5K-05

間接相関ルールを用いた代替食材の抽出

川口 美香 † 尾崎 知伸 †
† 日本大学 文理学部

1. はじめに

SNS の発展と健康志向の高まりにより，クックパッドや楽天レシピなど，利用者によるレシピ投稿サイトが注目を集めている．本研究では，レシピ投稿サイトからの知識発見の一つとして，代替可能食材の発見を対象とする．代替可能食材を探す目的は，アレルギーへの対応の他，より安価で健康的な食材を利用したい，より簡単に調理を行いたいなど多岐にわたる[1)~3)]．これらの多様なニーズに応える一般的な枠組みとして，本研究では，間接相関ルール[4)] を用いた代替可能食材の発見を基本とし，料理手順や食材の組み合わせなど，種々の観点に基づく評価関数を提案する．

2. 代替可能食材の発見

レシピ $R = \langle id_R, I_R, S_R \rangle$ を，識別子 id_R，利用する食材の集合 I_R，調理手順のリスト S_R の 3 つ組で表す．レシピの集合 $D = \{R_1, \cdots, R_{|D|}\}$ をレシピデータベースと呼ぶ．

2.1 間接相関ルール

食材集合 P の D における支持集合を $occ_D(P) = \{\langle id, I, S \rangle \in D \mid P \subseteq I\}$，支持度を $sup_D(P) = |occ_D(P)| \,/\, |D|$ と定義する．食材間の相関ルール $X \to y$ とは，条件 $y \notin X$ を満たす食材（の飽和）集合 X と食材 y の関係を表すルールであり，その支持度と確信度はそれぞれ $sup_D(X \to y) = sup_D(X \cup \{y\})$，$conf_D(X \to y) = sup_D(X \cup \{y\})/sup_D(X)$ と定義される．

同一の本体部を持つ 2 つの相関ルール $M \to c_a$ と $M \to c_b$ の対が，支持度，確信度，帰結の共起に関する 3 つの閾値 σ, τ, θ に対して下記の条件を満たすとき，この対を間接相関ルール[4)] と呼び，$r = M \to \{c_a, c_b\}$ と表記する．

$$sup_D(r) = \min\left(sup_D(M \to c_a), sup_D(M \to c_b)\right) \geq \sigma \wedge$$
$$conf_D(r) = \min\left(conf_D(M \to c_a), conf_D(M \to c_b)\right) \geq \tau \wedge$$
$$sup_D(\{c_a, c_b\}) \leq \theta$$

間接相関ルール $r = M \to \{c_a, c_b\}$ は，c_a と c_b はそれぞれメディエータ（条件）M と強い関係があるにもかかわらず，c_a と c_b の間には共起の関係がないという状況を捉えており，条件 M の元で c_a と c_b が同じ役割を果たしている，もしくはライバルの関係にあることを表していると考えられる．本研究ではこの考え方をレシピデータベースに適用し，代替可能食材発見の基礎として，「食材集合 M の元での代替可能食材の対 c_a と c_b」を表す間接相関ルールを抽出する．

2.2 間接相関ルールのランキング

間接相関ルールの全体集合を $\mathcal{IAR}$ とする．2 つの間接相関ルール $r_1 = M^1 \to \{c_a^1, c_b^1\}, r_2 = M^2 \to \{c_a^2, c_b^2\} \in \mathcal{IAR}$ に対し，r_1 の方が r_2 より望ましい場合に真となる関数を $\phi(r_1, r_2)$ とする．このとき $\mathcal{IAR}$ 中の ϕ に関する間接相関ルール $r = M \to \{c_a, c_b\}$ の順位を $rank(r, \mathcal{IAR}, \phi) = |\{r' \in \mathcal{IAR} \mid \phi(r', r)\}| + 1$ と定義する．

本研究では，種々の観点から関数 ϕ を具体化することで，目的に応じて様々な間接相関ルールのランキングを実現する．

2.2.1 調理手順数の削減量に基づくランキング

食材代替えの一つの目的として，調理手順の簡略化が考えられる．この目的を反映し，食材を変更することで調理手順数が大きく削減できる（変わる）ルールを優先する関数 ϕ_{step} を提案する．ルール $X \to y$ の支持集合での平均手順数は，$S(X \to y) = \sum_{\langle id, I, S \rangle \in occ_D(X \cup \{y\})} |S| \,/\, |occ_D(X \cup \{y\})|$ と計算できる．これを利用し，間接相関ルール r の調理手順削減量を $DS(r) = |S(M \to c_a) - S(M \to c_b)|$ と定義する．また関数 $\phi_{step}(r_1, r_2)$ を $DS(r_1) > DS(r_2)$ のとき真となる関数とする．

2.2.2 食材の多様性に基づくランキング

多様な食材を利用した方が望ましいという考えに基づき，間接相関ルールの評価基準として，食材間の非類似性を利用する．本研究では，(1) 食材オントロジー[5)] による非類似性 ont 及び (2) 共起に基づく非類似性 cos の 2 つを考える．またそれぞれにおいて，(1) 代替食材（c_a と c_b）そのものの非類似性と (2) 食材を置き換えることで得られる多様性を考え，計 4 つの尺度を導入する．

食材間の非類似性を d としたとき，レシピ R における食材集合 I_R の食材 a に対する多様性を

$$V_d(I_R, a) = \sum_{x \in I_R \setminus \{a\}} d(x, a) \,/\, |I_R \setminus \{a\}|$$

と定義する．これを利用し，間接相関ルール r において食材 c_a を c_b に置き換えたときに得られる多様性の差分を

$$DV_d(r, c_a, c_b) = \frac{\sum_{R \in occ_D(M \to c_a)} V_d(I_R \setminus \{c_a\}, c_b) - V_d(I_R, c_a)}{|occ_D(M \to c_a)|}$$

と計算する．また r による得られる多様性を

$$DV_d(r) = \max\left(DV_d(r, c_a, c_b),\, DV_d(r, c_b, c_a)\right)$$

と定義する．以上を用い，$\phi_{ont}(r_1, r_2)$ と $\phi_{cos}(r_1, r_2)$ をそれぞれ $ont(c_a^1, c_b^1) > ont(c_a^2, c_b^2)$，$cos(c_a^1, c_b^1) > cos(c_a^2, c_b^2)$ のときに真となる関数，また ϕ_{ont}^{all} と ϕ_{cos}^{all} をそれぞれ $DV_{ont}(r_1) > DV_{ont}(r_2)$，$DV_{cos}(r_1) > DV_{cos}(r_2)$ のときに真となる関数とする．

2.2.3 食材の一般性・希少性に基づくランキング

食材の一般性・希少性は，代替可能食材を考える際の重要な要素である．間接相関ルール r に対し，c_a, c_b の出現頻度に基づく一般性の差を $DP(r) = |\, sup_D(\{c_a\}) - sup_D(\{c_b\})\,|$ とする．関数 $\phi_{pop}(r_1, r_2)$ を $DP(r_1) > DP(r_2)$ のときに真となる関数と定義する．

Finding replacable ingredients by indirect association rules
by Mika Kawaguchi and Tomonobu Ozaki (College of Humanities and Sciences, Nihon University)

3. 評価実験

提案手法の有効性を確認するため，クックパッドデータセット ☆ 中のタイトルが「オムライス」または「ハンバーグ」で終わるレシピを対象に評価実験を行った．表 1 にデータセットの基本情報を示す．なお，手作業も含め，食材の表記ゆれに対する処理を行っている．また食材オントロジーに関しては，文献[5] を基に独自に作成した．

表 1 データセットの概要

データ名	レシピ数	食材種数	食材数平均	手順数平均
オムライス	4535	2275	11.8	6.5
ハンバーグ	14277	4530	12.5	6.2

3.1 間接相関ルールの抽出

$\sigma = 0.005$ とし，τ, θ を変化させながら，間接相関ルールを抽出した．得られたルールの例として，オムライスにおける $\{$ 牛乳, バター, ゴハン, タマゴ $\} \rightarrow \{$ トマト, ネギ $\}$ などがあげられる．また，オムライスにおいて「バジルとパセリを置換する」といった細かい置換に関するルールや，ハンバーグにおいて「ソースとウスターソースを置換する」といった食材の特殊化（具体化）に関するルールも抽出された．得られたルール集合の基本情報を表 2 に示す．表より，設定によっては得られるルール数が少なくなく，何らかの基準で順位付けすることが必要であることが分かる．

表 2 得られた間接相関ルールの概要

	オムライス				ハンバーグ			
τ	θ	$\mathcal{IAR}$	$\#M$	$\#C$	θ	$\mathcal{IAR}$	$\#M$	$\#C$
0.1	0.05	61688	9184	1579	0.05	155265	23889	365
0.1	0.005	3300	176	680	0.01	172	138	17
0.3	0.05	61688	9184	1579	0.05	155265	23889	365
0.3	0.005	3300	176	680	0.01	172	138	17
0.5	0.05	32534	1291	1546	0.05	2049	1483	86
0.5	0.005	3215	108	675	0.01	5	5	1

※ $\sigma = 0.005$, $\#M, \#C$ はそれぞれ M と $\{c_a, c_b\}$ の異なり数

3.2 ランキング結果の比較

得られた間接相関ルール，オムライスデータ 3300 件（$\tau = 0.1, \theta = 0.005$）とハンバーグデータ 2049 件（$\tau = 0.5, \theta = 0.05$）を対象に，提案した関数によるランキングを行い，その違いをスピアマンの順位相関を用いて評価した．また，ランキング上位の類似性を確認するため，各ランキング上位 50 位タイまでのルール集合に対する Jaccard 類似度を計算した．結果を表 3 と表 4 に示す．これらの表中における ϕ_s と ϕ_c は，それぞれ支持度，確信度を利用した場合に相当し，$sup_D(r_1) > sup_D(r_2)$ と $conf_D(r_1) > conf_D(r_2)$ のとき真となる関数である．

ハンバーグデータでは，ϕ_{pop} と ϕ_{cos} の間に強い正の関係が，ϕ_{ont} と ϕ_{cos} 及び ϕ^{all}_{ont} の間に強い負の関係が認められた．また ϕ^{all}_{ont} は，ϕ_{pop} と ϕ_{cos} に正の相関関係が認められるが，Jaccard 類似度は高くなく，ランキング全体としては相関があるが，上位のルールに関しては類似性が低いという結果となった．一方オムライスデータでは，ϕ_{cos} と ϕ^{all}_{cos} に相関関係が認められるものの，それ以外の基準の間にほとんど相関がないことが分かった．

☆ http://www.nii.ac.jp/dsc/idr/cookpad/cookpad.html

表 3 ランキング間の順位相関：オムライス（左下），ハンバーグ（右上）

	ϕ_{pop}	ϕ_{step}	ϕ_{cos}	ϕ^{all}_{cos}	ϕ_{ont}	ϕ^{all}_{ont}	ϕ_s	ϕ_c
ϕ_{pop}		0.00	0.76	-0.02	-0.42	0.44	0.17	0.19
ϕ_{step}	0.03		0.02	0.09	0.10	-0.11	-0.06	-0.11
ϕ_{cos}	0.08	0.17		0.13	-0.63	0.45	0.25	0.24
ϕ^{all}_{cos}	0.03	0.02	0.42		0.19	-0.21	0.08	-0.14
ϕ_{ont}	0.18	-0.10	-0.12	-0.14		-0.62	-0.11	-0.25
ϕ^{all}_{ont}	0.00	0.06	0.12	-0.04	-0.30		0.12	0.29
ϕ_s	-0.02	-0.12	-0.01	-0.02	0.00	-0.02		-0.08
ϕ_c	-0.19	0.13	-0.16	-0.20	-0.13	0.11	0.01	

表 4 ランキング間の Jaccard 類似度（上位 50 位タイ）：オムライス（左下），ハンバーグ（右上）

	ϕ_{pop}	ϕ_{step}	ϕ_{cos}	ϕ^{all}_{cos}	ϕ_{ont}	ϕ^{all}_{ont}	ϕ_s	ϕ_c
ϕ_{pop}		0.00	0.74	0.10	0.00	0.01	0.08	0.00
ϕ_{step}	0.00		0.00	0.05	0.01	0.02	0.00	0.13
ϕ_{cos}	0.09	0.01		0.00	0.00	0.00	0.09	0.00
ϕ^{all}_{cos}	0.02	0.00	0.19		0.00	0.04	0.00	0.02
ϕ_{ont}	0.00	0.00	0.00	0.00		0.00	0.00	0.08
ϕ^{all}_{ont}	0.00	0.02	0.00	0.00	0.00		0.00	0.00
ϕ_s	0.03	0.00	0.00	0.00	0.00	0.00		0.00
ϕ_c	0.00	0.02	0.01	0.01	0.02	0.02	0.02	

以上の結果から，支持度や確信度も含め，全体として各ランキング間での順位に関する相関が小さく，また Jaccard 類似度も低いことから，本研究の目的である多様なルールのランキングが実現されたと考えられる．

4. おわりに

本研究では，代替可能食材を表す間接相関ルールの評価関数を複数提案し，種々の目的に応じたランキングを実現した．今後の課題としては，食材の重要度や調味料との相性を考慮した代替食材の提案や，評価関数の合成によるランキングの多様化・個人化などがあげられる．

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」を利用した．

参考文献

1) 志土地由香, 井手一郎, 高橋友和, 村瀬洋: 料理レシピマイニングによる代替可能食材の発見, 電子情報通信学会論文誌, J94-A(7), pp.532-535, 2011.
2) 野沢健人, 中岡義貴, 山本修平: word2vec を用いた代替食材の発見手法の提案, 信学技報, vol.114, no.204, DE2014-30, pp. 41-46, 2014.
3) 早川知道, 西川智佳, 榎優一, 鈴木涼, 伊藤孝行: アレルギー対応給食管理支援システムのための代替食材提案機能について, 第 29 回人工知能学会全国大会論文集, 2015.
4) P.-N. Tan, V. Kumar and J. Srivastava: Indirect Association : Mining Higher Order Dependencies in Data, *Proc. of the 4th PKDD*, pp.632-637, 2000.
5) 土居洋子, 辻田美穂, 難波英嗣, 竹澤寿幸, 角谷和俊: 料理レシピと特許データからの料理オントロジーの構築, 信学技報, vol.114, no.204, DE2014-30, pp.41-46, 2014.