社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－ 93 (14)
2009－ＮＬ－189 (14)
2009／1／23

文脈規定に寄与する要素に関する考察

佐藤 進也[†] 福田 健介[‡] 栗原 聡[††] 廣津 登志夫[‡‡] 菅原 俊治[†††]
[†]NTT 未来ねっと研究所 [‡]国立情報学研究所 [††]大阪大学 [‡‡]豊橋技術科学大学 [†††]早稲田大学

ある語 (が指し示す概念) に特有な文脈を発見する方法として，文書集合を文脈規定に寄与する要素に着目して構造化し，その構造上で当該語の出現分布を調べるというアプローチを提案する．実際に，文脈規定に寄与する要素として，時間，年周期性，人のつながりを選び，文書集合としてブログを用いて文脈発見を試みた結果を示す．

A study on factors for defining contexts

Shin-ya SATO[†] Kensuke FUKUDA[‡] Satoshi KURIHARA[††] Toshio HIROTSU[‡‡]
and Toshiharu SUGAWARA[†††]
[†]NTT Network Innovation Laboratories, [‡]National Institute of Informatics, [††]Osaka University,
[‡‡]Toyohashi University of Technology, [†††]Waseda University

We present a method to discover contexts peculiar to a given term (a concept represented by a given term) by analyzing a document set. In our approach, first, a factor that contributes to definitions of contexts is selected. Then, the document set is organized (structured) on the basis of the factor. Finally, the distribution of term occurrence over the structured documents is analyzed. We also show some actual examples of context discovery where time, 1 year periodicity, interrelationships among people are respectively used as context factors.

1 はじめに

ことばの意味 — たとえば，ある語（固有表現）が何を指し示しているのか— を理解するためには，多くの場合，そのことばが置かれている文脈の把握が必要である．語として人名を例にとると，人名と人物の対応には同姓同名に由来する曖昧性がある．よって，ある文書中に記された人名をそれが指し示す人物に対応させるためには，その記述の文脈を把握し，この曖昧性を解消する必要がある．実際，人名の曖昧性解消を目的として，文脈把握のために有用な情報（当該文書に出現している，人物特定に効果的な語など）の抽出方法が盛んに研究されている [1].

本論文では，文脈把握というテーマに焦点を当てながらも，従来の曖昧性解消を目的とした研究とは異なったアプローチを試みる．具体的には，ある文書中である語が置かれている特定の文脈を

理解するという目的から（当面）離れ，語にとって特徴的な文脈，語から想起される状況 — たとえば,「西瓜」という語に対する「夏」 — を見つけ出す問題を考える．

2 アプローチ

この問題を解くために，ここで例として挙げた『「夏」を「西瓜」に対応させる』ということがどういうことなのか，この対応をどのようにして導き出すことができるのかを考えたい．

まず,「夏」という語からは，その特徴を与える気候や様々な行事などが想起されるが，この概念を定義する場合，時間軸上（一年の中）のある期間（一般に，だいたい 6〜8 月）として説明されることが多い．時間という物理量を基準とすることには解析（データの処理やその結果の客観的解釈など）上の利点が多いので，この定義に基づいて議論をすすめることにする．このとき「夏」が「西瓜」に対応付けられるということは,「西瓜」にとっては 6〜8 月という期間が（他の期間より）特別な意味を持つということとして捉えることができる．そして，このように捉えると，上記の主張が妥当であることを定量的に示すことができる．図 1 は,「西瓜」でブログ検索[1]を行い，その結果を日毎に集計したものである．この語の出現頻度（エントリ数）が 7〜8 月を中心として明らかに高くなっていることがわかる．

ここで示した手法は，時間軸で語の出現頻度を調べ，その分布状況（出現の集中）からその語特有の文脈を見つけ出すというものである．このアプローチが有効であるのは，その語（例では「西瓜」）に関する（ある）文脈を規定するうえで，時間という要素が重要な意味を持っているからである．このように考えたとき，例に示した手法は次のように一般化できる：

> 文脈発見の戦略
> 文脈規定に寄与する要素を選び，その要素を基準として文書集合に関係を導入し，構造化する．そして，構造化された文書集合における語の出現分布を調べ，集中的に出現している個所を抽出する．

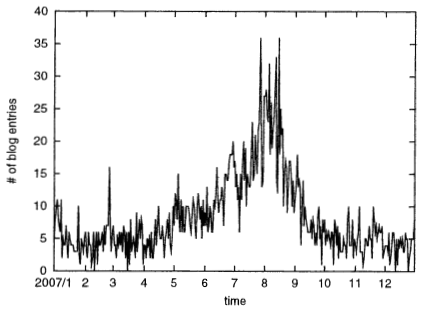

図 1: ブログにおける「西瓜」の出現頻度（2007 年 1 月 1 日より 1 年間）

この一般化に「西瓜」の例を当てはめると次のようになる．まず，文脈規定に寄与する要素として時間を選び，ブログエントリ間に時間的近接性という関係を入れる．このとき，ブログエントリは 1 次元の構造をもつことになる（時間軸上にエントリを並べているイメージ）．その構造上，語の出現分布を調べると，集中的に出現している個所として「夏」に対応する期間を取り出すことができる．本論文では次章以降，いくつかの具体例を通してこの戦略の適用可能性を探っていく．

なお，工学的観点からすれば，時間といった（本文の記述に対して）付加的な情報が取得可能であることも本アプローチを成立させるうえで重要である．これまで，文脈把握のため，文書中から抽出した関連語を利用する手法などが試されてきたが，これは，本文の中に文脈のヒントを見つけ出そうとするアプローチである．近年では，ブログなどのオーサリング環境が充実し，本文を作成した時刻（タイムスタンプ）などのデータも自動的に付与されるようになっている．これらのデータは本文からは得られない情報を含んでいるため，文脈把握のためには，本文に劣らず有用であると考えられる．

[1] http://blog-search.yahoo.co.jp/

3 年周期性

本章では，周期性，特に年周期性を文脈規定要素として選び，2章で述べた文脈発見戦略の適用を試みる．ここでも，前章同様ブログを文書集合として用いる．つまり，簡単に言うと，与えられた語のブログでの出現パターンが周期的であるか否かを判別し，周期性がある場合には，そのピークとなる期間を見つけ出す．なお，2章で扱った「西瓜」の例については，そもそも「夏」自体が年周期を前提とした概念であり，まず年周期性に着目するのが自然なアプローチと言えるだろう．しかし，説明の都合上，時間を文脈規定要素として選択したケースとして紹介した．

さて，まず，いくつかの具体例に対して出現パターンの周期性を調べてみる．そのために，ブログ検索サービスを利用し，3年の間（2006年10月1日〜2008年9月30日）のブログにおける語の出現頻度（エントリ数）を日単位で調べた．図2は「西瓜」の出現パターン（時系列）であり，年単位の周期性が見て取れる．

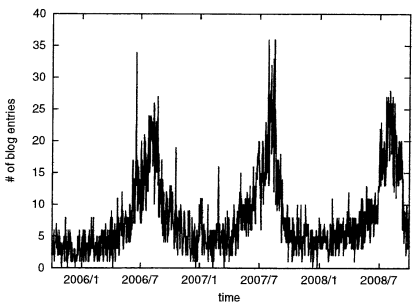

図2: ブログにおける「西瓜」の出現頻度（2006年10月1日より3年間）

周期性は，時系列の周波数成分を調べることにより，定量的に把握することができる．図3は図2の時系列のトレンドを除去した後にパワースペクトルを計算した結果（の低周波部分）であり，年周期に対応する周波数3でピークを示していることが確認できる．

「西瓜」以外にも，いくつかの名詞に対してパワースペクトルのピークから年周期性を判定した．その結果を表1に示す．「ゴディバ」「デュブッフ」はそれぞれチョコレートとワインのメーカーである．これらの語の出現が周期性を持つのは，それぞれが生産する商品が特別な意味を持つイベント，すなわちバレンタインデーとボジョレーヌーヴォ解禁，が毎年行われているからである．また，「発光ダイオード」もまたイベントに付随して周期性が生じているもので，クリスマスシーズンのイルミネーションに関する記述の中でこの語を見つけることができる．「ミネラル」に関しては，実は，多くの場合この語単体で用いられるのではなく，「ミネラルウォーター」という複合語の一部としてブログに現れている．そして，この語の周期性は，ミネラルウォーターが夏期によく消費されることに起因すると考えられる．

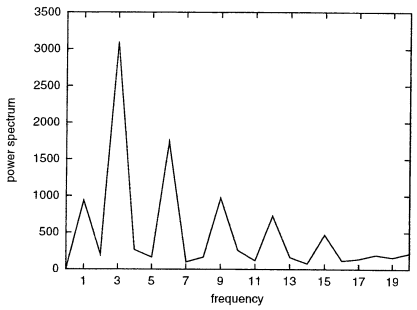

図3: 「西瓜」の出現頻度のパワースペクトル

表1: 年周期性の有無

年周期性あり	イチゴ，ゴディバ，デュブッフ，ミネラル，発光ダイオード
年周期性なし	グーグル，シェラトン，スモールワールド，年金，日本

さて，次に，年周期性が認められた語に対して文脈発見戦略を適用する．戦略の定義に従い，まず，年周期を基準として文書（ブログエントリ）間に関係を導入する．文脈規定要素として時間を選んだ場合，関係の導入はブログエントリを時間

軸という直線上に並べることに相当したが，年周期の場合は円上への配置と考えることができる（図4）．

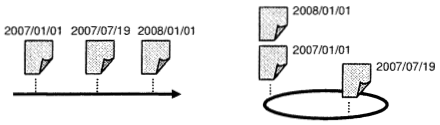

図 4: 文脈発見戦略における関係導入により文書集合を構造化した状態を摸式的に表した図．数字は各文書のタイムスタンプを示す．左右はそれぞれ文脈規定要素として時間と年周期を選んだ場合．

そして，語の分布を調べることは，円周上の同一点ごとに，当該語を含むブログエントリを集計することに対応する．「イチゴ」に対して，この集計処理を行った結果を図5に示す．グラフ中，実線で示したのが集計結果である．1年というスケールでの大まかな変動パターンを把握するため，実際のデータをスプライン曲線で平滑化してある．3年間の時系列の中から1年間分を抜き出すことにより同様なパターンを得ることもできるが，複数年のデータを加算することにより，各年に共通した特徴は強調される一方で，各年固有のゆらぎ（ノイズ）は弱められるという効果が得られる．

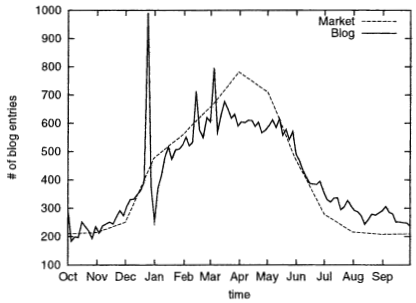

図 5: 「イチゴ」のブログにおける出現頻度の変動パターン

集計により，とりあえず，ブログから変動パターンが得られたが，いったいこの情報はどれだけ実社会の状況を正しく捉えているのだろうか? この疑問に答えるため，東京卸売市場の統計データ[2]，具体的には築地市場における「いちご類」の2007年10月から2008年9月までの取扱実績と比較した．図5で破線が統計データのグラフであり，ブログから得たデータとの比較を容易にするため，Y軸方向のスケールは適宜調整してある．2種類のデータの変動パターンには明らかな類似性が認められる．この結果から，文脈発見戦略によってブログから獲得した情報は「イチゴ」の「旬」という文脈をかなり正確に捉えていると言えるだろう．

さらに，図5のグラフで特徴的なのは，「旬」のピークから若干離れた12月下旬にあるピークである．この時期のブログエントリを実際に調べてみると，クリスマス用菓子の素材としてのイチゴの記述を多く見つけることができる．つまり，このピークは，「旬」という果物が共通して有する広く知られている文脈とは異なった，「イチゴ」に固有な文脈の存在を示すものである．

4 コミュニティ

本章では，時間や年周期性とは異なったタイプの文脈規定要素を選び，文脈の発見を試みる．その要素とは，人と人とのつながりである．実社会，そしてWebなどの仮想空間においても，同じ興味を共有する人々，同じ目的を持っている人々は，多くの場合，その活動の過程で互いのつながり—いわゆるコミュニティ—を形成する．メンバー間で共有されている興味や目的は，そのコミュニティ固有の文脈と捉えることができる．よって，ある語（概念）について興味を持っているコミュニティを発見することは，その概念に固有な文脈の発見につながる．これが，文脈発見のために人のつながりに着目する理由である．

前章までと同様，本章でも文書集合としてブログを用いる．ブログの場合，文書（ブログエントリ）と著者（ブロガー）を明確に対応付けることができる．さらに，ブログによる情報交換の状況を解析することにより，ブロガーどうしの関係を

[2] http://www.shijou-tokei.metro.tokyo.jp/ より入手可能（2008年12月現在）．

推定することも可能である．本論文では，コメントのやり取りに基づいてブロガー間の関係を推定し，文脈発見に用いる．

図6は，ブロガー間のコメント交換関係を表すネットワーク（コメント交換ネットワーク）のうち「ワイン」というトピックに関係する部分を抜き出し，可視化したものである．各ノードは個々のブロガーに対応し，リンクは，ブロガーが互いにコメントを述べ合うという双方向の関係に対応している（詳細は [2] 参照のこと）．描画には Fruchterman-Reingold のアルゴリズム [3] を用いているので，基本的に，より密な関係のあるノードどうしが近接して配置されている．

図 6: ブロガーのコメント交換ネットワーク

さて，文脈発見のために人のつながりに着目したのは，人々の相互のつながりが固有な文脈に対応すると考えられるからであった．まず，このことを図6のネットワークで確認しよう．具体的には，ネットワークの中で疎な個所[3](E) と稠密な個所[4](C) に属するブロガーがそれぞれのエントリ中で言及しているワインのタイプ（生産地）を比較する．(C) ではメンバ間での文脈の共有があり，その結果として，そこで話題にのぼる銘柄が (E) に比べより特定のタイプに集中していることが期待される．図7はタイプの分布を示したグラフであり，予想通り，(C) では特定のタイプに集中していることが確認できる．なお，(E) と (C) では記事を書く目的にも違いが認められる．(C) では，ワインを飲んだ感想を記したものがほとんどである．一方，(E) ではワイン販売店による商品情報の提供を目的としたブログが見受けられる．

[3]他とつながりをもたない孤立したノード群
[4]多くのつながりを持ちクラスタ性の高いノードの近傍

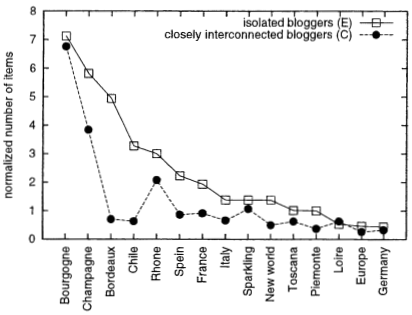

図 7: ブログ中に現れるワイン銘柄をタイプ（生産地）ごとに集計したグラフ

いよいよ，文脈発見戦略に則り，ブログエントリをコメント交換ネットワーク上に配置することを考える．ここでは，各エントリをそれを著作したブロガー（に対応するノード）に対応付けるという単純な配置方法を試みる．

語の分布に関しては，当該語を含むエントリの数をノード単位で集計し，その数値の分布を調べることもできるが，ここでは，語を含むエントリの有無だけに着目する．図8は，いくつかの語（ワインの銘柄; (a), (b) はフランスブルゴーニュ産，(c) はイタリア産）について，その分布を示したものである．各ブロガーの，当該語についての言及の有無を，それぞれ●と○で示してある．

ここで注目すべきは，ワインの種類（産地）と分布が集中している個所との間に対応関係が認められるということである．また，この分布から銘柄の認知度を読み取ることもできる．実は，(a) と (b) は同じ産地ではあるものの，認知度に差がある（後者は有名な高級銘柄である）．その違いが●で示したノードの数，そして，その分布範囲の広さの違いとして現れている．

さて，図8の (a) のように，ある銘柄 x の分布が特定の個所に集中しているということは，x について興味を持っている人々がコミュニティを形成しているということである．このとき，x はコミュニティメンバ間で共有される嗜好と合致していることが期待できるだろう．そして，このコミュニティがまた別の銘柄 y に価値を見い出したならば，

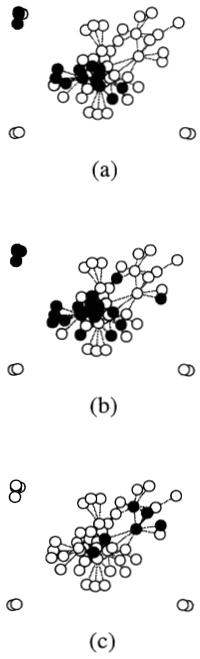

図 8: コメント交換ネットワーク上の語の分布

x と y は共通した特徴（味わい，香りなど）を持つと考えられる．この場合，ブロガーのコミュニティは（ある特定の）嗜好という文脈を提供していると考えられる．

5 むすび

本論文では,「西瓜」に対する「夏」のようなマクロな文脈を文書集合を解析することにより発見する戦略を提案し，いくつかの例を通してその有用性を確かめた．その戦略とは，(1) 文脈規定に寄与する要素を選び，その要素を基準として文書集合に関係を導入して構造化し，(2) 構造化された文書集合における語の出現分布を調べ，集中的に出現している個所を抽出する，というものである．

戦略の (1) は解析の視点を選ぶということである．本論文では時間，年周期性，人のつながりという 3 種類の視点を選び，ブログエントリという文書集合を解析した．その結果，それぞれに特有の文脈を発見することができた．

一般に，新しい視点は解析対象の未知の側面に光を当て，より深い理解をもたらす．文脈発見においても，文脈規定に寄与する新たな要素を見つけ出すことが鍵となる．今後，その手法についても検討していきたい．

参考文献

[1] 関根聡. Web 検索における人名の曖昧性解消技術の動向 — 同姓同名のクラスタリング —. 情報処理, Vol. 49, No. 5, pp. 573–578, 2008.

[2] Shin-ya Sato, Kensuke Fukuda, Toshio Hirotsu, Satoshi Kurihara, and Toshiharu Sugawara. Co-occurrence Analysis Focused on Blogger Communities. Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence. pp. 372–376, 2008.

[3] Thomas M. J. Fruchterman and Edward M. Reingold. Graph Drawing by Force-directed Placement. Software – Practice and Experience, Vol. 21, No. 11, pp. 1129–1164, 1991.