Vol.2016-NL-228 No.2
2016/9/29

料理レシピサービスにおける検索語の意味変化に関する分析

深澤祐援[†1,†2,a]　原島純[†2,b]

概要：我々の毎日の食事について, その傾向を分析するためにこれまで様々な研究が行われてきている. こうした研究の多くはアンケートやインタビューによる分析を行っている. これに対して, 本研究では料理レシピサービスであるクックパッド内に蓄積された検索ログによる分析を行なった. 具体的には, 検索ログの共起語ネットワークに対する表現学習によって各語に関する分散表現を獲得し, 2015 年 9 月を起点としてどれだけの意味変化が生じているのかを分析した. 分析の結果, 味覚表現に関する単語が最も意味変化が大きく, 加工食品に関する検索語の意味変化が最も小さいことなどがわかった.

1. 序論

我々の日常的な行動である食事について, その食生活の変化を捉えようとする研究は数多く行われてきている. 例えば, 仁藤[1]は家庭内でどのようなメニューが喫食されているのかを十年間に渡って調査している. また, 池田ら[2]は中学生の十年間における食生活変化をアンケート結果から分析し, その結果と健康状況の変化とを対応付けて調査している.

本研究では, 食生活の変化を分析するためにクックパッド[1]の検索ログを用いる. クックパッドは日本で最大の料理レシピサービスで, 2016 年 9 月の時点で 245 万品以上のレシピが検索できる. 同サービスには日々, 膨大な検索ログが蓄積されている. 検索ログは、先行研究で分析されてきたアンケートやインタビューとは異なるリソースとして食生活の変化を分析するのに利用できると考えられる.

クックパッドのデータを用いて食生活の変化を捉えようとした同様の研究として桐本ら[3]の研究が挙げられる. 同研究では, 「つくりましたフォトレポート(通称, つくれぽ)」と呼ばれるレシピに対するフィードバックの頻度を用いて分析を行なっている. 一方, 本研究ではクエリ中に現れる各単語の分散表現を用いて分析を行う. 具体的には, ユーザの検索ログにおける共起関係を用いて共起語ネットワークを構築し, ネットワーク表現学習によって分散表現を抽出する. そして, 得られた各単語の分散表現及びそれを用いた意味変化のデータについて, 時系列的に生じている変化を分析する.

2. 先行研究

ネットワークを解析する場合, データが巨大になるほど, その特性を視覚的に理解するのが困難になる. そこで, 近年, ネットワークの解析を容易にするため, ネットワーク表現学習と呼ばれる手法が提案されている. これは, 何らかの予測問題を解くことでネットワークのノードを分散表現として抽象化する手法である. 2014 年に Bryan ら[4]が提案した DeepWalk は従来の表現学習の手法に比べてラベル推定などのタスクで大きく向上した精度を発揮した. これに続いて LINE[5]や GraRep[6]などの DeepWalk 以上の精度を発揮する手法が提案されてきている.

DeepWalk の高い性能はそのベースである Word2Vec[7]の寄与によるところが大きい. Word2Vec はある単語から周囲の単語を予測するタスクを解くことで各単語の分散表現を獲得する手法である. DeepWalk ではまず, ネットワーク上をランダムウォークするエージェントを用意する. そして, エージェントが到着したノードのラベルを単語とみなして文章を生成する. 最後に, これらの文章に対して Word2Vec を用いて, ノードのラベルの分散表現を獲得する.

Word2Vec によって得られた分散表現を用いて単語の意味変化の分析を行っている研究として Vivek ら[8]のものがある. この研究ではまず Google Book N-gram などの文章データを利用して 1900 年から 2005 年まで 5 年ごとにおける単語の分散表現を Word2Vec によって獲得する. 得られた分散表現を同一空間内に写像して 1900 年の分散表現からの距離を計算する. そして, その距離を単語の意味変化を示す時系列データとする. さらに, その時系列データを用いて Mean Shift アルゴリズムを用いた変化点検出を行い, gay という単語がこの百年間で大きく意味が変化していることを指摘している.

本研究では以上の研究を踏まえて, DeepWalk を用いて一ヶ月ごとの検索ログの共起語ネットワークから単語の分散表現を獲得する. そして, それらを同一空間内に写像し, その距離を各単語の意味変化として捉える.

†1 東京大学
†2 クックパッド株式会社
a) fukasawa@css.t.u-tokyo.ac.jp
b) jun-harashima@cookpad.com

1. https://cookpad.com

2.本論文において記述する各手法は, R の{igraph}と python の{langchangetrack}を用いて実行した.

3.本研究では GNU Parallel によって並列処理を行った[9].

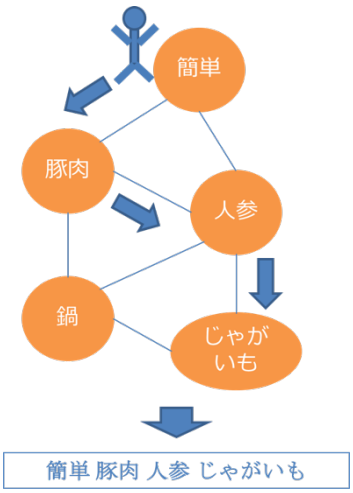

図 1. ランダムウォークによる文章生成

3. 分析手法

分析手法は大きく五つのフェーズに分けられる.[2,3]

(1) 共起語ネットワークの形成
(2) ノードのラベルの分散表現獲得
(3) 分散表現の整形
(4) 分散表現からの時系列データ抽出
(5) データに対する分析

以降, 各フェーズについて詳述する.

3. 1. 共起語ネットワークの形成

本研究で用いたデータはクックパッドの検索ログにおける単語間の共起関係を示すデータである. 本研究では 2015 年 9 月から 2016 年 6 月までのデータを使用した. 各月において, 各単語をネットワーク上のノードとし, 共起関係になっている単語についてネットワークのエッジを結ぶ. 各エッジの重みはその共起関係が出現した回数とする. このようにして, 月ごとに 1 個のネットワークを構築する. つまり、10 ヶ月で 10 個のネットワークを構築する.

3. 2. ノードのラベルの分散表現獲得

DeepWalk を使用してノードのラベルの分散表現を獲得する. まず, ネットワークからノードをランダムに選択して, そのノードにエージェントを配置する, そして, そのノードを開始位置としてエッジの重みを考慮したランダムウォークを行う. ランダムウォークするノード数は開始位置を含めて 10 とした. また, ランダムウォークは 100,000 回を行なった. 一回のランダムウォークで 10 個のノードが得られる. 各ノードにはラベルが付いているため, 一回のランダムウォークで 10 個の単語が得られる. これを文章とみなして, Word2Vec で単語の分散表現を獲得する. 本研究では分散表現の次元を 200 とした. 以上述べた処理について具体的な例を図 1 に示している.

3. 3. 分散表現の整形

ここまでで, 各単語について 10 個の分散表現が獲得される. しかし, 語の集合は毎月異なるため, 分散表現の特徴空間も毎月異なる. そこで, 先行研究[8]にならって, 10 ヶ月分のネットワークに共通する語で特徴空間を揃える. 全時点における共通語集合がKである場合において, t 時点のある単語wの分散表現についてマッピング前を$\theta_t'(w)$, マッピング後を$\theta_t(w)$とする. このとき式(1)で示されている最適化問題を解くことによって, 時点 t の分散表現を時点 t'の分散表現に合わせてマッピングする線形変換$W_{t\to t'}(w)$を求めることができる.

$$W_{t\to t'}(w) = \underset{W}{\operatorname{argmin}} \sum_{w_i \in K} \| \theta_t'(w_i)W - \theta_t(w_i) \|_2^2 \qquad (1)$$

3. 4. 分散表現からの時系列データ抽出

前節の線形変換で分散表現をマッピングした後, 0 時点目の分散表現(2015 年 9 月の分散表現)と t 時点目の分散表現(e.g., t = 2 の場合、2015 年 11 月の分散表現)の距離を算出することで、各単語の意味変化を表す時系列データを構築する. ある単語wについてt時点における距離λ_t(w)は, t時点においてマッピングされた後の分散表現行列を$\theta_t(w)$とした場合に式(2)として表現される.

$$\lambda_t(\mathrm{w}) = 1 - \frac{(\theta_t(w)W_{t\to 0}(w))^T\theta_0(w)}{\| \theta_t(w)W_{t\to 0}(w) \|_2 \| \theta_0(w) \|_2} \qquad (2)$$

3. 5. データに対する分析

以上の処理を通じて得られた単語の分散表現・時系列データに対する分析を行う. 分析は全部で 5 つのパートに分かれている.

(1) 単語のカテゴライズ
(2) 時系列データを用いた意味変化の分析
(3) 時系列データに対するピーク検出による分析
(4) 単語分散表現を用いた意味変化の分析
(5) 類似単語の変化に関する分析

以降, 各パートについて詳説する.

最初の分析として, クックパッド内で構築・運用されている単語分類辞書を用いて各単語を材料・メニュー・加工食品・目的・調味料・菓子・味覚表現・調理器具に分類する. 二つ目の分析として, 時系列データを用いた分析を行う. カテゴリごとに標準偏差を算出し,カテゴリによる意味変化の傾向を分析する. 三つめの分析として, 各カテゴリの季節性について分析するため, 時系列データに対してピーク検出による分析を行う. 本研究ではある

表 1. 各カテゴリの代表単語

カテゴリ名	代表単語
材料(ingredient)	豚肉, 玉ねぎ
メニュー(menu)	カレー, 親子丼
加工食品(processed_food)	カップラーメン
目的(purpose)	簡単, 夕飯
調味料(seasoning)	塩, オリーブオイル
菓子(sweets)	チョコ, ケーキ
味覚表現(taste)	ピリ辛, こってり
調理器具(tool)	オーブン, レンジ

単語wの時系列データにおいて, t時点の距離$\lambda_w(t)$は式(3)を満たした場合にピークとみなす.

$$\frac{d\lambda_w(t)}{dt} <= 0 \ and \ \frac{d\lambda_w(t-1)}{dt} >= 0 \quad \lambda_w(t) > 0.01 \tag{3}$$

四つ目の分析として, 分散表現を用いた分析を行う. まず, 分散表現空間上において単語の分布状況が時系列でどのように変化しているのかについて分析を行う. 最後に, 分散表現の性質を活かした分析として, カテゴリ毎の各単語についてコサイン類似度が近い 10 単語を月ごとの分散表現に基づいて抽出する. そして, それらの単語がどのカテゴリに属する単語なのか, その比率がどのように推移しているのかを分析する.

4. 分析結果・考察

4. 1. 単語のカテゴライズ

各カテゴリの代表的な単語を表 1 に示した. また,今回のデータで扱う全単語がどのカテゴリに所属するのか, その数を示したグラフを図 2 に, 各カテゴリに属する単語の一ヶ月間における平均出現回数を図 3 に示した. 複数カテゴリに所属する単語はその中で単語数が少ないカテゴリに合わせている. 全体的に単語数が少ないカテゴリは一つ一つの単語の平均出現回数が高く, 単語数が多いカテゴリは規模が大きいため出現回数が少ない単語も多く含んでいる. そのため, 平均出現回数が少なくなる傾向を持っている. 例えば, メニューカテゴリは「サラダ」「スープ」といった出現回数が多い単語を持ってはいるもののそれ以上にカテゴリに含まれる単語の範囲が広い. その分, 出現回数が少ない単語を数多く含んでしまうことが, 平均出現回数が低い理由だと思われる. これに対して, 調理器具カテゴリは「レンジ」「フライパン」など汎用的な語のみで構成されており, カテゴリとして含む単語の範囲が狭いことが, 単語数が少なく, 平均出現回数が多いことの理由だと考えられる.

4. 2. 時系列データを用いた意味変化の分析

一次元データの標準偏差を算出し, それをカテゴリごと

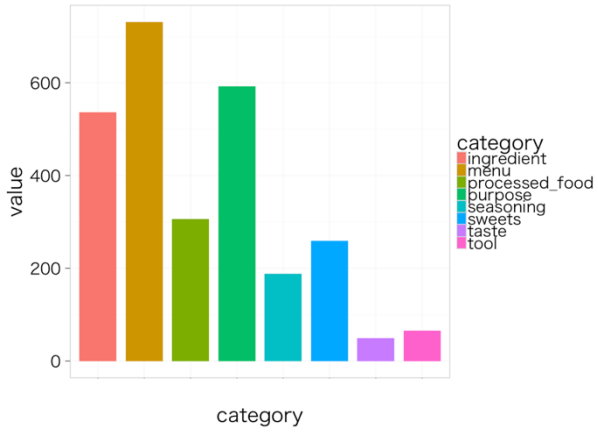

図 2. 各カテゴリに含まれる単語数

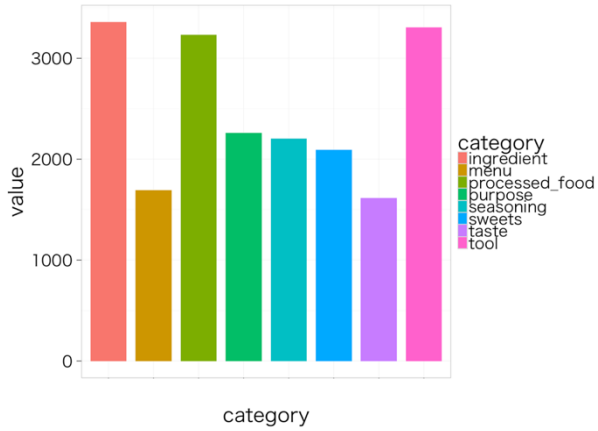

図 3. 単語の平均出現回数

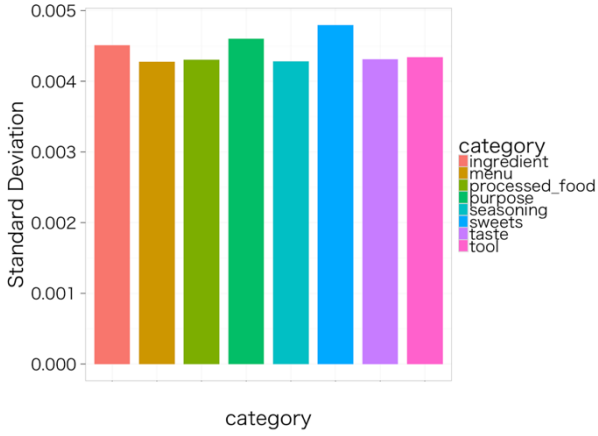

図 4. 時系列意味変化データの
カテゴリごとの平均標準偏差

に平均した. その結果を図 4 に示した. 最も標準偏差が大きいのは菓子カテゴリだった. 対して標準偏差が小さいのはメニューカテゴリだった. 具体例として両カテゴリで最も標準偏差が大きかった「クリスマスケーキ」と

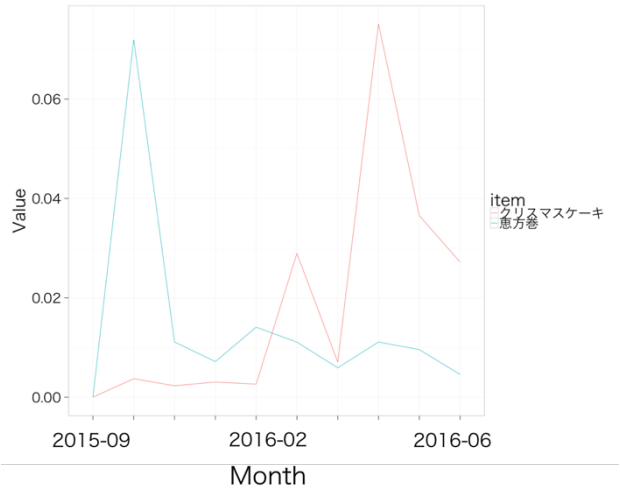

図 5. 具体例:クリスマスケーキ(菓子カテゴリ)と
恵方巻(メニューカテゴリ)

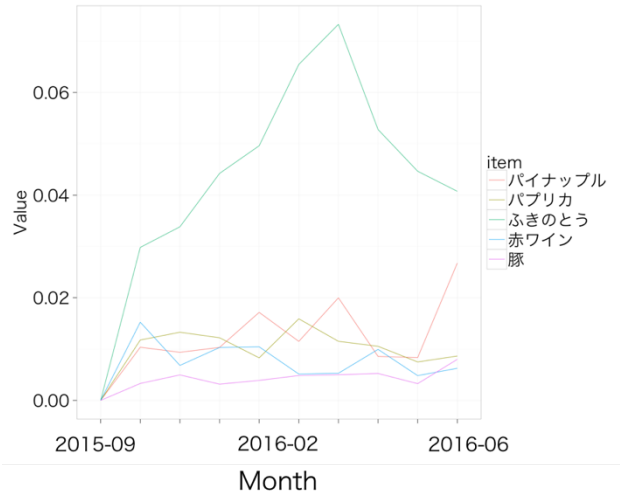

図 6. ピーク数に関する材料カテゴリの具体例

「恵方巻」のデータを図 5 で示した. 全体的にカテゴリ間における大きな差は見られなかった.

4. 3. 時系列データに対するピーク検出による分析

時系列データに対してピーク検出を行った結果について, まず, 具体例としてピーク数が 0,1,2,3,4 それぞれの場合の具体例を図 6 に示した. ピーク数 0 の具体例とした「豚」は季節性がなく汎用的な語であると言える. ピーク数 1 の具体例である「ふきのとう」は 3 月にピークが来ておりその月における季節性があることがわかる. ピーク数が 2, 3, 4 の「パプリカ」・「赤ワイン」・「パイナップル」は季節性が感じられるものもあるが, 意味変動のぶれが大きく出現する文脈が定まっていないとも言える.
このように, ピーク数によって各単語の季節性・意味変動の安定性を見ることができる. 本節ではこの点に着目し,ピーク数が 0, 1, 2, 3, 4 の単語がカテゴリ内において

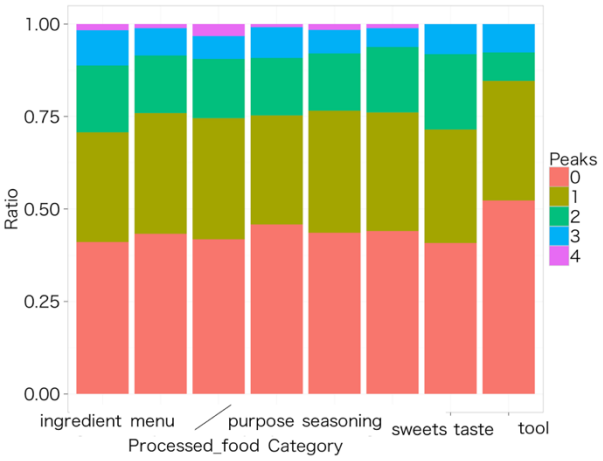

図 7. カテゴリごとのピーク数の分布

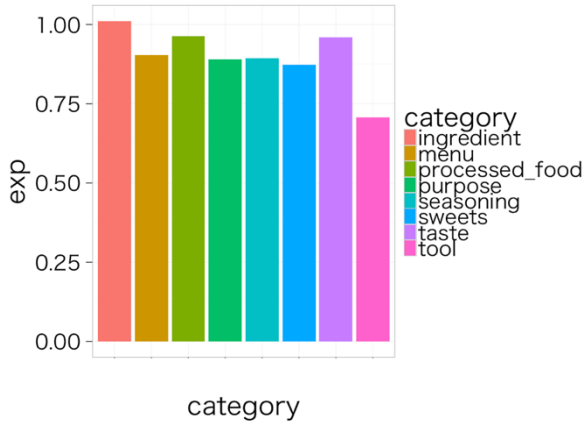

図 8. カテゴリごとのピーク数の期待値

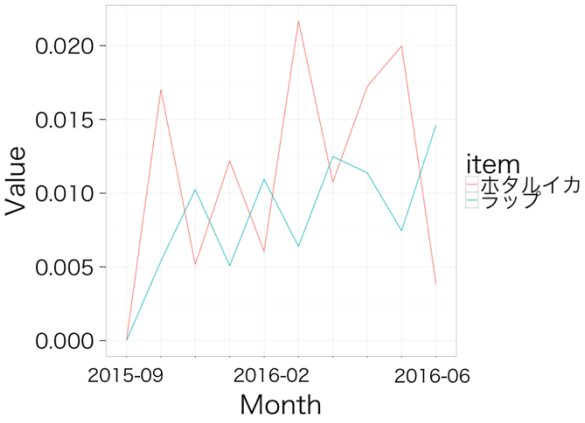

図 9. 具体例:ホタルイカ(材料カテゴリ)と
ラップ(調理器具カテゴリ)

どれくらいの割合で分布するのかを分析した. その結果を図 7 に示す. 図 7 において, 縦軸はそのカテゴリの中でピーク数$a(0 \leq a \leq 4)$の単語がどれくらいの割合存在しているのかを示している. また, 横軸はカテゴリ名を示している. 図 7 のピーク数の分布を確認してみると, カテゴリごとの違いはあまり見受けられないように見える.

この分布を用いて, カテゴリごとに一単語についてピークが幾つ検出できるのか, その期待値を算出した. 期待値はカテゴリごとに得られたピーク数の分布割合を用いる. カテゴリ x のピーク数の期待値をE_x, ピーク数$a(0 \leq a \leq 4)$の分布割合をD_aとし, 式(4)によって期待値を算出する. カテゴリごとに算出した結果を図 8 に示す. ピーク検出の期待値が最も高いのは材料カテゴリ, 最も低いのは調理器具カテゴリだった. 具体例として両カテゴリでピーク数が最も高い「ホタルイカ」と「ラップ」のデータを図 9 に示す. 材料カテゴリの期待値が高かったのはそれだけ材料カテゴリに季節性を有する単語が多く含まれていることの現れだと言える. 対して, 調理器具カテゴリについては, 季節性がある単語が少ないという知見が得られた.

$$E_x = \sum_{a=0}^{4} a * D_a \tag{4}$$

4.4. 単語分散表現を用いた意味変化の分析

今回の分析を通じて得られたデータの一つに, 月ごとの各単語に関する意味を表現した 200 次元の分散表現がある. これらの分散表現を t-SNE で二次元に圧縮した. 図 10 と図 11 は、2015 年 9 月と 2015 年 10 月のデータを可視化したものである. カテゴリごとにおおまかな分布を各色の円で囲んで示している.

両図を比較すると, カテゴリごとにデータの分布具合が変化していることがわかる. これは, 時間経過によって単語間の類似度が変化していることを示している.

これについてカテゴリごとにどれだけ意味変化が生じているのか, 各月ごとに共分散を算出した. またその推移に対して標準偏差を算出して, 変化の大きいカテゴリ・そうでないカテゴリについて分析を行った. その結果を図 12,13 に示す. これらの図を見ると加工食品カテゴリの変動性が最も小さく, 味覚表現カテゴリの変動性が最も大きいことがわかった.

4.2 及び本節で得られた意味変動に関する分析結果について, 特にメニューカテゴリに着目してみる. すると, 時系列一次元データの場合には標準偏差が比較的小さいという結果だったのに対して単語分散表現での分析においては共分散の変動性が味覚表現カテゴリに次いで高いことがわかる. つまり, 時系列的な意味における変化と単語同士の関係性変化は同一で捉えられるものではなく別の意味があるものだと捉えるべきだと考えられる. メニューカテゴリを例に挙げると, ベクトル空間では別のカテゴリ(味覚表現や目的など)の単語分布状況が変化しているため, それに依ってメニューカテゴリ自体の分布状況も変化している. ただし, 周囲の単語が変化しているわけ

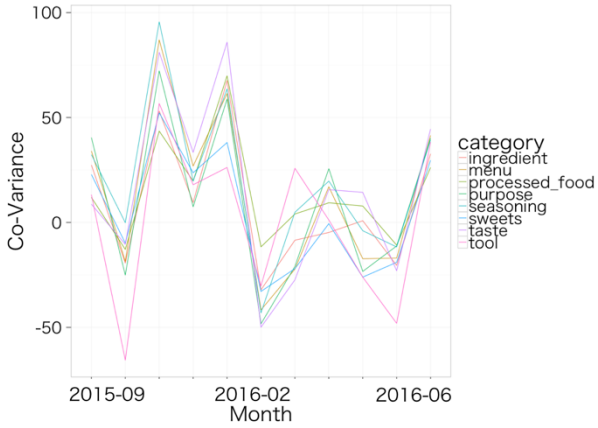

図 12. カテゴリごとの共分散の推移

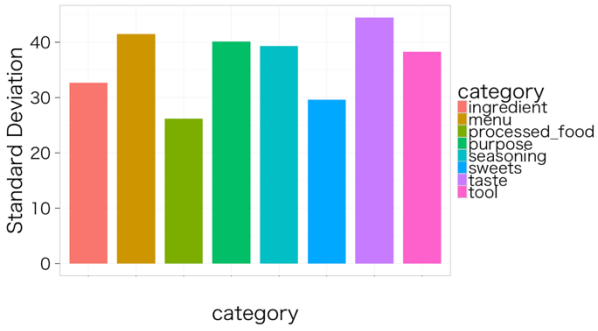

図 13.共分散推移の標準偏差

ではないのでt時点における 0 時点からの意味変化が小さく算出されやすい, と考えられる.

4.5. 類似単語の変化に関する分析

最後に, 類似単語の変化に関する分析を行う.各カテゴリの単語について, その単語と近い 10 単語がどのカテゴリに属しているのか, その割合の推移を図 14 に示した. 図 14 において, 八つのそれぞれのグラフは, 各上部に示したカテゴリ内の単語に関するグラフである.縦軸の'Ratio'はある月において, 上部に示すカテゴリ名の単語と類似する 10 単語がどのカテゴリに属するのか, その割合をカテゴリ全体で集計したものである. 例えば, 材料カテゴリのグラフであれば, 全体として最も類似単語として出現しやすい単語のカテゴリはメニューカテゴリであるが, 2016 年 2 月あたりで目的カテゴリが一度最も出現しやすいカテゴリになっていることがわかる.

以上を踏まえて結果を概観してみると, 材料・メニュー・加工食品・目的カテゴリのように単語数が多いカテゴリは類似単語に関する変化が小さいように見受けられる. 対して, 調味料・菓子・味覚表現・調理器具カテゴリ, 中でも味覚表現・調理器具カテゴリは特に類似単語の変化が大きいと思われる.

このような変化が見受けられた原因として, いくつか理

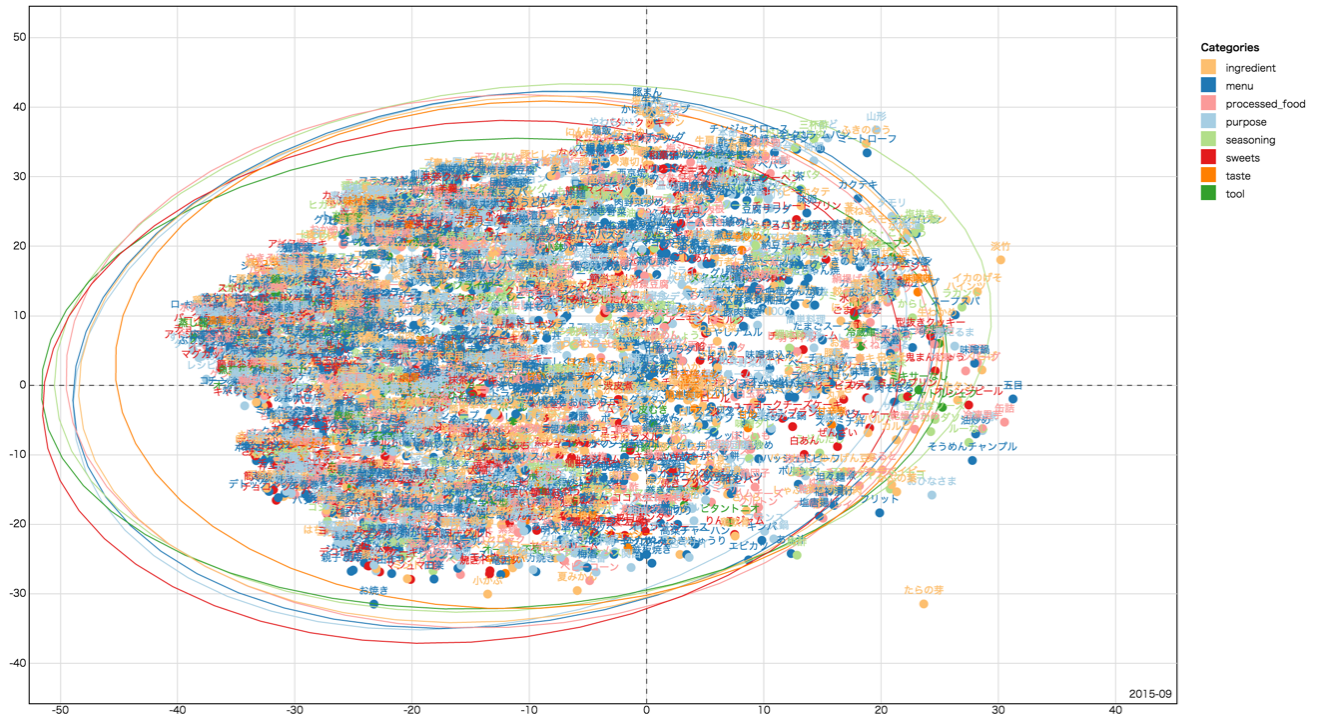

図 10.t-SNE で二次元に圧縮した 9 月のデータ

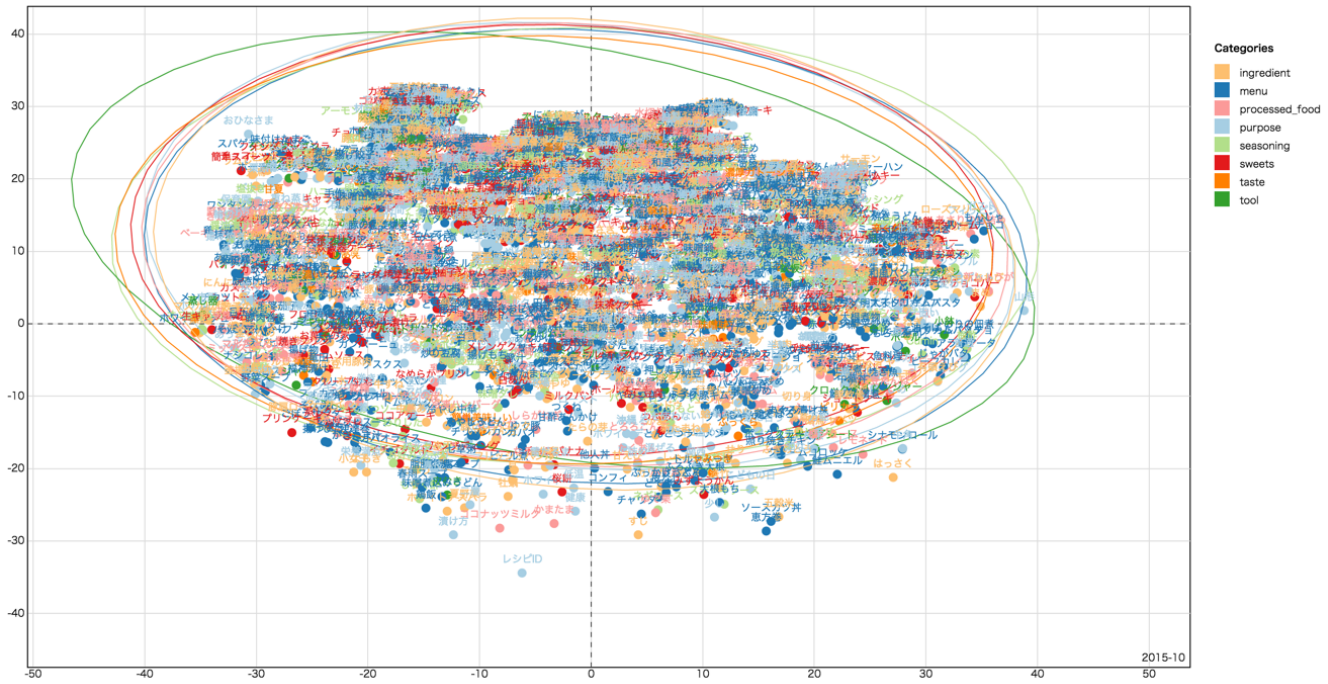

図 11.t-SNE で二次元に圧縮した 10 月のデータ

由が考えられる. 味覚表現カテゴリについては, 4.4 節の単語分散表現に関する分析において, 図 13 に示されているように, 分布状況が最も変化しやすいカテゴリであった. つまり, 他のカテゴリと比較して様々なカテゴリの単語と類似性を持ちやすいカテゴリであると言うことができる.

また, 調理器具カテゴリについては, 味覚表現カテゴリほど分布状況が変化しやすいわけではなく, むしろ全体的に変化が少ないカテゴリだと言える. 4.3 節のピーク検出による分析において, 図 8 に示されているように調理器具カテゴリは最もピーク数の期待値が低いカテゴリであった. また一方で, 図 2 と図 3 で示されているように, カ

図 14.類似 10 単語のカテゴリの変化の推移

テゴリ内の単語数は他カテゴリと比較して非常に少ないが,単語の平均出現回数が多いカテゴリである. 以上から, 調理器具カテゴリの単語は分布の変動性・季節性ともに小さいが, 一つ一つの単語は多様なカテゴリと組み合わされて出現するものが多い, と考えられる.

5. 結論

本研究ではクックパッドの検索ログを用いて, 単語の月ごとの分散表現を獲得し, それらを時系列的に分析した. 単語の意味変化に関する分析として, 時系列一次元データ・分散表現を用いた分析を行った. 時系列データを分析した結果, 菓子カテゴリでは意味変化が大きいことが, メニューカテゴリでは意味変化が小さいことがわかった. また、単語の分散表現を分析した結果, 味覚表現カテゴリの関係性変化が大きいことが, 加工食品カテゴリの関係性変化が小さいことがわかった.

続いて, 時系列一次元データに対してピーク数を検出し, カテゴリごとのピーク数の分布と期待値を算出した. 結果として材料カテゴリが最もピーク数の期待値が高く, 調理器具カテゴリが最も少ないという結果を得た. 最後に, 類似単語の変化に関する分析を行った. 結果, 材料・目的・メニュー・加工食品は類似単語の入れ替わりが小さく、調味料・菓子・味覚表現・調理器具は入れ替わりが大きいということがわかった.

続いて, 今後に向けた課題について二点述べる. 一点目として分類体系をさらに精緻化することが考えられる. 例えば, メニュー・目的のカテゴリに対して味覚表現・調理器具のカテゴリは単語数に大きく差がある. メニューや目的のカテゴリをさらに細分化することで, より詳細な分析が可能になる. 例えば, 目的カテゴリはイベントや季節, キャラクタ等のカテゴリに細分化できると考えている.

二点目としてデータを拡充することが考えられる. 本研究においてはデータ蓄積状況の関係から 10 ヶ月分のデータを用いて分析を行った. しかし, 一年分のデータを用いれば一年間を通してどれだけの意味変化が生まれたかを検出することが可能であると考えられる.

参考文献

[1] 仁藤清孝. 10 年間に見る食生活の変遷, 日本調理科学会誌, Vol.28 (1995) No.3, pp.196-204

[2] 池田順子, 河本直樹, 米山京子, 完岡市光. 中学生の 10 年間における食生活・生活状況と健康状況の推移, 日本

公衛誌, Vol.50 (2003) No.5, pp.420-434
[3] 桐本宙輝, 風間一洋. Cookpad のつくれぽ数の時間変動に基づく類似レシピ抽出法の提案法, DEIM Forum (2016) C7-1
[4] Bryan Perozzi, Rami Al-Rfou and Steven Skiena. DeepWalk: Online Learning of Social Representations, In Proceedings of the 24th International Conference on World Wide Web (2014), pp.625-635
[5] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan and Qiaozhu Mei. LINE:Large-scale Information NEtwork Embedding, In Proceedings of the 24th International Conference on World Wide Web (2014), pp.1067-1077
[6] Shaosheng Cao, Wei Lu and Qiongkai Xu. GraRep: Learning Graph Representations with Global Structural Information, In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (2015) ,pp.891-900
[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality, Advances in neural information processing systems 26 (2013), pp.3111-3119
[8] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi and Steven Skiena. Statistically Significant Detection of Linguistic Change, In Proceedings of the 24th International Conference on World Wide Web (2014), pp.625-635
[9] Tange (2011): GNU Parallel - The Command-Line Power Tool, ;login: The USENIX Magazine, February 2011:pp.42-47