オーラルセッション — フルペーパー

大規模ブログデータベースを用いた食の流行現状把握システム開発

情報システム研究機構 / ホットリンク 特任助教　株式会社ホットリンク 営業本部 第3営業部　株式会社ホットリンク 開発本部研究開発グループ・マネージャー
渡邊 隼史　小森 あゆみ　榊 剛史

要約

大規模データの解像度を活かし，これまでより客観的・即時的・網羅的な流行の把握や予測を行いたいというニーズが存在する。しかし，そのニーズに対して，完全に満足できるような手法・サービスはいまだ確立されていない。そこで，本研究では，そのニーズに応える第一歩のプロトタイプとして「食の流行や関心の現状分析・予測システム・サービス」の提案を行った。具体的には，本論文では，第一に，過去7年間・約30億記事のブログを解析することで，流行物の早期発見に対して，「キーワードの書き込み頻度の増加の持続性」への着目が有効である可能性を示唆した。次に，それを実行・運用可能な具体的なシステムとして実装した。特に，二点目については，ニーズが明確でなく確立された技術がない課題に対して，リソースが少ない中小企業で，どのように研究開発を進めたか経験も踏まえ報告する。

キーワード

ソーシャルメディアデータ，流行解析，大規模データ，研究開発

I.　背景及び既存研究

I. はじめに

近年，企業や様々な組織において，公式アカウント等を利用した広報・コミュニュケーション活動など，ソーシャルメディアのマーケティング活動での利用はもはや当たり前になりつつある。ソーシャルメディアのマーケティングでの利用法の一つにソーシャルリスニングがある。ソーシャルリスリングとは，ソーシャルメディアデータ(SNSデータ)を用いて社会や生活者・消費者の“思い”を調査する活動である。企業活動においては，消費者の生に近い声やリアルタイムの反応を知れる等の長所があるため，広告の効果測定，製品の感想調査，経営リスク発見等様々な用途に利用されている（大西，2015）。

一方，社会の期待に十分応えられていないソーシャルリスリングに関するニーズも存在する。その一つが流行把握と予測のニーズである。流行把握と予測のニーズとは，「世間全体での流行の現状や今後の行方を大規模データを用いて，これまでの手法より，客観的かつ精密に即時的に知りたい」というニーズである。このニーズは，例えば新商品開発部門においては「流行の兆しを6か月前までに知ることにより，商品開発の生産性をあげたい」というより具体的なニーズに結び付いてくる。

これまでもこのようなニーズや解決法も存在したが（コトラー，2014），その高度化の手段としてSNSデータが期待される理由は以下の特徴からである：（1）分野やジャンルに依存しない広い情報を含んでいる[広く社会全体の流行の全体像を知れる]（2）売上データ等の結果のデータでなく，動機や思いの情報を含んでおり，それを定量化できる可能性がある[流行に至る理由やその背景を知ることができる]（3）半リアルタイムに情報を収集できる[今の状態を知れる]。

そこで本稿では，SNSデータによる流行把握予測ニーズに応える第一歩のプロトタイプとして「食の流行や関心の現状分析システム・サービス」の提案を行う。ただし，II節以降で示す通り，現在の科学技術のレベルをもってしても，上にあげた期待に完全に答えることは非常に困難である。そこで，本論文では始めから技術的に完璧なものを目指すのでなく，技術や資源の制約のなか上記の理想やニーズに近づいていく現実的な（実行可能な）接近法を重視し

たい。

以下，第Ⅰ節では，類似研究について概観する。続く第Ⅱ節では，本研究の目的及び開発の進め方の方針について述べる。第Ⅲ節では，Ⅱ節で示した目的をどのような発想に基づき実現するか?今研究における流行の候補をとらえるための技術的な背景や原理について概観する。第Ⅳ節では，その原理をどのように運用し，顧客や社内で利用可能にしていくか?手法のシステム・サービス実装について記述する。第Ⅴ節では，Ⅳ節実装されたサービス・システムを様々な資源制約やニーズ不明確の状況化でどのように社会につなげていくか?その社会実装の過程について，私たちの経験を踏まえ記述する。最後の第Ⅵ節では，論文全体をまとめ，提案法の限界と今後の展望について述べる。なお本稿では読者の広汎性を想定し個別詳細より全体像や広い観点での記述を方針とした。

2. 本研究の類似サービス：

ここでは，主要な本研究の類似研究・サービスを国内外で2つあげておく。1つ目は，クックパッド社の「たべみる」である。これは同社のレシピ検索データをもとに日本全国の食に関する関心を前年同月比，季節，地域や性別等のデモグラフィック情報，共検索語の情報を用いて食のトレンド動向を調査できるWebサービスである（中村，2015）。2つ目はグーグル社によるグーグル検索の変化を用いた流行トレンド解析である。2015年に服飾，2016年に食品に関する簡単なレポートが発表されている（Zimmeret, 2016）。内容は検索ワードの過去3年程度の時間変化を6クラスタに分類しそれぞれのクラスタな代表的な食べ物を報告し，さらにそれらについて共検索語や検索場所等を報告している。上記の研究と比べた本研究の特徴は（1）類似研究が検索データを元にしているに対し，本研究がブログデータを元にしていること（検索は「興味」の意味がほとんどなのに対し，ブログは多様な意思情報を取り出せる可能性がある）（2）類似研究が現状把握ため，主に現状関心が高い事物の関心変化に着目するのに比べ，本研究は商品開発者向けに流行を早くとらえることを目指し開発が進められてきた背景があり，人々の関心がまだ低い事物の関心変化の扱いに特に注力していることがあげられる。

Ⅱ. 本研究開発の目的

1. サービス全体の開発の目的

我々は流行解析サービスの理想は，流行や関心・欲求の時間変化が（1）これまでどのような経緯を経て（2）現状どのような状態にあり（3）今後どのようになっているかを，客観的・俯瞰的・即時的かつ高解像度でとらえられること；加えて，それらを商品開発や広告出稿等マーケティング活動に役立つ形の情報で提供できること，と考えている。

本研究は，上にあげた理想の第一歩になるような「食の流行に関する流行や関心の変化の現状・予兆分析システム・サービス」の実行・運用可能なプロトタイプの構築とそれを背後で支える手法の開発を目指す。データは2006年11月からの30億国内ブログ記事データを用いる。なお，本研究において「食」に限定した理由は，第一は，具体的な社会ニーズがあったこと−本研究は，雑誌やTVにでないような流行を早く客観的に知りたいという大手の食品メーカーのニーズに端を発している—；第二は，「食」領域はブログデータと相性がよいこと—「食」は比較的ブログ記事数も多く，かつ，お店・場所・商品・素材・レシピ等その他の消費分野でもありうるような話題の広さがあるジャンルである—という二点からである。ただし，本研究の手法・フレームワークやデータは基本的に本質的な部分は「食」以外の分野でも適応可能な一般的な方法論をとっているため，様々なジャンルの流行や関心の変化全般に拡張することも比較的容易であると考えている。

2. 開発の制約と現実的な研究の進め方—人間と技術との関係—

一般に大規模データにおける流行に関する解析は，例えば「○○が××日位以内に流行する確率何%」というよ

うな非常に精度が高い形での定量的な予測が期待されることが多い。しかし，現状の最先端の科学技術をもってしても上記を機械的に信憑性高く実現することは困難である。例えば「予測」に関しては，主要な接近法のうち一つ目の物理学的接近法は（現象の理解に基づくホワイトボックス的な接近法，ニュートン力学が100年先の日食を予測するように予測能力が高い場合がある），モデルの元となるSNSデータの動力学研究は進みつつあるものの（高安，2012, Ishii, 2015, Watanabe, 2016），その全体像はほとんどあきらかになっておらず，現状ではまだ利用できる段階ではない。一方，もう一つの代表的接近法である統計科学・機械学習的な接近法（汎用的なブラックボックス的な接近法，現象の理解を伴わずデータ構造から予測）も，本質的に外挿問題である長期予測に不向きなことや，多様な時間変化パタンやニュースに起因するような例外的なはずれ値が多いなど学習を困難にするSNSデータ特性からかなりの困難が予測される。

そこで，本研究では上記の技術的困難さに対して現実的な接近法をとった。その接近法とは，機械ができることはできる限り機械が行い，機械の困難な部分は人間が行うという接近法である。そして，研究や開発がすすむごとに，人間が行っている部分を順次機械化していく（計算機による処理への置換え）という立場である。このようにすることで限定された開発コストとリスクのもとサービス化を意識しながら技術開発が可能になる。実際この方針でどのように研究開発が進められたかは第Ⅴ節で扱う。

3. 本研究における，人間と機械の仕事訳と技術的な目標

本稿では第一段階のプロトタイプとして，技術的と人間の以下のようにわけた（流行の予兆候補発見タスク）：

（1）機械は，日本中のブログに書かれる食べ物に関する語から，人知れず，長期的に関心が長期間増加し続けている言葉を探し，それを流行語候補語として提案する。

（2）人間は，その候補について流行りそうな事項の絞り込みやその根拠の把握（裏どり）をする。

より一般的な表現では，（1）機械は増加・減少傾向というブログの書き込み件数に関する単純な特徴量で絞り込み（例：流行の忘却を調べたい場合は，書き込み件数が徐々に減少している語を探す），（2）その後，深い解析―流行や関心の背後にはどんな社会状況があるのか?今後この流行はどのようになっていくか?等―は人間が行う；という仕訳である。

例えば，図表2bに示した「塩麹」の時系列は，TV等に取り上げられた2011年から2012年にかけて急速に書き込み件数を増やしている。一方，2008年からその時期になるまでの3年間も書き込み件数が徐々に増加していることも確認できる。本研究の予兆発見タスクは，このような急増以前の増加量がわずかだが継続的な増加を検出することでTV等によりメジャー化する以前に流行食語候補を見つけてこようという発想である。なお，本手法は長期の継続性に着目するため，トレンド（中長期な流行やニーズ，スタイルやファッションの変化）に重点をおき，現状では予測が難しいファッド（短期的な流行や熱狂）の早期検知には主眼を置いていない。

III. 機械的に流行の食に関する語の候補を絞り込む手法

本節では，前節で示した機械部分の目標：「日本中のブログに書かれる食べ物に関する語から，人知れず，長期的に関心が増加し続けている食に関する言葉を探す」をどのように実現するかを述べる。ここでは上記のタスクを以下の2つのサブタスクに分解して考える。

Step 1: 食に関するキーワード辞書生成（食語発見タスク）

Step 2: 食に関するキーワードの件数時系列なから徐々に件数が増加している語［関心が増えている語］を流行語候補として抽出・順序づけして出力する（絞り込み・流行候補語ランキングタスク）

以下では，上記の2つのタスクをそれぞれについて記述

する。ただし，紙数の都合と論文誌の性質上，ここでは概要だけを述べ技術的詳細については今後発表予定の情報工学の専門誌に譲りたい。

I. Step I: 食に関するキーワード辞書生成―ブログデータを用いた食べ物に関する新語発見―

食べ物に関する流行やニーズを知るためには，まず，この世の中にどのような食べ物に関する言葉（「食語」）があるか知っておく必要がある。しかも，「食語」は固定的でなく，随時新たに生まれてくるので，常に情報を更新しつづけることが要請される。「食語」の辞書を得る一番簡易的な方法は，ウィキペディア等のカテゴリつきの大規模辞書を利用する方法があり，本研究でもこの方法は併用している。しかし，この既存辞書を用いる方法のみでは，辞書に新語が登録されるまで時間がかかることや，関心が増えていても登録されない単語がある等，流行の把握タスクに対して十分でなかったため，新たに以下に示す「ブログを用いた新語候補発見法」を開発した。

(1) Step 1-1: 食新語候補発見法―ブログの文書を単語に分解し，新語候補リストを作る―

ブログに書かれた文書から「食語」を探すためには，まず，ブログの文書を単語に分ける必要がある。一般に，文書の単語への分解は自然言語処理において「形態素解析」といわれるほぼ確立した手法やツールが存在する。しかし，今回の「食語」発見のタスクでは，「食語」特有の複合語が形態素解析で分解されてしまい一つの単語として認識できない問題があった。例えば標準的な形態素解析ツール（MeCab+ipadic）用いると「うまい棒」は「うまい｜棒」に，「炊きこみご飯」は，「炊き｜こみ｜ごはん」に，また「味噌｜カレー｜牛乳｜ラーメン」，「ちょい｜飲み」，「裏｜難波」，「孤独｜の｜グルメ」と過剰に分解されてしまう。

この複合語の問題は，形態素解析器は解析器に登録された単語辞書を元に単語を分解していることに起因する（奥村，2010）。そこで以下の解決策をとった。まず，（1）標準的な形態素解析器で単語に分解する。次に，（2）その分解された単語のうち「名詞」-「名詞」（例：牛乳カレー），「名詞」-「動詞連用形」（昼飲み），「副詞」-「動詞連用形」（ちょい飲み），それらの繰り返しや組み合わせ（例：味噌カレー牛乳ラーメン，炊き込みご飯），等の複合語作る品詞の並びをすべて結合し「単語候補文字列」として収集する。最後に（3）この収集した「単語候補文字列」のうち「吉川先輩歌うますぎ」や「外苑～明治神宮～外苑周回～青山1～豊川稲荷～四ツ谷～市ヶ谷～曙橋」のような多数の単語とは言えない文字列を除去する。除外条件は「3年連続で年間の総頻度が増加している語を除き除去する」とする。この除外条件を用いる理由は，単語でない文字列は一般的に一度や一時期や断続的にしか出現しないため除外され，また，「こんにちは」「ありがとう」のような新語でない一般化した単語も時間変化が少ないため除外され，大まかには新語候補のみ抽出されると期待されるからである。

(2) Step1-2: 食の新語候補発見法―単語リストから食べ物に関する語を選別する―

前のステップで取り出された新語候補は「食語」以外の語も含むため，そこから「食語」のみを選別する。本研究では，着目語が「食語」か「食語」でないかは，基本的には，着目語を含む文書中での着目語の前後の単語群をもとに機械学習（回帰分析）を用いて判定している。例えば，「昼飯に”ラーメン”を食べた。おいしかった」と「線形代数は”統計解析”においてよく用いられる数学である」を比較したとき，”ラーメン”は「昼飯」，「食べる」，「おいしい」と食べ物と共に使われそうな単語が周辺に多くあるので「食語」と判定し，一方，”統計解析”の周りには「線形代数」や「数学」等あまり食べ物と一緒に使われそうにない単語が多いため「食語」でないと判定している。なお「食語」のまわりに出現しやすい語は，1万語程度の「食語」と「非食語」のブログの約300万記事（キーワード「ブーム」を含むブログ記事集合）を用いて事前学習している（回帰のパラメータ推定）。ただし，実際の学習モデルでは，精度の向上のため上記のような周辺単語を直接用いるのでなく，周辺単語の意味情報を数学的に圧縮

したベクトル表現を用いて学習器を構成している。具体的には，潜在的意味解析（LSA）を用い周辺単語のベクトル化を行い，分類にはロジスティック回帰等を基本に，いくつかの分類器を組わせて多数決をとるような方法を用いている（佐藤，2015）。最後に，この機械的に得られた「新食語候補リスト」を人目で確認することで「新食語辞書」を得られる（人間確認は2000語程度で月一回1時間程度）。

なお，実運用では，上記の手法が大半を占めるが，網羅性を改善するため，『』に囲まれた単語を新語候補にする等いくつかの補助的な手法も組合わせている。

2. Step2 キーワード時系列の時間的変動の解析―人知れず継続的に伸びている時系列を探す―

次の課題は，これまでに得た「食語辞書」のうちに人間が流行解析のできる程度に流行候補語を絞り込むことである。まず，ブログの食語辞書に含まれる過去7年間の各単語の出現頻度の日次の時系列を取得する（口コミ@係長APIサービスを利用）。次に，この時系列データから人間が流行解析できる程度に流行語候補ワードを絞込み，順序をつけて出力する。この絞込みや順序づけに，時系列情報を元に計算できる，以下の長期，短期，関心量の3つの特徴量を利用した。3つの特徴量は（A）長期の指標：時系列の増加・減少の継続期間（B）短期指標：時系列の直近で1年での伸び率（C）認知・関心の指標：直近一か月の平均件数（関心の指標）とした。例えば，II節で示した「人知れず，長期的に関心が増加し続けている」に対して，特徴量（A）は，「長期的に関心が伸び続けているもの」に対応し，特徴量（C）が「人知れず（件数が小さい）」の特徴づけに対応する。特徴量（B）は文書には明示的に含まれていないが，例えば，同じ期間書き込み件数が継続して伸びている語について，直近の増加の勢いが大きいキーワードのほうを小さいキーワードより優先して候補順位をあげたいとき等に利用できる。以下，特徴量の計算方法を記述する。

(1) 前処理―特徴量を計算する前に―

特徴量を計算する前にまず前処理を行う。前処理の目的は，第一にブログの全数など利用する既存のブログ収集システムに依存する量やシステムエラーを吸収すること，第二は，ニュースによるはずれ値や強い季節性等長期トレンドを計算するためにかく乱要因（計算を狂わせる要因）を除去することである。具体的に，前処理は，(i)ブロガー数増減効果（収集ブログ数変化）の除去[ブログ全数で除算し使用率相当量に](ii)異常値の除去および補間(iii)短期効果及びニュースはずれ値の除去[刈込み30日移動平均](iv)季節性の除去，という4段階をとっている。

(2) 特徴量の計算

(A) 長期の指標（増加・減少の継続性）

まず，一つ目の時間の長期変化の指標「増加・減少の継続性」について述べる。この「増加・減少の継続性」は本研究の流行語候補検出のメイン指標であり，本研究の最大の特徴でもある（多くのSNSデータを用いた時系列的方法は短期的のスパイク的な増加―忘れられやすい―に着目するのに対して）。この継続性は，基本的に「何か月間書き込み数が連続的に上昇しているか」という「連続上昇月数」で測かる。例えば，「18か月連続上昇（下降）」等である。ただし，これのみだと，同じ18か月連続上昇でも，毎年1%増加してきたか，3%増加してきたかの，増加の大きさの相違が区別できないため，実際の特徴量には月次上昇率の重みを加味した「重み付き上昇月数=連続上昇月数×上昇期間中の月次成長率の中央値（確率化）」を用いている。また，減少については，負の増加として表現し，連続期間にマイナスをつけて表現する。なお，現実の特徴量計算では，時系列のランダム成分により連続上昇計算がかく乱されるため，それをランダム拡散モデルの方程式（Watanabe, 2016）で補正している。また，重みに当る月次上昇率は（B）で述べる方法で確率標準化を行っている。

(B) 短期の指標（確率化年間増加率）

基本的には直近1年の書き込み増加率(今年の書き込み数÷前年の平均書き込み数）を用いる。しかし，単純に年間増加率だけで単語をランキングすると上位は平均書

き込み数が小さい単語ばかり現れるという問題が発生する。その理由は，同じ倍率の成長でも，書き込み数が少ない単語より大きい単語のほうが起こりにくいことによる。例えば，前年の平均的書き込み件数1件／日程度の単語が今年の書きこみ数が2件／日になることに比べ，同じ増加率2倍でも，前年の書き込み数1000件／日の単語が今年に2000件／日になることははるかに起こりくい。そこで，そのバイアスの補正法として，私たちは確率による年間増加率の標準化を採用した。具体的にいうと，前年の平均書き込み数 x(t-1) に対応する観測した成長率 r(t)=x(t)/x(t-1) の価値（確率の逆数）―平均書き込み x(t-1) で条件付けした時の増加率 r(t) の累積条件付確率の絶対値 $|\log(P(r(t)|x(t-1)))|$ ―を指標とした（減少の場合は負符号として表現する）。こうすることで確率という書き込み件数に依存しない共通の尺度で年増加率を比較することができる。なお，この確率化に用いた確率分布は，データ解析より，形状は自由度1.3のt分布（平均1で，分布の幅は平均書き込み数 x(t-1) の -0.2 乗に比例させる）を用いている。この特性は平均と標準偏差の自明でないべき乗則など複雑系科学的に興味深い性質があるため2015年度日本物理学会秋季大会で発表されている。

(C) 認知・関心の指標（年間平均件数）

認知・関心の特徴量は，単純な直近1年の平均書き込み数を用いる。関心や認知が多い単語ほど書き込み数が多いと考えているからである。ただし，これはあくまで近似であることに注意されたい。件数と着目物の関心は完全には対応しない。例えば，同じ”シオコウジ”を表す語でも，「塩麹」と「しおこうじ」の表記では件数は異なる。この問題は，提案の新語発見は代表的な言い回しをとってくる傾向が高いため，多くの場合は回避できるが，やはり粗い近似であり，将来期には，共起語を用いた認知指標に置き換える必要があると考えている。

3. 結果

本分析法がどのような流行語候補リストを出力するかを概観する。結果の出力には，2007年11月～ 2015年9月までのデータを用いた。当時の全「食語」数は，19448語である。ランキングは，短期指標と長期指標の重み付き相乗平均（の対数）を用いた。具体的には，ランキング指標 =(1-Q)・log(短期指標)+Q・log(長期指標)とした。ここでのランキングでは，「短期指標」は前節で述べた「確率化年次増加率」，また，「長期指標」は「（確率化）重み付き連続上昇月数（重み付き連続上昇数を確率化したもの）」を用い，重みはQ=0.85を採用している（書き込み数が減少局面にあるときは，負の符号をつけた値として表現する）。なお，重みQは，解析のニーズ応じてランキング短期と長期の重みQを変更される。例えば，短期的に関心が増加している事項を特に拾いたい場合は，短期のみ（Q=0）とすればよい。

図表1左は，上記のパラメータでの流行候補リストの上位60語である。2015年9月現在での流行語候補に相当するものと解釈している。実際，当時流行が報道された「チアシード」，「スーパーフード」，「糖質制限」が確認できる（週刊女性6月30日号等）。また，2015年のデータではあるが，2016年中ごろにネット等で話題になった「水素水」（2016年4月23日Fisco等）や，現在広がりつつあると報道されている「ランチ会」（2015年8月28日産経ニュース）等も確認できる。ただし，食語発見のミスにより「よもぎ蒸し」「アルバムカフェ」等、食語でない語も含んでしまっていることも確認できる（なお実務上はこのようなミスもあえて残し食外流行の参考用に利用している）。

図表2は，流行の早期検出可能性を過去データで検証したものである。具体的には，過去2010年11月データを用い，図表1aと同様な条件に加え，2010年11月での年平均件数が少ない語のみ（30件／日以下）を出力したリストである。平均件数が少ないという条件は，第II節の「人知れず…」に対応する。表にあげた語の半数以上が2010年11月からその後1年以上件数がのびつづけていることがわかる（2010年11月から約1年以上件数が連続的に増加しつづけたもの下線で示されている）。また，図表2bと図表2c は，実際に「塩麹」と「糖質制限」の単語時系列を確認したものである。図中で網掛けしてい

る部分が2010年11月である。2つの単語共に，件数増加の前半で流行語候補として検出できていることが確認できる。なお，なぜこの表にあげた多くの単語が1年以上の増加し続けている理由を示唆しているのが図表2dである。この図より，長期間増加を続けてきた語ほど，次の期間に件数が増化する確率が高い傾向がみてとれる（大まかにいうと関心が伸びている語ほど未来に伸びやすい傾向がある）。ただし，この経験的事実の裏にどのような動力学的（現象的）な背景があるか不明であり，今後さらなる解析が必要である（この解析が本手法を「予測」につなげていくことが期待される）。

本手法は増加と同様に関心が徐々に（人知れず）減少している語も検出できる。図表1右は2015年11月の連続減少数が大きいものの順にランキングしたものであり，韓国料理や外食チェーン店が目立つ。実際に韓国関係の食べ物の売り上げの減少（POSデータ）は2015年9月1日の産経ニュースで報道されており，この報道と本手法の結果とは矛盾していない。

IV. システム・サービス構成と人間とのインターフェース

前節まで今研究の基本的な原理をのべた。しかし，原理があっても，実際に動かなければサービスや社会応用につなげられない。そこで，本節では，前節で示した方法のシステム・サービス実装やその運用について述べる。特にここでは人間とのインターフェースについて記述する（本研究は確認・検証など人間前提のシステムのため）。

システム実装・運用の概略は図表3bに示す（詳細は図内説明参照）。このシステムと人間（内部利用者）の接点に対応するのが，図表3aのWebインターフェース（「食の現状ボード」）である。システムが計算したランキングを表示する。このボードは人間が役割や力を発揮できるよう以下の特徴がある―（1）顧客のニーズ等に応じて営業担当者が試行錯誤しがならランキングルールをチューニング・新規拡張できる（2）レポート作業等で必要とされる人間の検証･確認を補助する―。具体的には，例えば，「詳細設定機能」を用いることで，顧客ごとに異なるニーズ―短期重視・長期重視，意外性重視・確実な流行重視等―に合わせて営業担当者が独自のランキングルールを試行錯誤（特徴量重みを調整）しながら構築でき，また，そのランキングの特性を「グラフ・小グラフ表示機能」を用いて生件数時系列データに戻りつつ簡易的に確認できる。さらに，そもそも顧客が要求する言葉が「食」でない等辞書に入っていない場合も「任意語検索機能」で着目した言葉がランキングで何位に相当するかを表示することが可能になっている。その他確認補助機能は図内説明を参照されたい。

最後にレポートについて述べる。レポートは営業担当者と顧客との接点の一つである。例えば，レポートでは，システムに示した候補から顧客の興味等に合わせてキーワードを絞り込み，その語ついて，ブログ本文や他のデータやニュースの情報等を用いて，なぜ件数が増加しているかどのように興味がもたれているか等の定性的な情報をつけることで，今後の展望や発展性を示唆する。レポートの例を図表3cに示した。

V. 社会実装

1. 研究開発方針

ここでは研究成果をどのように社会やサービスとして出していくか?私たちの経験を述べたい。特に，本研究は，これまでにあまりない新しい試みであるため（特に開発中は），どこにニーズがあり，それがどのようなもので，さらに，そのニーズの量（市場規模）もよくわからない，という問題；加えて，中小企業での開発のため開発営業リソースが共にあまり割けないというリソース上の制約もあった。

本研究では，上記のような困難を緩和するため，3つの研究開発の方針をとった：（1）研究開発は小さくはじめ，プロトタイプやサービス化等を随時進めニーズを確認しつつ，その規模を順次広げていく[市場を確認しながらすす

むことで，研究開発がニーズから大きく外れることや開発コストが大きくなることを防ぐ]（2）技術的に研究開発するのはできる限り本研究以外に使える汎用的な手法の開発をベースにする[失敗してもある程度は無駄にならないようにする]（3）初めから全て機械化しないで，できないところは人間の力を借りる[難しすぎるタスクに挑戦し開発全体が失敗リスクを防ぐ，人間の知見を開発に落とす]。以下は，実際に3つの方針の下，どのように研究成果を社会実装・サービス化に近づけてきたか私たちの経験を述べる。この節は我々の例で必ずしも開発の進め方の科学的な正解というわけではない。しかし，今後，マーケティングの大規模データ化がすすみ先端の技術研究とサービスを結び付ける重要さがさらに増しくることも鑑みて経験を共有したい。

（1）初期：コンサルサービス（開発：1人，利用者：営業1～2人）　出力：パワーポイント資料

本研究は大手食品メーカーの雑誌やTVのような流行の兆しをソーシャルメディアから知りたりというニーズを起点としている。本研究開発は，まずこのニーズに答えることから始まった。このときの開発規模は営業担当コンサルタントが1名（補助的にもう1名加わることもある），加えて，開発者が1名の実質2人であった。開発では，初期の開発期間が1か月と非常に短かったため，研究開発部分は既存の研究や開発物が応用しやすい「徐々に件数が増加する語をみつけること」のみに集中し，その他の部分は人間の力に任せることにした。具体的には以下のようなコンサルティングサービスになった：まず（1）機械により「食語リストから徐々に件数が増加している語のリスト」を作成する（2）そのリストの中から興味のあるものをクライアントが選んでいただき（3）その語がどのような振る舞いになっているかを定性的に人間が裏どりをし，それをパワーポイント資料として納品するという，形である。これにより，機械は大量の単語リストから候補を絞り込みに集中し，ほかの困難な部分を人間が行うことで開発コストを抑えつつある程度ニーズに答えるというひな形ができた。（なお，この時点では，食語リストは，Web辞書と直近数か月の「激ウマ」を含むブログ記事に含まれる単語から人間が探すという人間ベースの方法であった）。

（2）ボード化（開発:1人，利用者:営業グループ等）出力:ボート，レポートや一覧ファイル

その後，衣料関係や版権関係等から引き合いがありある程度のニーズがあることがわかったため，コスト等の削減を目指し手動で行っていたことの自動化を進めていった。具体的には，データ取得更新，リスト作成・表示部分を半自動化し，図表3abに示した流行の現状ボードシステムを開発した。これにより，技術的には，常に最新の状態を表示できること，系統的に過去のランキングの確認し比較できること等の手動で案件ごとに対応していたときに比べて可能なことが増えた。また，営業上も，これまでは主担当者のみがパワーポイント資料をもってニーズがあるところに出向いて説明することも多かったが，ボードという簡単にみせられるものができたため，主担当者以外も，他サービスの説明のついでに，情報提供としてプロトタイプのボードを顧客にみせることにより幅広い顧客に本研究を認知させることも成功した。それにより，研究開発としても，より広く感想やニーズ等の情報を得られるという利点があった。加えて，営業担当者らが随時更新していくボードをみて確認することにもつながり，それによりボードや手法の不充分さもわかり，その改良にもつながった。例えば，新語発見法は，営業担当者が独自に調べた語や顧客先等の要望により必要になった語がランキング辞書に入っていない不備に気づく，研究開発者は，その不備を埋めるための改良の繰り返すことで改善が進んでいった。その他にも，例えば情報源がブログデータだけと不安だという要望に対して，外部サービスであるgoogleトレンドと簡易な相互チェックする方法等の営業活動が起因となった様々な実践的な確認法をボードに加えていった。この際，一人での開発しシステム全体を一人が把握していることが要望を柔軟に取りいれることや素早い改良のための強みになった。

（3）研究と開発の分離（開発件数：研究開発グループ，営業：営業グループ）

現在，技術的な目途や運用の仕方に目途がついてきたため，ボードをきちんと専門の開発エンジニアの手で再開発

し直そうという方針になっている。この方針の意図は（1）専門の開発者によるコンピュータシステム構築によるシステムの安定化やデザインの洗練（2）開発運用作業からある程度解放されることで，研究開発者は方法論や基礎部分の研究に集中できる（3）プログラムコードを整理・合理化することで，食以外の分野への展開や機能拡張しやすくする（4）開発システムの研究者の属人生を排除し，研究手法の会社の資産化を図ること，である。そのため，現在では開発者向けに仕様の文書化を行い担当研究者以外でもある程度開発がすすめられるような準備を行っている。

2. 開発方針の運用上の困難さ

しかし，できるところから小さく開発し徐々に大きくしていく開発ポリシーには難しさもある。最大の問題は，必ずしも顧客の期待と技術的に可能なことが一致しないことである。例えば，一般に顧客の期待は「6か月後流行確率何%」を示してほしい等非常に高い場合が多いが，本稿で提示したシステムでは現状ではまだ不可能である。また，同じ流行リストでも食べ物の専門家と食べ物に興味がない人では，「当たり前」「流行っている」「流行ってない」の認識がかなり異なり，後者の期待には答えられても，前者にまで意外性をもってとられるような流行物を精度よく出すことは現状のシステムではまだ難しい。もし，期待度が高いまま受注等をしてしまうと期待に答えられずプロジェクトが炎上してしまう。そのため本開発手法においては，研究者は営業担当者に現状で技術的な可能ことをしっかり伝え，営業担当者は，それを把握し顧客の期待度を十分にコントロールすることが重要になってくる。

VI. まとめと展望

本論文では，ブログ大規模データベースを用いた現実に実行可能な食に関する流行の現状把握システムをどのように構築してきたかを示した。本研究の意義は（1）約30億記事の大規模ブログデータ・約2万語の食に関する語の網羅的に解析により，「件数の増加の継続性」という特徴量に着目することで流行事象の早期発見できる可能性を示唆したこと(図表2)（2）現実のマーケティングニーズと最先端科学知識や技術を結びつける研究開発フレームワークの例を構築したこと（II, IV ～ V節）にあると考えられる。特に，（2）については，本研究フレームワークの特色は，機械ができないことを人間が行いサービス化と開発をある程度同時に進めることにある。この特色は，ニーズや社会的要請を踏まえつつ技術的な可能なところから随時開発・改良していくことを可能にする（新技術は随時システムにプラグインされていく）。この方法は，一端基礎ができれば，大学等の基礎研究者からみれば技術や研究のわかりやすいショーケースになり（また切実な現場のニーズにさらされることで，研究者にない着想や目標・厳しい評価基準も得られる），ビジネスからみれば最先端技術を素早く利用するフレームワークになっている。このような長所は，大規模データ時代―技術の進歩・盛衰が早い情状況化で，人間・社会という一般に複雑かつ多様で扱いにくい課題を解くことが求められる―における先端学術と産業の連携に関する若干の示唆を含んでいるかもしれない（例えば，理系の基礎研究者である著書にとっては，技術中心に見てしまい，人間との連携により技術を超えたサービスを実現し，そこからさらに着想をえるという発想はあまりなかった）。

しかし，本稿で示したシステムは，あくまで流行解析予測・サービスとしての第一のプロトタイプであり，社会の高い期待―例えば，製品開発ならば「6か月後に流行るものを知りたい」や「食の専門家も知らないようなブームの兆しをしりたい等」―にきちんと答えられるレベルまでには至っていない。そこで，今後の課題として以下のものをあげ，随時技術的に可能なところから開発研究を進めていきたい。

（1）ブーム早期発見の客観的かつ厳密な精度評価法の確立―件数の増加予測率のような統計学観点，流行をどのような段階で発見できたかなど流行的観点，商品開発有用度など応用的観点での評価手法の確立（本手法の流行の早期発見可能性を説得力をもって示す）―

（2）流行・関心度の定量化―例：「マイブーム」「密かなブーム」[初期],「話題の」「流行の」[盛期],「ちょっと古い」「昔流行った」[衰退期]など件数でなく共起語の使用の変化による関心度・ブーム度の定式化（キーワード件数と社会関心が必ずしも一致しない問題や非専門家や専門家の認知の違いの問題の解消）―
（3）定量的流行予測法の確立―例：数100万語の日本語の単語時系列を網羅的，増加･減少パタンの明らかにし，それを本手法へ応用（6か月先の流行を見つけることに向けて）―
（4）流行の質や理由をより明確化する技術の確立―共起語の時間変化可視化による関心増加理由発見支援，デモグラフィック・趣味などブーム状態・書き込み傾向等の人を軸にした流行状態評価，イノベータ理論の精密化・イノベータ自動発見，単語になる前の関心増加の概念発見等（関心の増加の理由・状態がわかることでアクションによりつなげやすいサービスへ）―
（5）価値・欲求の長期変化の動力学の研究と汎用未来予測システムへの拡張―マーケティング分野で鍛えた技術を他分野へ応用―

謝辞

本研究は株式会社ホットリンクに関わる多くの方々との連携や助けからできたものです。ここでこれらの方々のうち論文の内容に特にかかわりがあった方に対する謝辞述べさせていただきます。まず，第一に横江淳次氏に感謝いたします。本研究の基盤である初期のコンサルティングサービスを構築や営業等，横江氏なしに本研究がはじまることはありえませんでした。次に，本研究のプロトタイプの利用等研究の進め方に関する助言をいただいた内山幸樹社長に感謝いたします。加えて，コンサルティング，営業，研究開発，サポート部門目線で様々な貴重な助言をいただいた，神子島隆仁氏，久保田暁氏，宮田洋毅氏，平野真理子氏，セーヨ・サンティー氏，水木栄氏，佐藤弘和氏，神田麻衣子氏，公平奈都美氏，管理開発部門の方々に感謝いたします。

参考文献

大西浩志（2015）「レビュー:ソーシャルメディアとマーケティング研究」（その2）『MJ 2015 WINTER』34(4) pp. 58-68。

フィリップコトラー他（2014）『コトラー＆ケラーのマーケティング・マネジメント 第12版』丸善出版。

中村 耕史（2015）『「少し先の未来」を予測する クックパッドのデータ分析力』日本実業出版社。

Zimmeret, O. et. al. (2016) "Food Trend 2016" http://think.storage.googleapis.com/docs/FoodTrends-2016.pdf.

高安美佐子他（2012）『ソーシャルメディアの経済物理学』日本評論社。

Watanabe, H. et.al (2016), "Statistical properties of fluctuations of time series representing the appearance of words in nationwide blog data and their applications", arXiv:1604.00762.

Ishii, A. et.al (2015), "Mathematical model for hit phenomena and its application to analyze popularity of weekly tv drama", arXiv:1501.00758.

奥村学（2010）『自然言語処理の基礎』コロナ社。

佐藤一誠 (2015) 『トピックモデルによる統計的潜在意味解析』コロナ社。

図表—1　提案手法出力するランキング

左：流行候補語。右：関心が長期的に落ちている語 (2015 年 9 月現在；ランキング方法の詳細は，III-3 節参照)

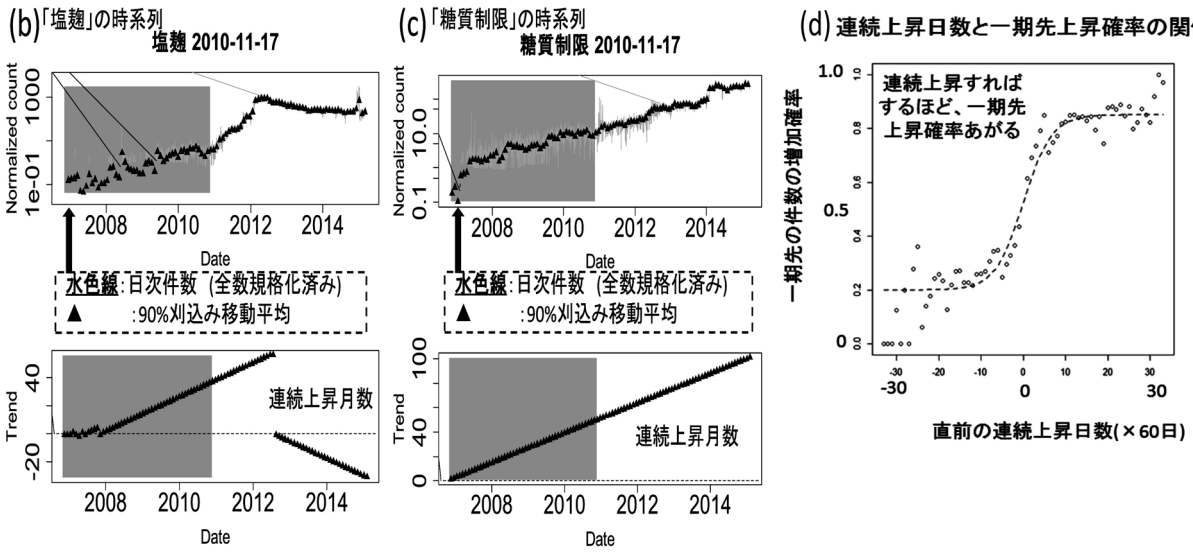
順位	ワード名	連続上昇月数	年増加率	年平均件数
1	SALON	106	1.3	1305
2	アヒージョ	105	1.3	158
3	アイシングクッキー	98	2.0	419
4	ホットペッパービューティー	80	1.7	196
5	ランチ会	106	1.3	454
6	いただき	106	1.1	31269
7	マカロンタワー	106	1.5	100
8	塩対応	73	1.6	44
9	台湾まぜそば	104	1.7	34
10	アルバムカフェ	58	1.3	53
11	うちサロン	78	1.2	102
12	熟成肉	76	1.6	40
13	メイソンジャー	38	5.6	107
14	よもぎ蒸し	91	1.3	146
15	もぎ蒸し	91	1.3	144
16	まぜそば	103	1.3	91
17	酵素シロップ	74	1.7	28
18	アイシングクッキー教室	51	2.8	121
19	メニュー	105	1.1	10899
20	多肉	105	1.2	427
21	プレ花嫁	70	1.5	44
22	グルテンフリー	87	1.6	50
23	いただきます	104	1.0	8992
24	ぐでたま	29	5.1	119
25	スーパーフード	59	2.7	55
26	ロースイーツ	95	1.8	18
27	飯テロ	103	1.6	44
28	多肉さん	103	1.2	41
29	ドリンク	104	1.0	39
30	アイシングクッキー作り	83	2.1	18
31	大豆ミート	97	1.2	20
32	糖度 17 度	33	6.8	53
33	コインパーキング	98	1.2	232
34	短期集中ダイエット	44	2.3	80
35	ガパオライス	72	1.4	34
36	飾り巻き寿司	74	1.2	33
37	出汁	96	1.2	1052
38	ファスティング	65	1.5	36
39	糖質	57	1.2	781
40	味変	77	1.2	35
41	水素水	59	1.5	171
42	置き換えダイエット食品	55	2.3	26
43	ハンドメイドマーケット	72	1.9	34
44	ランチ付	71	1.2	39
45	カルディ	92	1.1	103
46	糖質制限ダイエット	71	1.2	59
47	低糖質	59	1.4	89
48	昼飲み	96	1.4	22
49	フライングタイガー	29	1.3	51
50	チアシード	45	5.6	121
51	飲む点滴	94	1.3	20
52	脱毛クリーム	62	1.4	86
53	腸内環境	65	1.4	157
54	飲食代	62	1.5	182
55	酵素玄米	69	1.5	49
56	お茶会	57	1.4	761
57	低糖質パン	61	1.5	9.3
58	いきなりステーキ	27	2.1	25
59	ワンコインランチ	103	1.3	28
60	糖質制限し	65	1.5	20

順位	ワード名	連続上昇月数	年増加率
1	ショットバー	−39	0.9
2	ロールちゃん	−33	0.9
3	おそば屋さん	−28	1.0
4	タイ焼き	−28	1.0
5	ヤキソバ	−27	0.8
6	ドラ焼き	−27	0.8
7	食い過ぎ	−27	0.9
8	しょうが	−26	0.9
9	カルーアミルク	−25	1.0
10	食べるラー油	−25	0.8
11	バーテン	−25	1.0
12	リポビタンD	−25	0.9
13	土間土間	−24	0.6
14	和民	−24	0.7
15	ミンティア	−24	0.8
16	かっぱ寿司	−23	0.8
17	眠眠打破	−23	0.8
18	トマト鍋	−23	0.7
19	焼き肉	−23	0.9
20	ドーナッツ	−23	1.0
21	チオビタ	−23	0.9
22	ミスタードーナッツ	−22	0.9
23	ケンタ	−22	1.0
24	牛角	−22	0.7
25	バーミヤン	−22	0.8
26	白ココア	−22	0.5
27	ラタトゥーユ	−22	0.9
28	SUBWAY	−21	0.9
29	たこやき	−21	0.9
30	調理実習	−21	0.8
31	HENRI CHARPENTIER	−21	0.8
32	焼き肉屋	−21	1.0
33	マッコリ	−21	0.8
34	お好み焼き	−21	1.9
35	宅飲み	−21	0.9
36	シェーキーズ	−21	0.6
37	都路里	−21	0.9
38	マック	−21	0.7
39	魚民	−21	0.7
40	エビカツ	−20	0.8
41	やきそば	−20	0.9
42	韓国料理	−20	0.8
43	チヂミ	−20	0.9
44	ガスト	−20	0.8
45	飲みほし	−20	1.0
46	ひな野	−20	0.9
47	なっちゃん	−20	0.9
48	みずな	−20	0.9
49	吉牛	−20	0.9
50	ポムの樹	−20	0.8
51	ホットケーキ	−20	0.9
52	ポポラマーマ	−20	0.8
53	鍋パ	−20	0.8
54	ミルクティ	−19	0.8
55	スンドゥブ	−19	0.8
56	らでぃっしゅぼーや	−19	0.9
57	トッポギ	−19	0.8
58	つけめん	−19	0.9
59	甘党	−19	1.0
60	すた丼	−18	0.9

図表—2　過去データをもちいた流行の早期発見の検証

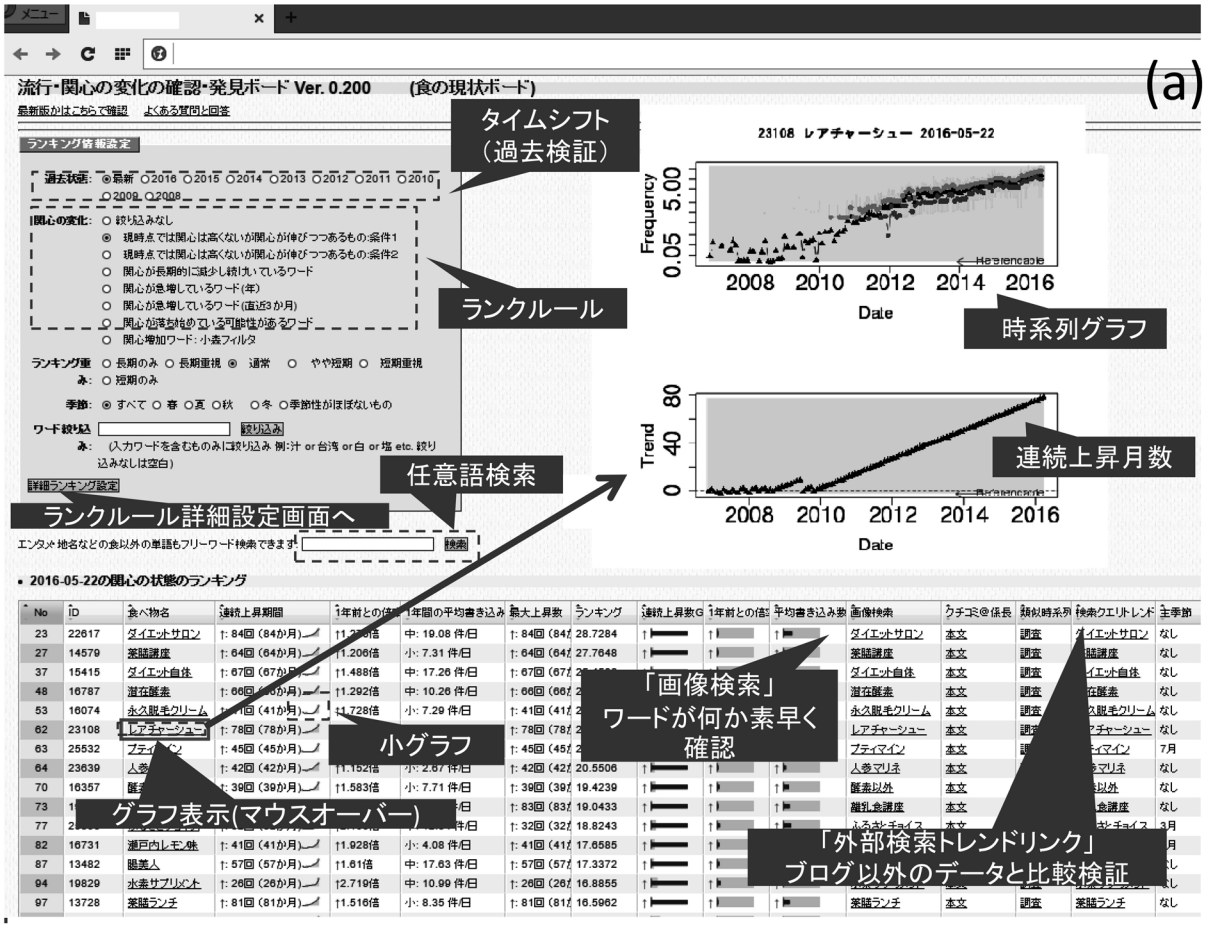
(a) 2010年11月時点での件数が大きく継続的に書き込みを増やしている語（下線は着目時点の2010年以降1年伸び続けた語。「最大上昇月数」は2015年11月までに到達した最大連続上昇月数)。

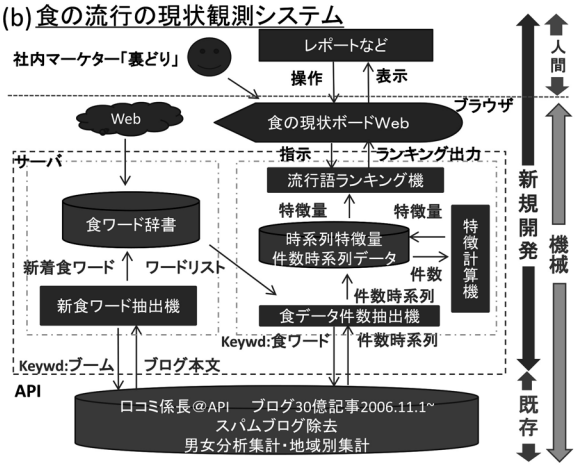
順位	ワード名	連続上昇月数2010	年増加率2010	年平均件数2010	最続上昇月数2015	順位	ワード名	連続上昇月数2010	年増加率2010	年平均件数2010	最大上昇月数2015	順位	ワード名	連続上昇月数2010	年増加率2010	年平均件数2010	最続上昇月数2015 (a)
1	**オムツケーキ**	**48**	3.25	16.4	**82**	21	川越シェフ	18	16.3	17.3	26	41	**お家ごはん**	**43**	1.32	12.6	**68**
2	ケークサレ	41	4.97	19.4	47	22	**焼き小籠包**	**29**	4.75	7.3	**45**	42	**とんてき**	**49**	1.25	5.56	**65**
3	**バイセン**	**48**	2.21	7.4	**96**	23	**塩麹**	**36**	1.95	4.9	**56**	43	**まぜそば**	**46**	1.23	14.3	**103**
4	川越達也	39	7.71	14.3	47	24	**アヒージョ**	**48**	1.80	6	**105**	44	**牛骨ラーメン**	**36**	3.68	2.37	**64**
5	赤飯さん	31	3.50	19.7	34	25	バーニャカウダソース	45	1.67	7.8	49	45	クリームチーク	31	1.62	5.57	40
6	**旦那さん弁当**	**49**	2.11	6.96	**77**	26	**マカロンタワー**	**49**	1.61	5.82	**106**	46	玉ねぎスープ	47	1.58	6.38	53
7	**変形フレンチ**	**47**	2.02	12.0	**78**	27	**アイシングクッキー**	**41**	1.89	11.9	**98**	47	**プレモル**	**34**	2.42	5.51	**81**
8	かりんとうまんじゅう	38	3.94	6.00	42	28	**酵素ドリンク**	**28**	4.38	9	**68**	48	**チーズトッポギ**	**48**	2.60	1.50	**61**
9	**SABON**	35	1.71	19.4	53	29	**料理男子**	**29**	2.87	2.1	**51**	49	米粉100%	42	1.27	3.96	43
10	**塚田農場**	**44**	3.33	4.4	**82**	30	**ホワイトフレンチ**	**49**	1.71	4.9	**86**	50	GOPAN	12	53.14	18.25	16
11	**かりんとう饅頭**	**30**	4.23	8.0	**43**	31	メガポテト	47	1.67	4	48	51	**酵素ジュース**	**31**	2.05	4.43	**72**
12	**はま寿司**	**47**	2.58	6.8	**94**	32	バラ焼き	45	1.44	6	49	52	**酵素液**	**26**	4.42	5.95	**55**
13	**糖質制限**	**49**	1.37	11.2	**104**	33	**A5ランク**	**49**	1.41	11	**87**	53	**大豆ミート**	**40**	1.71	3.13	**97**
14	**アジング**	**49**	1.28	17.6	**86**	34	**餡掛け**	**49**	1.26	9	**97**	54	**吉田類**	**43**	1.64	5.68	**63**
15	**グリーンスムージー**	**38**	5.19	12.9	**61**	35	**トンテキ**	**49**	1.23	15.0	**70**	55	**丸源**	**47**	1.53	13.32	**63**
16	**着丼**	**33**	6.36	4.5	**88**	36	**アガベシロップ**	**31**	2.07	2.84	**66**	56	**純豆腐**	**48**	1.45	15.51	**59**
17	炭酸パック	33	2.98	17.6	44	37	ダイバーケーキ	38	1.81	3	47	57	**多肉さん**	**46**	1.45	6.92	**103**
18	朝ラー	48	2.76	8.4	49	38	サムギョプル	26	1.74	0.13	34	58	すた丼	46	1.33	13.73	51
19	濃厚つけ麺	49	1.33	4.0	61	39	おんどる	49	3.89	7.5	60	59	かりん糖	49	1.31	6.25	61
20	森崎友紀	26	22.2	3.7	37	40	**塩唐揚げ**	**49**	1.8	2.5	**73**	60	ビッグボーイ	49	1.8	18.6	53



図表—3　システムと流行語の表示

(a)　食の現状ボード Web　(b) 流行解析システムのシステム構成　(c) 流行予兆候補レポートの例。詳細はⅣ節参照。





- 月次更新タスク（運用タスク）:
  - 新食語発見・件数データ更新・特徴量更新計算（1~2週間程度）
- 人間作業：システム起動，新語のチェック，データ更新の整合性の確認
  - 障害等がなければ2～3時間程度
- 更新タスクは、Web画面からのマウスクリック操作のみで可能。
  - 開発と運用の並列作業の軽減，また，将来誰でも運用可能にするため