O-026

レシピサイトにおける提供者と使用者の嗜好抽出と可視化
Extraction and Visualization of Preference of Recipe Providers and Users on Recipe Sharing Websites

江本 守† 大澤 幸生†
Mamoru Emoto Yukio Ohsawa

1. 初めに

現在、ユーザ投稿型レシピサイトには、大量のレシピが投稿されている。例えばレシピサイトの一つであるクックパッドにおいて、クッキーカテゴリーに属するレシピだけでも登録件数は 3000 件にのぼる[1]。レシピサイトにおいては、レシピの提供者と、実際に調理するレシピの使用者が存在する。レシピそのものは、提供者の嗜好、意図を反映した情報を含有していると考えられる。そのような観点に立ち、レシピ一つ一つから提供者の意図を読み取ることが可能であるという仮定の元で、KeyGraph による可視化を行った。また、各レシピのレビューも取得可能であるので、レビューもまた可視化を行い、レシピの使用者の嗜好を理解することを試みた。

2. KeyGraph

2.1. 概要

本研究においてレシピサイトの情報を可視化するにあたり、KeyGraph を使用した。KeyGraph は大澤らが提唱した非常に古い技術であり、当初は文書データからのキーワード抽出の為に提案されたが、現在では事象の共起関係可視化手法として適用されることが多い[2,3]。

2.2. KeyGraph の処理手順

D=[a1,a2,a4,a5,…,a10],[a4,a5,…],[a1,a2]…とデータ D が与えられたとする。a1,a2 などのデータを構成する最小単位をアイテムと呼ぶ。

2.2.1. データの洗浄

指定されたノイズアイテムの集合を D から削除する。

2.2.2. 島の抽出

頻度の高いアイテムを上位から一定(M1)個取り出す。次にその中で、共起度 c0(w1,w2)の高い上位 M2 対の 2 アイテム(w1,w2)を実線のリンクで結んで共起グラフ G を得る。次に、単連結のパスを切除することにより、強く結びついているアイテムの塊だけを残す。この段階で、含まれるアイテムのうち、どの 2 アイテムにもその間に実線リンクからなるパスが存在するようなアイテムの塊を島と呼ぶ。

2.2.3. 橋の抽出

データ D の中の全てのアイテム w について、2.2.2 で取り出した各島 g との共起度 c1(w,wg)を求め、この値が上位となる M3 対の w,g の間に橋があるとみなす。ただし wg は、w が g に含まれない場合は、D 中にある島 g 内のアイテム全てを wg というアイテム名で置き換えたものとし、w が g に含まれる場合は D 中にある島 g 内のアイテムのうちの w 以外を wg という語で置き換えたものとする。

2.2.4. ハブの候補抽出

データ D の中の全アイテム w について 2.2.3 の橋を介して w と呼ばれる全ての島 gi との共起度 c1(w,gi)を合成した式(1)の key(w)とする。c0(w1,w2), c1(w,gi)の計算法は KeyGraph の初期版以降、様々な目的に応じて多様な手法が開発されてきた。ここでは経験的に c0 としてリフト値、c1 としてジャカード係数を用いる。

if c1(w,gi)>0 となる i が 2 個以上:

$$key(w) = 1 - \prod_i (1 - c1(w, gi)/f_0)$$

else $key(w) = 0$ (1)

ただし f0 は正規化係数で、c0(w,gi)の最大値が 1 を超える場合に c0(w,gi)の最大値、それ以外は 1 に設定する。key(w)の上位 M4 個のアイテムの集合を K（キーアイテム集合）とする。キーアイテム集合 K の中から 2.2.2 の高頻度アイテムに含まれないものをハブと呼び、赤ノードで描く。

2.2.5. グラフ化

K 内の各アイテム w と、w の共起度が上位となる 2 つの島 g を点線で結ぶ。これが主な橋と島の間の繋がりを描画していることに当たる。また、g 内で w と最も共起度の高いアイテムを選んで w の間に点線を引く。

以上の手順に基づき、可視化を行った。

3. レシピデータ可視化

本研究にて対象としたのは、レシピ共有サイトクックパッドにて、'ゼラチン'というキーワードで検索取得した、29000 件程度のレシピと、そのレビューである。

- 各レシピをバスケット単位として入力
- polaris にて形態素解析
- KamisibaiKeyGraph による KeyGraph 出力

という手順に基づき可視化を行った。黒ノード数を 40、黒リンク数を 10、赤ノード数を 30 とした。

3.1. レシピ

ゼラチンレシピを可視化した図をに示す。各レシピをバスケット単位とし、タイトル、概要、使用材料、コツ・ポイント、レシピの生い立ちのデータを抽出した。

から、ゼラチンを使用したレシピは、分類すると

・夏に合う果汁を使用した清涼感のあるゼリー
・ヘルシー指向な料理レシピ
・ゼラチンを使用したスイーツ

のレシピが主であることが読み取ることが出来る。

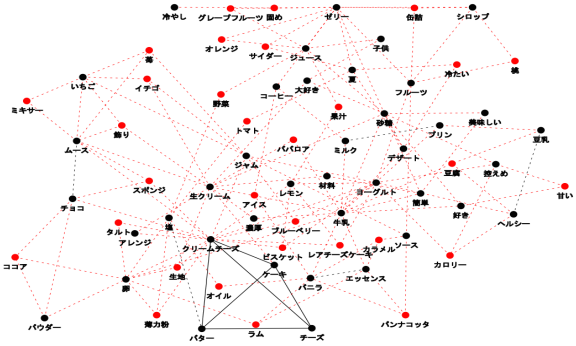

図 1 ゼラチンレシピ

また、黒ノードとして「簡単」、「子供」といったワードが

† 東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo

出現しており、子供が喜ぶ、手軽に作れるスイーツレシピを目的としてレシピを提供していることも読み取れる。

3.2. レシピレビュー

各レシピに対するレビューの可視化結果を図 2 に示す。

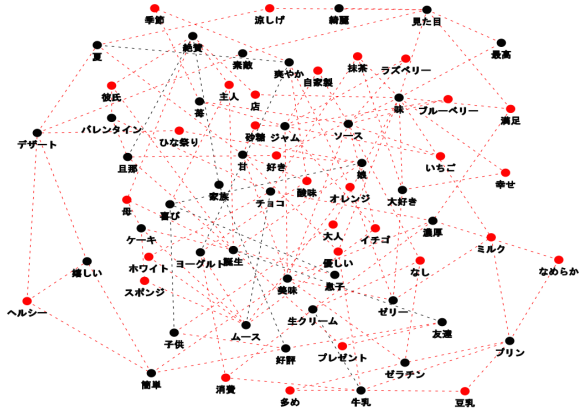

図 2 ゼラチンレシピレビュー

図 2 の結果から以下の様な事情が推測される。

- バレンタイン、ひな祭りのようなイベント時に、彼氏、夫、子供のために調理を行った。
- 夏にぴったりな清涼感のあるゼラチンレシピは、見た目の綺麗さから爽やかな印象を受けており、甘さと酸味のバランスがポイントとなっている。

4. 結果と考察

KeyGraph による可視化結果から得られた仮説推論の検証の為、Google トレンドを用いて、検索ワードを「ゼラチン」として検索された人気度動向について、2010 年 1 月から 2015 年 6 月までの変化を図 3 に示す。

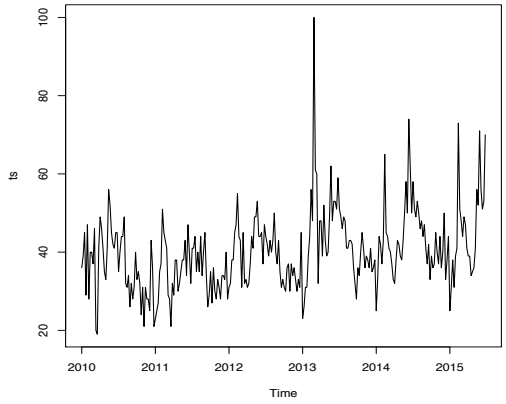

図 3 Google トレンド時系列データ

ただし人気度とは、検索ボリュームをシステムで正規化して処理した相対値のデータであり、総検索ボリューム、対象地域、設定期間などの要素で正規化されている。取得したデータの場合においては、設定期間におけるピーク値を 100 とした時に、他の年月ではそれと比べどの程度であったか比較出来るような正規化がなされており、過去の実検索回数に関わらず検索ワードのトレンド変化を知ることが可能である。

なお、図 3 のデータにおいては 2013 年 2 月 24 日から 2013 年 3 月 2 日の値がピーク値を示している。

次に、取得した時系列データの分解を、R を用いて行った結果を**図 4** に示す。観測値＝周期変動＋トレンド＋残差と分解した。トレンドは比較的に滑らかな長期変動、周期変動（季節成分）は時節とともに一定のリズムの周期的な変動をする成分である。

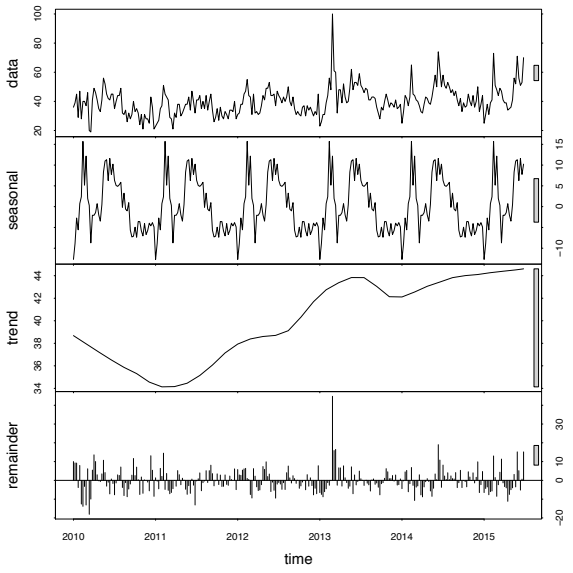

図 4 Google トレンドデータ分解

図 4 の seasonal 成分は周期変動成分を表し、2-3 月前半と、6-8 月あたりの数値は大きく、特徴的な山が形成されていることが読み取れる。この結果は、図 2 のゼラチンレシピのレビューを可視化した KeyGraph を読み取り推測した、ゼラチンレシピの使用者が「バレンタインのお菓子作りの為にレシピを利用した」、「ゼラチンを用いた清涼感のある料理を夏に作る為にレシピを利用した」という仮説と一致している。以上のことから、KeyGraph を用いたレシピレビューの可視化による、レシピ使用者の嗜好抽出は人気度の変動データと一致する。レシピデータの繋がりを KeyGraph により可視化することで、各レシピに使用者が求めている潜在的な嗜好を顕在化することが出来る為、ニーズに基づく商品展開の検討にも活用出来ると考える。今後の課題としては、レシピ市場における実際の分布に合う意見収集を行うことにより、高評価、低評価を公平に反映した検討を行うことを挙げたい。

謝辞

本研究において、解析対象となるレシピデータをクックパッド株式会社からご提供頂いた。記して感謝します。

参考文献

[1] レシピ検索 No.1/料理レシピ載せるならクックパッド,http://cookpad.com/

[2] 大澤幸生,ネルス E.ベンソン,谷内田正彦, ”KeyGraph:語の共起グラフの分割・統合によるキーワード抽出”, 電子情報通信学会論文誌,Vol.J82-D-1,No2.pp.381-400,1Feb,1999.

[3] 大澤幸生. チャンス発見のデータ分析:モデル化+可視化+コミュニケーション→シナリオ創発.東京電機大学出版局

[4] Google トレンド, https://www.google.co.jp/trends/