サイバースペースと仮想都市研究会　CSVC2014-12
テレイマージョン技術研究会　TTS14-2-6
香りと生体情報研究会　SBR2014-10
(2014.6.12･13)

VennStreams：2つのトピックの対比と共起を可視化するストリームグラフ

塩澤 秀和 [1)]，Marian Dörk [2)]，Sheelagh Carpendale [3)]
1) 玉川大学 工学部
2) Potsdam University of Applied Sciences
3) The University of Calgary

あらまし：ストリームグラフは，ある集合の全体の要素数の時間的な増減と，それを構成する各サブカテゴリの要素数の時間的な増減を直感的に表示するのに適した可視化技法である．本論文では，SNS の投稿やニュースサイトの記事を対象とし，ストリームグラフにベン図の表示方法を応用することで，2 つのトピックの情報量の時間的な変遷を対比させ，さらに語句の共起による重なりも同時に可視化する技術（図 1）を提案する．最後に本手法を用いた興味深い可視化をいくつか例示する．

VennStreams: A Stream Graph Visualizing Contrast and Cooccurrence of Two Topics

Hidekazu Shiozawa [1)]，Marian Dörk [2)]，Sheelagh Carpendale [3)]
1) College of Engineering, Tamagawa University
2) Potsdam University of Applied Sciences
3) The University of Calgary

Abstract: Stream graph is a visualization technique that is suitable for showing chronological trends of both the number of the total elements and the numbers of elements in its subcategories. This paper focuses on postings on an SNS and articles in a news site and proposes a stream-graph-based visualization method (Fig. 1) that can display both the contrast between two different topics and their overlap by cooccurrence of the two sets of topical words. Also, we show some interesting examples made by our visualization.

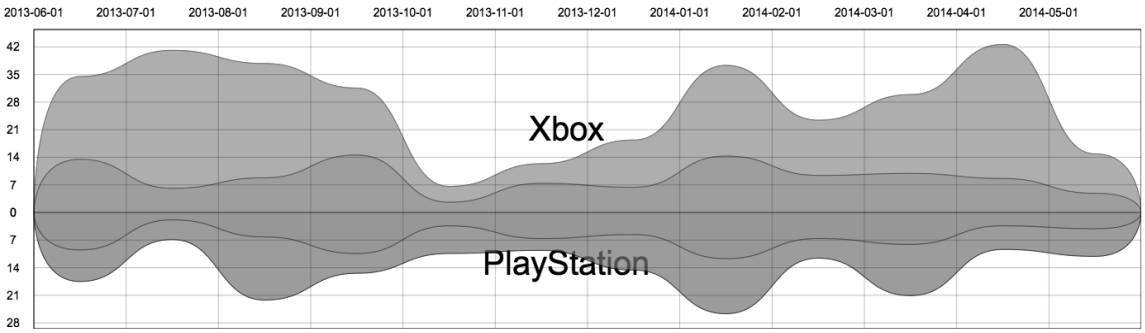

図 1　VennStreams による新聞記事数の可視化（Xbox 対 PlayStation，2013 年 6 月～2014 年 5 月）

1.　はじめに

インターネットにおける企業や個人の情報発信・情報収集の活動はますます加速している．最近では，そのようなネット上での活動のデータを収集し，時間的な変化を分析することによって，現在話題となっているトピックを判断したり，事件やイベントの社会的な影響度を評価したり，流行などの今後の動向を予測したりする手法が注目されている[1].

このような目的では，データの統計的な処理による分析だけでなく，情報可視化やデータ可視化といった技術を応用してデータの動向をグラフなどによって直感的に表示し，人間の理解を支援する技術も研究されている. 可視化の手法としては，シンプルな折れ線グラフから情報の伝搬をアニメ

ーションによって表示するものまで，多様なものが存在する．

本論文では，SNS の投稿やニュースサイトの記事を対象とし，ストリームグラフと呼ばれる可視化技術にベン図の表示方法を応用することで，時間的な変遷に伴う 2 つのトピックの対比とそれらの共起を同時に可視化する技術（図 1）を提案する．さらに，その手法を用いた興味深い可視化をいくつか例示する．

2. 関連研究

2.1 トピックのトレンドの可視化

Google トレンド[2]には，全世界のユーザが Google に入力した検索語の記録をもとに，ある語句に関する検索数が時間とともにどのように変化したかを折れ線グラフで表示する機能がある．ユーザは複数の語句を入力して，それらの検索数の時間変化を比較することもできる．Google の検索語の検索数の動向は，社会的な事件や病気の流行をよく反映すると言われている[1]．

SNS の投稿を対象に同様の折れ線グラフを表示するものとして，Topsy が提供する Twitter のトレンドのグラフ化[3]がある．Topsy は独自に Twitter の投稿（ツイート：つぶやき）を収集し，検索サービスを提供している会社である．ユーザはツイート本文を検索するだけでなく，3 つまでの語句を入力して，それらが含まれるツイート数の時間変化を可視化することができる．

これらの手法を用いることで，ユーザは，複数の折れ線グラフによって複数のトピックの動向を対比して表示させることができる．また，複数の検索語をつなげて入力すれば，それらの共起数の時間変化を可視化することもできる．しかし，折れ線グラフによる表示は，対比と共起を同時に可視化したい場合にはあまり直感的とはいえない．

2.2 ストリームグラフ

図 2 は，「ストリームグラフ」という呼称が定着しつつある情報可視化技術の例である[3]．ストリームグラフは，ある集合の全体の要素数（例えば，大学入学者数）の時間的な増減と，それを構成する各サブカテゴリの要素数（例えば，出身地別の学生数）の時間的な増減を直感的に表示するのに適している．

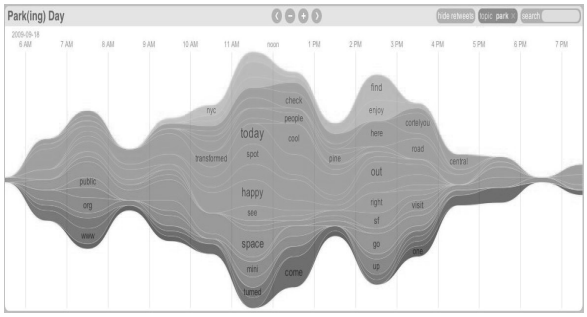

図 2 Visual Backchannel のストリームグラフ

ストリームグラフでは，各サブカテゴリの数量を時間軸にそって幅が変化する帯（stream：流れ）として表示し，それを層状に積み上げるようにして，全体の数量の時間変化を表示する．

この研究のさきがけは，文書コレクションにおける話題の動向を可視化した ThemeRiver [4]である．これは特許文書や歴史文書などの大量の文章をテーマに分類し，時間的な流れにおけるテーマの移り変わりや個々のテーマの文書の増減を直感的に可視化するものである．

現在では，ストリームグラフは Twitter 等の SNS の投稿の可視化としてもよく用いられている．図 2 は，Visual Backchannel の Twitter のストリームグラフ部分である．

構成要素の要素数の時間的な増減を可視化するというストリームグラフの特徴が分かりやすい例としては，新生児の名前の数の時間的な増減を可視化した NameVoyager [6]がある．

このような従来のストリームグラフは，トピックの時間的な移り変わりや対比を表示するのには適するが，重複関係が表現できないので共起によって同時に複数のトピックに属すべき要素の可視化が難しいという問題がある．

3. VennStreams の提案

複数の集合とそれらの共通部分（積集合）を図示する代表的な方法として，ベン（Venn）図があ

る (図 3). ベン図では集合を円などの領域で表し，共通部分を領域の重なりによって図示する．さらに，全体集合を長方形で表し，各集合の要素数を領域の面積に反映させることも一般的である．

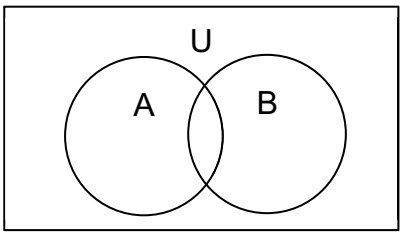

図 3 ベン図

本研究では，ストリームグラフにこのベン図の考え方を組み合わせ，2 つのトピックの重複部分をストリーム同士の重なりとして，直感的に可視化する手法を提案する．ユーザが指定した２つのトピックを表すストリームは，上方と下方に対比して表示され，上下両方のトピックに属する要素は中央のストリームとして表示される．

その際，中央のストリームの描画位置は，上下対称ではなく，上下のトピックの要素数の比率によって位置をずらして表示するのが好ましいと考えた．つまり，図 4 に示すように，中央の共通部分の要素を上下のトピックの総数の比に応じて上下に分配したとみなして描画する．

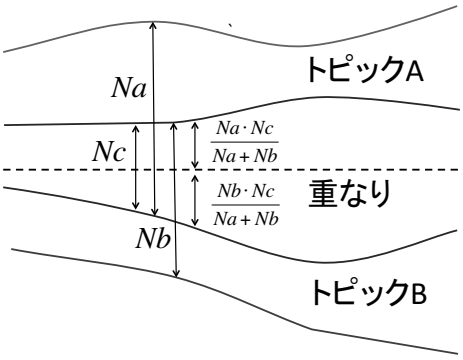

図 4 中央ストリームの描画位置

本方式によって，2 つのトピックの盛り上がりを対比して表示しながら，同時に共起によって両方のトピックに属する要素数の変化も直感的に表示することができる．本方式の制限としては，2 次元平面上の表示では３つ以上のトピックには対応できないということである．

4. SNS 投稿の可視化

4.1 ツイートの検索と可視化

VennStreams の最初のバージョンは，2011 年から 2012 年にかけて, Twitter の投稿（ツイート，つぶやき）を対象として開発された．

図 5 が実装した Web アプリケーションの画面である．ユーザはページ左上に示されたベン図のそれぞれ右の赤い円と左の青い円の横の入力欄にトピックを示す語句を入力する．さらに，全体集合を表す長方形の上の「About」入力欄には，全体の検索結果を絞り込むための語句を入力することができる．

ユーザがこれらの検索語を入力して「Search」ボタンをクリックすると，赤，青，紫に色分けされた 3 つの領域を持つストリームグラフが表示される．上の赤い領域はベン図の左の語句のみを含むツイート，下の青い領域は右の語句のみを含むツイート，中央の紫色の領域は両方の語句を含むツイートの増減を表す．絞り込み語句が指定されているが場合は，それも各検索条件に加わる．

可視化の対象とする期間は「days」入力欄で指定するができ，その期間で現在までがストリームグラフの時間軸の定義域となる．

ユーザがグラフ領域上でマウスポインタを動かすと，対応する期間で検索の際上位となるツイートがグラフ領域の下に一覧表示され，それらのデータ数に応じてベン図の円の大きさおよび重なり具合が変化する．この際，円同士の重なりの面積から中心間の距離を解析的に求めるのは難しいので，反復によって数値解を求めている．

また，試験的な実装であるが，各ストリーム上に検索語以外の頻出語句をラベルとして表示させることと，各ストリームの幅を要素数の対数にして表示することができる．

4.2 作成した可視化の保存と共有

本システムでは，ユーザが作成した可視化を他のユーザと共有できる機能を開発した．

ユーザは検索語やパラメータを変更して興味深い可視化が得られたら，「Share This View」ボタ

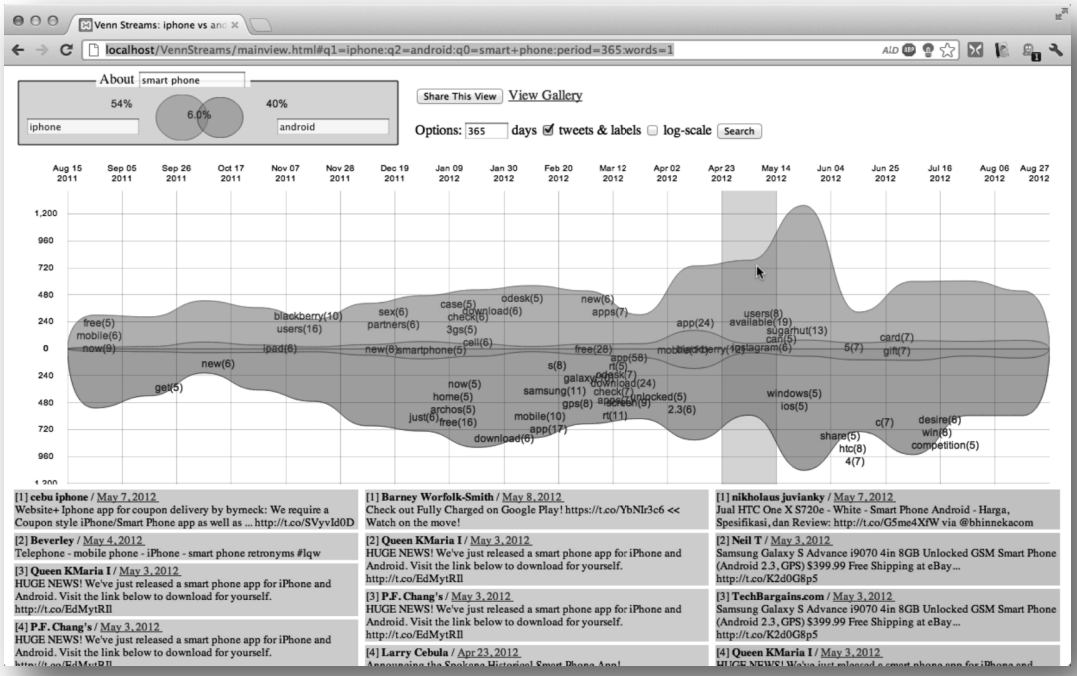

図 5　VennStreams による Twitter 投稿の可視化（smart phone を含む条件で iphone と android を比較）

ンを押してそれをシステムに保存することができる．このとき，可視化には id と URL が割り振られ, SVG による描画データと作成時の各種パラメータおよびデータがサーバに保存される．

サーバに保存された可視化の一覧は，ギャラリーページで見ることができる．ユーザはボタンを押すことによって保存された可視化（の URL）を Twitter に投稿することもできる．

4.3 開発および動作の環境

Twitter の提供する API では過去 1 週間程度のツイートしか検索することができないため，本システムの開発には，ツイートの検索サービスである Topsy が（当時）無償で提供していた検索 API を用いた．ただし，Topsy は過去のツイートについては重要でないと判断したものを検索対象から除外していくようであるため，過去のツイート数については正確な値ではない．

可視化の描画とツイートの検索はクライアントサイドの JavaScript で行っており，描画には描画命令によって SVG を生成する Raphael というライブラリを利用し，検索処理には jQuery を利用した．サーバーサイドの共有機能とギャラリー機能のためには，PHP を利用した．

この Web アプリケーションは，完成後の公開実験を計画していたが，実運用前の 2013 年 1 月に Topsy 社が無料の API 提供を停止し，その後有料サービスの API も変更され, フリートライアルも停止されたため, 残念ながら開発は中止され, 現在では動作しない状況となっている．

5. ニュース記事の可視化

5.1 New York Times 記事の検索と可視化

上述のような経緯の後，今回，VennStreams を New York Times の過去記事データベースを対象として再開発した．図 6 がそのソフトウェアのインタフェース画面である．

New York Times のデベロッパ API [7]では 1851 年 9 月以降の全記事を検索・取得することができるので，過去 150 年以上の記事について動向を可視化するシステムを実現することができた．

可視化期間のモードとしては，任意の 100 年間における 10 年ごとの変化，10 年間における 1 年ごとの変化，1 年間における 1 ヶ月ごとの変化，約 3 ヶ月間における 1 週間ごとの変化，1 週間における 1 日ごとの変化を選択することができる．

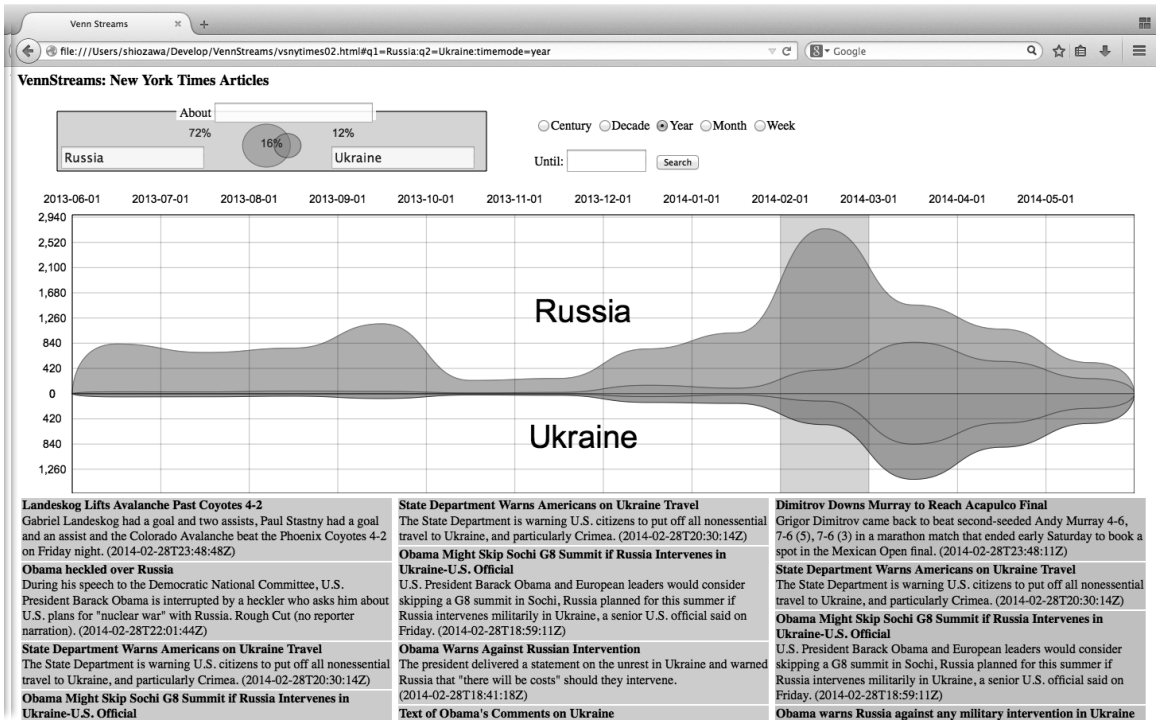

図 6 VennStreams による NY Times 記事の可視化（Russia 対 Ukraine, 2013 年 6 月～2014 年 5 月）

ユーザがグラフ領域上でマウスポインタを動かすと，記事が表示されベン図が変化するのはツイートの可視化と同様である．

5.2 開発・動作環境

本システムは基本的に 4 章で説明した Twitter 投稿の可視化ソフトウェアを改変して開発した．

記事の検索には，New York Times の記事検索 API バージョン 2.0 [7]を利用しているが，認証に必要な API キーごとの検索数の制限が厳しいので Web での公開は行えず，ローカルファイルを読み込んで起動するプログラムとなっている．このような制限のものでは，共有機能についても現実的でないので有効化していない．

5.3 ニュース記事の可視化例

新聞記事は含まれる単語数が多いために共起が起こりやすく，140 文字に制限されたツイートよりも本可視化の有効性が顕著となった．また，New York Times の過去 150 年以上の記事を検索できるため，下記のような非常に興味深いトピックの可視化することができた．

図 1 は，Xbox と PlayStation を検索語とし，最近 1 年間の動向を比較したものである。イベント等の際に記事の増加が見られる．

図 6 は，ロシアとウクライナに関する最近 1 年間の記事数を比較したものである．ウクライナ問題へのロシアの関与の大きさが分かる．

図 7 は，2004 年以来の地震と津波に関する記事数を比較したものである．東日本大震災だけでなく, 2004 年末のインド洋大津波の影響が分かる．

図 8 は，最近 100 年間について，資本主義と共産主義の記事数を比較したものである．冷戦初期のいわゆる赤狩りの共産主義の記事数が多い．

図 9 は，最近 10 年間について，ラーメンと天ぷらの記事数を比較したものである．

6. おわりに

本論文では，語句の出現数の時間変化を可視化するストリームグラフにおいて重複関係を可視化し，２つのトピックの動向の比較と分析を支援する VennStreams を提案し，いくつかの興味深い可視化例を紹介した．

今後は，他の情報源への適用，共有機能の再実現, システムの評価などを行いたいと考えている.

参考文献

[1] J. Ginsberg, et al.: Detecting influenza epidemics using search engine query data, Nature, Vol.457, pp.1012-1014 (2009.2)

[2] Google Inc.: Google トレンド, http://www.google.com/trends/

[3] Topsy Labs Inc.: Social analytics, http://topsy.com/analytics

[4] S. Havre, et al.: ThemeRiver: Visualizing thematic changes in large document collections, IEEE Transactions on Visualization and Computer Graphics, Vol.8, No. 1, pp.9-20 (2002.1)

[5] M. Dörk, et al.: A visual backchannel for large-scale events, IEEE Transactions on Visualization and Computer Graphics, Vol.16, No.6, pp. 1129-1138 (2010.11-12)

[6] Generation Grownup, LLC: The Baby Name Wizard: NameVoyager, http://www.babynamewizard.com/voyager

[7] The New York Times: Article search API v2, http://developer.nytimes.com/docs/read/article_search_api_v2

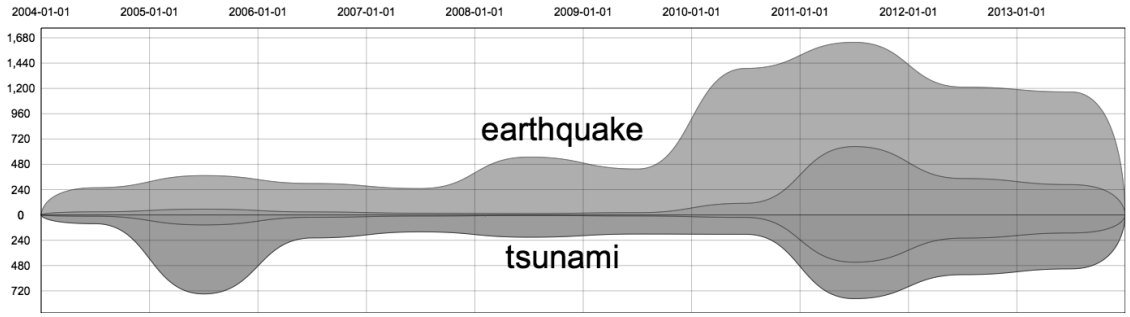

図 7 earthquake 対 tsunami, 2004 年～2013 年

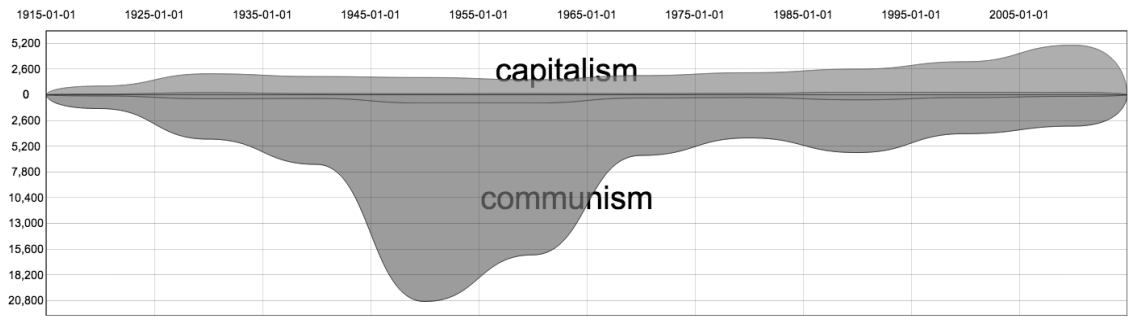

図 8 capitalism 対 communism, 1915 年～2014 年

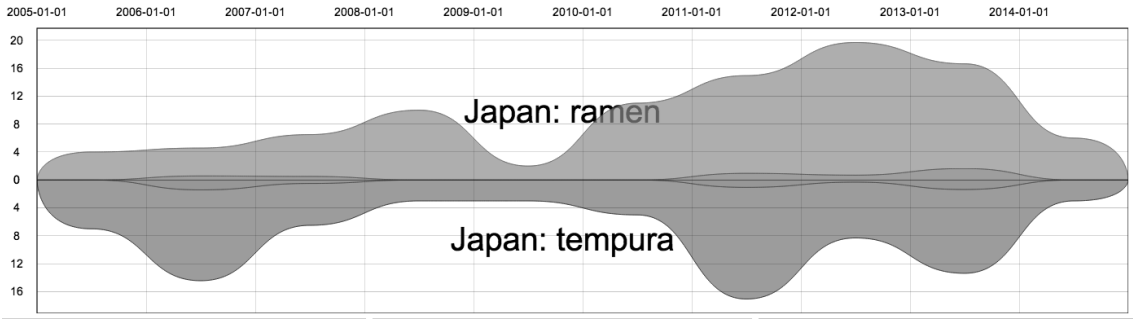

図 9 ramen 対 tempura（Japan を含む記事中), 2005 年～2014 年