では、食材集合から算出した集合間類似度に対して多次元尺度構成法を適用することにより、食材集合すなわちレシピデータをユークリッド空間に写像し、それらのデータを料理ジャンルごとに分類する。

① $n(n = 1, 2, \dots, N)$番目のレシピ$\boldsymbol{r_n}$を取得し、レシピ集合$R = \{r_1, r_2, \dots, r_N\}$を得る。
② レシピ$\boldsymbol{r_n}(n = 1, 2, \dots, N)$に含まれる食材名の表記揺れを、料理オントロジー[3]を用いて補正する。補正された食材集合を特徴集合$\boldsymbol{i_n}(n = 1, 2, \dots, N)$とする。この特徴集合$i_n$の集合を$I = \{i_1, i_2, \dots, i_N\}$とする。
③ レシピ$\boldsymbol{r_\alpha}$とレシピ$\boldsymbol{r_\beta}$の集合間類似度$s_{\alpha\beta} = Sim(i_\alpha, i_\beta)(\alpha = 1, 2, \dots, N, \alpha \neq \beta)$を算出する。
④ $N \times N$の類似度マトリックスMを定義する。ここで、マトリックス M の(α, β)要素を集合間類似度$s_{\alpha\beta}$とする。ただし、対角成分は 1 とする。
⑤ 類似度マトリックスMに対し、多次元尺度構成法を用いてレシピ$\boldsymbol{r_n}(n = 1, 2, \dots, N)$を$k(k = 1, 2, \dots, N)$次元のユークリッド空間上に写像する。レシピ$\boldsymbol{r_n}$が写像された座標を特徴ベクトル$\boldsymbol{v_n}$とする。
⑥ 特徴ベクトル$\boldsymbol{v_n}$の集合に対して k-means++法を適用し、m個のクラスタを得る。
⑦ 各クラスタに対して料理ジャンルを付与する。

3. 予備実験

3.1 レシピデータセットの作成

本研究では楽天レシピデータを使用した[4]。計 796,274 件あるレシピの中から、レシピタイトルに表 1 の料理ジャンル名が含まれるレシピをそれぞれ 250 件ずつ、計 5,000 件のレシピ集合を取得した。

表 1　料理ジャンル一覧

ラベル	料理ジャンル	ラベル	料理ジャンル	ラベル	料理ジャンル
A	ポトフ	H	ピーマンの肉詰め	O	麻婆豆腐
B	ソース焼きそば	I	ラタトゥイユ	P	肉じゃが
C	ナポリタン	J	親子丼	Q	中華丼
D	ゴーヤチャンプル	K	ビビンバ	R	ビーフシチュー
E	筑前煮	L	タコライス	S	豚汁
F	白和え	M	ニラ玉	T	ミネストローネ
G	冷やし中華	N	マカロニグラタン	-	-

3.2 多次元尺度構成法の次元設定

Jaccard 係数を用いて類似度マトリックスを算出する。ここで多次元尺度構成法では写像するユークリッド空間の次元数kを任意に設定できるため、k=3, 8, 16, 32, 64, 128, 256, 512, 1024 次元でレシピ分類を行い、最大平均 F 値を得た次元数における分類について着目する。最大平均 F 値 0.707 を得た次元数は，k=128 であった。

4. 多次元尺度構成法を用いたレシピ分類実験

128 次元の特徴量を用いた分類結果を表 2 に示す。表 2 のアルファベットは表 1 の料理ジャンルと対応している。行はある料理ジャンルのレシピが実際にはどのクラスタに分布していたのかを表し、列はクラスタ内に各料理ジャンルが何個含まれているかの分類結果を表している。また、最右列、最下行にはそれぞれ各料理ジャンルの再現率、適合率、最も右下のセルには再現率の平均を示している。最大値には赤、最小値には青の色付けをしている。

† NEC ネッツエスアイ株式会社　NEC Networks & System Integration Corporation
‡ 東京都立産業技術高等専門学校　Tokyo Metropolitan College of Industrial Technology

表 2　128 次元におけるレシピ分類の結果

料理ジャンル名 ＼ 割り当てられた料理ジャンル名	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	再現率
A	162	12	9	0	2	0	0	0	3	0	3	0	0	1	0	17	5	7	18	11	0.648
B	10	122	8	1	1	0	0	1	0	0	9	0	0	0	0	0	71	0	23	4	0.488
C	6	35	184	0	0	0	0	3	6	0	0	0	0	2	0	1	3	0	0	10	0.736
D	0	6	2	215	0	1	2	4	0	2	0	0	1	0	0	0	14	0	3	0	0.860
E	3	3	0	1	234	4	0	0	0	0	1	0	0	0	0	0	0	0	4	0	0.936
F	0	0	0	0	0	249	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.996
G	0	39	2	2	0	0	202	0	2	2	0	1	0	0	0	0	0	0	0	0	0.808
H	0	42	24	4	1	0	0	149	0	14	1	0	1	0	1	0	11	0	1	1	0.596
I	1	14	4	0	1	0	0	0	178	0	0	0	0	0	0	1	2	1	0	48	0.712
J	0	1	0	0	1	0	1	0	0	242	2	0	1	0	0	0	2	0	0	0	0.968
K	0	29	1	1	0	3	1	0	0	3	195	1	0	0	0	0	9	0	7	0	0.780
L	1	28	1	0	0	0	0	6	4	2	1	188	0	0	0	2	9	0	1	7	0.752
M	0	1	0	1	0	0	0	1	0	1	5	0	239	0	0	0	2	0	0	0	0.956
N	2	1	0	0	0	0	0	0	0	0	0	0	0	246	0	0	1	0	0	0	0.984
O	0	11	3	5	0	0	0	0	1	1	0	0	0	0	229	0	0	0	0	0	0.916
P	38	2	1	0	1	0	0	0	0	0	0	1	0	0	0	98	12	81	14	2	0.392
Q	4	53	2	2	1	1	0	1	1	1	51	2	2	0	0	0	73	0	55	1	0.292
R	15	6	11	0	0	0	0	0	5	0	0	1	0	0	0	3	1	192	2	14	0.768
S	2	28	0	2	0	1	0	0	0	0	0	0	0	0	1	16	36	0	164	0	0.656
T	53	6	11	0	0	1	0	0	14	0	0	1	0	0	0	2	5	2	1	154	0.616
適合率	0.546	0.278	0.700	0.919	0.967	0.958	0.981	0.903	0.832	0.903	0.728	0.964	0.980	0.988	0.991	0.700	0.285	0.678	0.558	0.611	0.743

5. 考察

再現率が最も高い料理ジャンルは F(白和え)の 0.996、再現率が最も低い料理ジャンルは Q(中華丼)の 0.292 だった。また、適合率が最も高い料理ジャンルは O(麻婆豆腐)の 0.991、適合率が最も低い料理ジャンルは B(ソース焼きそば)の 0.278 だった。さらに、F 値が最も高い料理ジャンルは N(マカロニグラタン)の 0.986、F 値が最も低い料理ジャンルは Q(中華丼)の 0.289 だった。

本実験での平均適合率は 0.773、平均再現率は 0.743 だった。しかし、適合率が 0.8 を超えた料理ジャンルは 11 個、再現率が 0.8 を超えた料理ジャンルは 8 個だった。これは、適合率、再現率共に値が 0.3 を下回る料理ジャンルが存在し、値に影響を与えていると考えられる。

再現率が最も高い F(白和え)と、F 値が最も高い N(マカロニグラタン)について着目すると、他の料理ジャンルとの食材の類似性が低いことが確認できた。特に N(マカロニグラタン)に含まれている食材である「マカロニ」は、他の料理ジャンルにはほとんど含まれていなかった。つまり、他料理ジャンルとの共通食材の少ない料理ジャンルにおいてはレシピ分類を行うことができ、かつレシピ分類性能は高い結果になるといえるが、共通食材が多い料理ジャンルは、類似度算出の時点で影響が出るためレシピ分類性能が悪い結果になるといえる。

F 値が最も低かった Q(中華丼)に着目すると、同料理ジャンル内でのレシピ間類似度が低いことが分かった。また、表 2 より、B(ソース焼きそば)にも Q(中華丼)と同等の数のレシピが分類されていることが分かる。そこで、「ソース焼きそば」と「中華丼」に含まれている食材を見ると、共通食材が多い上、この 2 つの料理ジャンルには他の料理ジャンルに比べて「もやし」、「イカ」、「エビ」という食材が多く含まれていた。このような共通点により、レシピ分類に相互的に影響を与えていると考えられる。つまり、同料理ジャンル内でも食材の類似性が低い場合は、分類性能を下げる影響を与えるといえる。

同料理ジャンル内での食材の類似性が低い原因は、食材の重複数が少ないほかに、差集合の要素数に左右されやすいという Jaccard 係数の特性によるものであると考えられる。この問題には、差集合の要素数による影響を防ぐべく閾値ありの Simpson 係数を使用するほか、その料理ジャンルにおける最頻食材に重みづけをするなどの対策が考えられる。

6. まとめ

本研究では、レシピの食材集合の集合間類似度に基づくレシピ分類の枠組みを提案した。実験では、20 料理ジャンル計 5,000 個のレシピ集合について、食材集合の集合間類似度に基づく多次元尺度構成法を用いたユークリッド空間上での分類を行い、その性能を確認した。

謝辞

本研究では、国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」 (https://rit.rakuten.com/data_release/) を利用した。

参考文献

[1] Nuno Silva, David Ribeiro, Liliana Ferreira, “Information Extraction from Unstructured Recipe Data,” ICCTA ’19:Proceedings of the 2019 5th International Conference on Computer and Technology Applications, pp. 165–168, (2019).

[2] Han Su, Ting-Wei Lin, Cheng-Te Li, Man-Kwan Shan and Janet Chang, “Automatic Recipe Cuisine Classification by Ingredients,” UbiComp ’14 Adjunct: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, pp. 565–570, (2014).

[3] 難波英嗣, “料理オントロジー構築プロジェクト,” 自然言語処理研究室 (難波研究室), http://nlp.indsys.chuo-u.ac.jp/cgi-bin/cooking/wiki.cgi, (2023).

[4] 楽天グループ株式会社, “楽天レシピデータ, ” 国立情報学研究所情報学研究データリポジトリ, https://doi.org/10.32130/idr.2.4, (2016).