局所性を用いた多様性を考慮したブログからのトピック抽出手法について

戸田　智子† 　横山　昌平†† 　福田　直樹†† 　石川　博††

† 静岡大学大学院情報学研究科　〒 432–8011 静岡県浜松市中区城北 3–5–1
†† 静岡大学情報学部情報科学科　〒 432–8011 静岡県浜松市中区城北 3–5–1
E-mail: †gs07037@s.inf.shizuoka.ac.jp, ††{yokoyama,fukuta,ishikawa}@inf.shizuoka.ac.jp

あらまし　ブログの爆発的普及により，ブログを用いて情報発信をする人，ブログを用いて情報収集する人双方が増加してきている．それに伴い，ブログを解析し，有益な情報を抽出しようとする試みが多くなされてきている．現在，このような試みはブログマイニングとしてさまざまな方面からのアプローチがなされている．これらの研究は，ブログの特徴である個人性や即時性，時系列データであることなどに着目している．本研究ではその中でも，時系列データであること，ブログの利用目的・ブログ著者には多様性があることに焦点を当てる．ここで課題となることの 1 つに，ユーザのニーズに応じて必要とされるトピックの切り出し方が異なってくるということがある．1 つのテーマは複数のトピックに分割することができるが，同一のテーマでも，視点が変化することによって，異なるトピックの集まりに分割されることがある．本論文では，ブログ記事よりトピックを抽出する際に，ユーザの多様な目的を意識し，多視点からトピックを抽出する手法に焦点を当てる．本論文では，ユーザが捉えたいと思う視点の違いが，文書中に出現する本来特徴語となるべき品詞などの特徴量の差異として出現するという仮説に基づき，品詞情報等の特徴量の差異を用いたブログ記事群からの多様な視点からのトピック抽出を実現する手法を提案する．

キーワード　ブログ文書，Locally Weighted Clustering，トピック抽出

Towards Adaptive Topic Extraction from Blogs Using Locally Weighted Clustering Diversities

Tomoko TODA†, Shohei YOKOYAMA††, Naoki FUKUTA††, and Hiroshi ISHIKAWA††

† Graduate School of Informatics, Shizuoka University　Johoku 1–2–3, Nakaku, Hamamatsu-shi, Shizuoka, 432–8011 Japan
†† Department of Computer Science, Faculty of Informatics, Shizuoka University　Johoku 4–5–6, Nakaku, Hamamatsu-shi, Shizuoka, 432–8011 Japan
E-mail: †gs07037@s.inf.shizuoka.ac.jp, ††{yokoyama,fukuta,ishikawa}@inf.shizuoka.ac.jp

Abstract　Doe to growth of blogs, There are certain needs for gathering useful and extracting information from them. There have been proposed approaches that are trying to extract topics from blogs. However, there are still difficult issues to cut out appropriate topics for various needs. It is needed to analyze the same topic from multiple aspects according to each situation rather than static aspects. Therefore, it is very important to extract topics from a large number of blog articles in several aspects. In this paper, We focus on topic extraction from blog articles that are demanded to be analyzed from multiple aspects. In this paper, we propose a method for adaptive topic extraction in various aspects from the specified blog articles by using grammatical characteristics of words that are related to the aspecs. Locally Weighted Clustering is applied for this purpose.

Key words　Blog, Locally Weighted Clustering, Topic Extraction

1. はじめに

ブログはここ数年において急速に発展し，新たな情報源として注目されている．ブログの普及により，ブログを用いて情報を発信する人の増加とともに，そういったブログの情報を利用して何かを行う人が増えてきている．現在のブログの特徴とし

ては，ウェブ上での個人の日記という側面と，特定のニュースやイベント・製品などに対する個人の意見を表現するメディアの 1 つという側面がある．また，従来の Web ページと異なり，時系列をそれなりの精度で追跡できるという特性も備えている．特定のニュースやイベントに対する個人の意見を表現するという側面においては，ブログから有意義な情報を抽出したいという要求が高まってきている．ブログを通して発信された情報が，ブログ閲覧者の行動決定に対して，一役を担ってきていると考えられる．Nifty の調査 [1] によると，ブログを情報源として利用したことがある人は，調査対象者の 96% にも上るとされている．また，ブログを閲覧する主な目的としては,「新しい情報を得るため」や「趣味や余暇の役に立つ」,「気分転換・ストレス解消」などが挙げられている．

このようなブログの個人性や時系列性に着目し，ブログ内での評判情報の抽出や話題の変遷の抽出を行う研究やサービスが，これまでに提案されてきている．ブログで扱われている話題や口コミ情報の変遷を可視化しているサービスなどがある [2], [3]．多くのブログ著者の間で記述されている話題について，関連語などを表示するサービスなどがある．また, [4] や [5] では，ユーザが入力したキーワードに対して検索されたブログ記事をクラスタ化して提示することなどを行っている．ブログ記事から評判情報や口コミ情報の抽出，評判情報の変遷に関する研究も行われている．Mei ら [6] では, 同様に確率モデルを用いてトピック抽出を行い, 肯定・否定に分類し, それぞれの時間的変化を可視化している．また，抽出された評判情報やその変遷を用いて，実世界の予測を行うような研究も行われている．Gilad ら [7] は, ブログで語られている評判から映画の興行成績を予測する手法，Liu ら [8] は，ブログ中から抽出した評判情報を用いて，商品の売り上げを予測する手法を Gruhl ら [9] は, ブログ記事の投稿数と書籍のランキングが関係していることを示している．これらから，ブログに記述された情報がブログ閲覧者に影響を与えるのではないかということが考えられる．また，ブログを利用した，別の側面からの研究としては，ブログから話題の空間的広がりを抽出するという試みもなされている．Mei ら [10] は, 確率言語モデルを用い, トピック抽出を行い, トピックの変遷を可視化する手法を提案している．

また，文献 [11] において，奥村はブログについて次のように述べている．ブログは通常の Web ページとは異なり，速報性，リアルタイム性のある新鮮な情報が発信されることから，掲示板同様有用な情報源と考えられるようになってきている．このブログを大量に収集し，収集したブログ集合をさまざまな手法で分析することで，一般の人々の「生の声」をうまく抽出することに現在関心が集まっている．

このようにブログが普及してきたことにより，ブログを閲覧するユーザの目的・意図が多様になってきているという一面もある．何かある製品について調べたいというような場合においても，その製品を買う場合にその製品の評判や口コミ情報について知りたいような場合や，その製品の詳細を知りたい場合，また，その製品を実際に購入した人，あるいは購入を控えた人はどの程度いるかなどを知りたい場合など一様ではないと予測される．

本論文では，このような，あるニュースやイベントに対する反応・話題を**トピック**と呼ぶこととする．同じ話題に関する事柄でも，性質の異なるものは，異なるトピックとして捉えた方が良い可能性があると考えられる．そのような状況において，ブログ閲覧者の目的の多様性にかかわらず，一様なトピック抽出では，ユーザにとって有益なトピック抽出を行うことは困難であると考えられる

我々は，文献 [12] において，ブログ記事群よりトピックを抽出し，抽出したトピックに対して，投稿されたタイムスタンプと記事数に基づいて，トピックの変遷を抽出し，可視化する手法を提案し, 文献 [13] では抽出したトピックに対して，その肯定・否定の時系列上の変化を可視化することにより，トピックの評判情報やその動向を抽出する手法を提案した．文献 [14] において，階層的クラスタリングを用いた多視点トピック抽出を試みた．本研究では，その改良として，局所性を利用したクラスタリングである Locally Weighted Clustering [15] を用いた多視点トピック抽出の試みについて報告する．

2. 多様な視点を考慮したトピック抽出

2.1 研究の目的

本手法では，ユーザが捉えたいと思う視点の違いが，ブログ記事中において，特徴語となるべき語の，品詞の差異として出てくるのではないかという仮定に基づいている．本手法では, 視点をそれぞれ品詞に対応付け，3 つの視点を定義することとした．名詞を中心に捉えることにより，そのトピックを構成する物事を中心とするトピック抽出，動詞を中心に捉える場合には,「行った」や「買った」などの行動中心のトピック抽出，また，形容詞を中心に捉える場合には,「良い」や「悪い」,「嬉しい」などの感想中心にトピック抽出を行うことができると考えられる．本手法では，品詞ごとに重み付けを行うことによって，物事中心，行動中心，感想中心など，ある 1 つのトピックに対しても異なる視点でトピック抽出を行うことを目的とする．

文献 [14] において提案した，多様な視点を考慮したトピック抽出手法を拡張する．文献 [14] においては，階層的クラスタリングを用いてトピック抽出を行っていたが，計算量や処理時間を考慮すると，非階層的クラスタリングを用いたトピック抽出を行いたい場合があると考えられる．特に，人手でのパラメータの調整や，記事群の選別のために何度もクラスタリングを繰り返す場合には，処理時間の短縮は重要な要素の 1 つとなる．

そこで，単純な凝集型階層的クラスタリングに換えて，非階層的クラスタリング手法である K-means 法を拡張した LWC(Locally Weighted Clustering) 法 [15] を用いたトピック抽出について試みる．

クラスタとは，似た特徴を共有しているデータの集合である．この定義はかなり直観的であるので，多次元のデータセットを意味のあるクラスタに分割するということは重要な問題である．データオブジェクトはクラスタリングアルゴリズム中では，特徴ベクトルとして表現される．しかし，その特徴空間は大抵の場合複雑であり，元のデータの特徴よりも小さくなってしまう．

加えて，データは大抵均一ではない．このことから，そのデータのそれぞれのサブセットの中で影響を与える次元が異なっていると考えられる．したがって，それぞれの特徴次元が，データ空間全体で均一に重要であるとは限らない．

LWC(Locally Weighted Clustering) 法では，非階層的クラスタリングの代表的手法である k-means 法を拡張し，クラスタごとに異なった重み付けを行うことによって，“意味のある”クラスタを抽出することを目的とし，評価実験によりクラスタリングの精度の改善が見られることが示されている．

2.2 トピックの定義

本研究では，ある話題を扱っているブログ記事の集合をテーマとして抽出する．テーマの抽出は，与えられたキーワードを含んでいるかどうかによって行う．抽出したテーマをまとまりごとに排他的分割することにより，トピックを抽出する．その分割の方法を視点とする．テーマ，トピック，視点の関係について，式 1 にて示す．ここでは，視点 a と視点 b があるものとする．

$$\begin{aligned} Theme &= tp_1^a + tp_2^a + \ldots + tp_n^a \\ &= tp_1^b + tp_2^b + \ldots + tp_m^b \end{aligned} \tag{1}$$

ここで，$Theme$ は与えられたキーワードに基づいて抽出したブログ記事集合を示す．tp_1^a〜tp_n^a は，視点 a において抽出されたトピック，tp_1^b〜tp_m^b は，視点 b において抽出されたトピックを示す．

2.3 LWC(Locally Weighted Clustering) 法

$\mathcal{R}^m$ は m 次元のデータ空間で，N 個のデータ $\vec{x}_i$ を含んでいる．$\vec{x}_i$ の j 番目の要素は x_{ij} である．K-means 法ではクラスタは重心 $\vec{c}_k \in \mathcal{R}^m$ で表わされ，与えられたデータはユークリッド距離や global マハラノビス距離などに基づき最も近い重心に割り振られる．しかし，こういった global distance metric は局所構造を捉えることに不向きである．

LWC 法では，異なるクラスタに対しては異なる重み付けを行った距離関数を用いている．具体的に言うと，重心 $\vec{c}_k$ と別に，そのクラスタ内に含まれる要素から導き出した重みベクトル $\vec{w}_k$ を用意する．データ $\vec{x}$ と重心 $\vec{c}_k$ の距離を重み $\vec{w}_k$ によって拡大や縮小している．本研究では，コサイン類似度を用いたため，類似度は式 2 に示す式で算出する．

$$\begin{aligned} &\mathcal{L}_{2,\vec{w_k}}(\vec{x},\vec{c_k}) = \\ &\qquad \frac{w_{k1}x_1 * w_{k1}c_{k1} + \ldots + w_{km}x_m * w_{km}c_{km}}{|\vec{x}| * |\vec{c}_k|} \\ &\text{ここで,} |\vec{x}| = \sqrt{(w_{k1}x_1)^2 + \ldots + (w_{km}x_m)^2} \\ &\qquad |\vec{c}_k| = \sqrt{(w_{k1}c_{k1})^2 + \ldots + (w_{km}c_{km})^2} \end{aligned} \tag{2}$$

それぞれのデータは類似度関数を適用することにより，最も類似度の大きいクラスタに振り分けられる．メンバシップ関数 ϕ_c として，データ $\vec{x}$ を K 個のクラスタのうちどのクラスタに振り分けるかについては式 3 に基づいて算出する．

表 1　LWC アルゴリズム

アルゴリズム Locally Weighted Clustering(LWC)

Require： $\vec{x}_i \in \mathcal{R}^m$，クラスタ数 K

Ensure： クラスタ K の重心 $\vec{c_k}$，重み $\vec{w_k}$

1: クラスタの初期重心を決定する．
すべての重みベクトルを 1 にする．

2: E-Step: N 個のデータすべてに対して式 3 を用いて $\phi_c(\vec{x})$ を算出し，K 個のクラスタの重心を導出する．

3: M-Step: それぞれのクラスタに対して式 6 により，重心を再計算し，重み $\vec{wk}$ を更新する．

4: 収束するまで 2 と 3 を繰り返す．

$$\phi(\vec{x}) = \arg \max_{1 \leqq k \leqq K} \mathcal{L}_{2,\vec{w_k}}(\vec{x},\vec{c_k}) \tag{3}$$

よって，k 番目のクラスタに属するすべての要素は次のように表わされる．

$$C_k = \{\vec{x} | \phi_c(\vec{x}) = k\} \tag{4}$$

最適なクラスタを得るために，重心の集合と一致するクラスタの重みはともに，そのすべてのデータとそれぞれ重心との類似度の平方和が最大になる必要がある．

$$\begin{aligned} &\sum_{i=1}^{N} \mathcal{L}^2_{2,\vec{w}_{\phi_c(\vec{x}_i)}}(\vec{x}_i, \vec{c}_{\phi_c(\vec{x}_i)}) \\ &\text{このとき,} \forall k \prod_{j=1}^{m} w_{kj} = 1 \end{aligned} \tag{5}$$

定理 1

式 4 で定義した問題に対して，最適な重心，および最適な重みベクトルは次の式によって算出する．$1 \leqq k \leqq K$，$1 \leqq j \leqq m$ とする．

$$c_{kj} = \frac{1}{|C_k|} \sum_{\vec{x} \in C_k} x_j \tag{6}$$

$$\begin{aligned} &w_{kj} = \frac{\lambda_k}{\sum_{\vec{x} \in C_k} |x_j - c_{kj}|^2} \\ &\text{このとき,} \lambda_k = (\prod_{j=1}^{m} (\sum_{\vec{x} \in C_k} |x_j - c_{kj}|^2)^{\frac{1}{m}} \end{aligned} \tag{7}$$

$LocallyWeightedClustering$ のアルゴリズムを表 1 にまとめる．

LWC の特性を調査するために，局所性を持たせたデータを作成し，それに対してクラスタリングする予備的実験を行った．また，LWC の効果を調べるために，単純な K-means 法でクラスタリングを行ったものと比較をした．

実験に際して，作成したデータを表 2 に示す．表 2 を用いて，

表 2　LWC 特性調査のための作成データ

$D_1 = (1.0, 0.7, 0.5, 0.4,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0, 2.0)$
$D_2 = (0.8, 0.6, 1.0, 0.4,\ 0, 2.0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)$
$D_3 = (0.5, 1.0, 0.7, 0.4,\ 0,\ 0,\ 0,\ 0, 3.0, 2.0,\ 0,\ 0)$
$D_4 = (\ 0,\ 0,\ 0,\ 0,\ 0,\ 0, 1.0, 0.6, 0.4, 0.7, 1.2, 0.8)$
$D_5 = (1.0, 2.0,\ 0,\ 0,\ 0,\ 0, 0.8, 0.5, 0.4, 0.8, 1.0, 2.0)$
$D_6 = (\ 0,\ 0,\ 0,\ 0, 3.0,\ 0, 0.7, 0.4, 0.5, 0.9, 0.8, 1.0)$

表 3　LWC を用いたクラスタリング結果

クラスタ 1	D_1, D_2, D_3
クラスタ 2	D_4, D_5, D_6

表 4　K-means を用いたクラスタリング結果

クラスタ 1	D_2, D_3
クラスタ 2	D_1, D_4, D_5, D_6

$K = 2$ で実験を行った．表 2 に対して，*Locally Weighted Clustering* を用いてクラスタリングした結果を表 3 に，K-means 法を用いてクラスタリングした結果を表 4 に示す．

表 2 に示すデータは,D_1, D_2, D_3 と D_4, D_5, D_6 がそれぞれ局所構造を持っているものである．表 3 によると，これらのデータに対して LWC 法を適用した際には，この局所構造をクラスタに反映することが可能である．一方，K-means 法を適用した際には，そのほかの一部の重みの高い要素に引っ張られ，局所構造がうまく反映されていないことがわかる．

2.4　トピック抽出

ブログ記事のタイトルと本文それぞれに対して形態素解析を行う．形態素解析ツールとしては，Sen [16] を用いる．Sen によって形態素解析を行った結果から，名詞・動詞・形容詞に加えて，名詞による複合語・新語を抽出する．名詞による複合語及び新語については後述する．抽出した語を用いて，文書ベクトルを生成する．ここで，文書ベクトル作成に用いる語には，非自立語，接尾語，数，代名詞を除くものとする．生成した文書ベクトルに基づいて，ブログ記事のクラスタリングを行う．得られたクラスタをトピックとし，トピック中から話題として扱えそうなトピックのみを，クラスタに含まれる記事数によって，自動的に選択することによって，ブログ記事群からトピックを抽出する．

2.4.1　複合語の抽出

連続して出現している名詞はもともと複合語である名詞が，形態素解析により，分割されたものと考えられるため，これらを結合し 1 つの名詞とみなした語もベクトル中に登録することとする．ブログ中においては，口語体のような記述がなされている記事や，句読点が十分に付加されていないような記事も多く存在するため，複合語以外にも名詞が連続して出現する場合がある．名詞が連続するすべての場合に結合を行うと，本来ベクトル中に登録したい，記事中に現れる特徴的な複合語以外にも，不必要に多くの語数が登録されることとなってしまう．結合を行う場合は，続いて現れる名詞が副詞可能，非自立語，数，代名詞でない場合のみに行うこととする．結合を行う語のうち，接尾語に関しては，直前に出現している名詞に結合する場合のみ登録することとし，単独での登録は行わないようにする．

また，Sen では人名は**姓名詞**と**名名詞**に分割されるが，**姓名詞**と**名名詞**が連続して出現するような場合には，一人の人物の姓および名であると考えられる．そのため，姓と名を結合した**人名名詞**も，**姓名詞・名名詞**とあわせて，文書ベクトル中に登録することとする．

表 5　抽出された未登録語

抽出新語	ドキュメント数
ブログ	4682
DVD	529
キャラ	484
CD	451
イイ	517
orz	486
プレイ	452
ヤバ	426
アレ	404

2.4.2　辞書未登語の抽出

ブログ記事中における特徴として，口語的表現が多いことや，新しく広まってきた語が多く用いられることが挙げられる．このような特徴により，ブログからのトピック抽出を行う際には，そのトピックの特徴的な語を，正しく抽出する必要性がある．したがって，形態素解析を行う際の辞書に登録されていないような語句を抽出することが必要となってくる．

本研究では，このような辞書未登録語を「未登録語」と呼ぶこととする．未登録語は，Sen では形態素解析の際に，未知語として出力される．よって，未知語として出力されたものを未登録語として抽出することとする．しかし，未知語として出力されるものの中には，形態素解析に失敗したものや，語尾を表すもの，顔文字の一部などが含まれる．よって，単純に未知語として出力されたものをそのまま未登録語として抽出してしまうと，不必要な語句が多く抽出されてしまうことになる．本手法では，トピックを表現する特徴的な語としての未登録語を抽出することを目的としているので，未知語として判定されたもののうち，未登録語として扱えそうなもののみを抽出することとする．以後，未知語として判定された語を未登録語候補語と呼ぶこととする．

名詞による複合語と同様に，連続して出現した未知語は辞書未登録語が分割されたものと考えられるため，これらを 1 つに結合したものも未登録語候補語に含めることとする．未登録語候補語のうち，1 文字からなる語句は未登録語として抽出しないこととする．また，2 文字以上からなる語句のうち，ひらがな・カタカナ・英字のみからなる語句のみを未登録語として抽出することとする．このように未登録語として抽出する文字の種類を限定することにより，ブログ記事に多く出現する，顔文字の一部などを除去することが可能であると考えられる．

予備的実験として, クローラで収集したブログ記事 92,988 件 (2007 年 11 月 6 日〜2007 年 12 月 16 日) に対して新語の抽出を行った．抽出した語の一部を表 5 に示す．

2.4.3 多視点トピックの生成

ブログ記事ごとの文書ベクトルの各要素には，一般的な TF-IDF に，各品詞ごとに重みづけを行ったものを用いる．名詞のみに重み付け，動詞のみに重み付け，形容詞のみに重み付けをそれぞれ行った文書ベクトルを生成する．あるブログ記事 E における語句 t の重み w_E^t は式 8 によって求める．

$$w_E^t = posw(t) \times \frac{\log(tf(t,E)+1)}{\log(M)} \times \log(\frac{N}{df(t)})$$
$$posw(t) = \begin{cases} \alpha, & (pos(t) = noun) \\ 1.0, & (pos(t) != noun) \end{cases} \quad (8)$$

ここで，α は各品詞ごとに行う重みづけの値を表す．$tf(t,E)$ はブログ記事 E 中に単語 t が出現する頻度，$df(t)$ は実験に用いた全ブログ記事中において単語 t が出現しているブログ記事数，N は実験に用いたブログ記事の総数，M はブログ記事 E より抽出された単語の種類数を示す．

2.4.4 記事のクラスタリング

本手法では，クラスタリングの手法として，Cheng らの提案する局所構造をより捉える事の可能なクラスタリング手法である *Locally Weighted Clustering*(LWC) [15] 法を用いる．LWC は K-means 手法を拡張した手法で，各クラスタに対して，重みベクトルを生成し，局所的な構造をより抽出することが可能となる手法である．クラスタリングの際の初期値の決定手法としては，*Subset Furthest First*(SFF) [17] 法を用いる．作成した文書ベクトル群に対して，LWC を用いて行う．文書ベクトル群を，あらかじめ決定しておいたクラスタ数 K に分割することにより，トピック抽出を行う．

本研究では，計算量の軽減のため，クラスタリングの際にはベクトル中のすべての語を計算に使用するのではなく，その語の TFIDF 値及び DF 値が次に示す条件を満たすもののみとする．計算対象とする語は，TFIDF に関しては，その語の TFIDF 値が閾値 $w_{TFIDF_{min}}$ 以上のもののみとすることとし，この閾値は分布によって決定することとする．同様に，DF 値に関しては，その語の DF 値が閾値 $w_{DF_{max}}$ 以下のもののみを使用する．この値は，実験に用いた全ブログ記事数の 3 分の 1 以下とする．

3. 実　　験

3.1 各種パラメータの設定

実験に際して，使用する各種パラメータの設定を行う．本実験において，使用するパラメータを次のように設定する．

トピックの抽出において，クラスタリングの際に使用する語句の閾値としては，0.4 とした．多視点トピック抽出のための，各品詞ごとの重み付け α は 2 とした．LWC を行う際のクラスタ数として，$K = 5$ とした．

3.2 実験に用いるデータセット

本実験では，データセットとして，クローラで収集したブログ記事 92,988 件 (2007 年 11 月 6 日〜2007 年 12 月 16 日) を使用する．

ブログ記事の本文中に「iPhone」の語を含む記事 174 件に対して実験を行った．174 件の内訳を表 6 に示す．表 6 によると，スパム記事と判定されるものが 116 件，その他の記事が 66 件であった．スパム記事としての判定は，機械的に生成されたような記事かどうかを，人手にて行った．スパム記事として判断したものに含まれていたものは，他のブログ記事や Web ページの一部の引用を自動的に取得して，記事を生成している “引用スパム”，文章をフレーズ単位で機械的に組み合わせて生成している “ワードサラダ” の 2 種類のスパム記事が含まれていた．また，スパム以外と判定されたブログ記事の内訳としては，iPhone について深く言及されている記事 29 件，iPodTouch について言及されている記事 7 件，新しい iPod について言及されている記事 6 件，広く携帯一般について言及されている記事 16 件，日記などの記事が 8 件であった．非スパムとして判定されたブログ記事のうち，日記の記事を除いた 58 件について，トピック抽出を行った．

表 6 「iPhone」を含むブログ記事 (174 件) の内訳

非スパム記事	iPhone について深く言及	29
	iPodTouch について言及	7
	新型 iPod について言及	6
	日記	8
	その他	16
	合計	66
スパム記事	引用スパム	80
	ワードサラダ	3
	その他	25
	合計	108

3.3 多様な視点を考慮したトピック抽出

実験によって得られたトピックについて表 7，表 8，表 9 に示す．

表 7 によると，名詞に重み付けを行った場合はそのキーワードのものは「携帯」や「電話」などが示すような “何なのか” ということや,「ソフト」や「アプリ」などが示す “何を持っているのか” などということが捉えやすいトピックが抽出されることがわかった．しかし，1 つのクラスタに偏りやすく，クラスタ内に関連性があまり強くないものも含まれてしまうことがわかった．一方，表 8 によると，動詞に重み付けを行った場合では，“何ができるのか”，“何をするのか，したいのか” などが捉えやすいトピックが抽出されることがわかった．また，動詞に重み付けを行った際には,「触る」や「持つ」などの主にキーワードが示す機器やその機能に対してユーザが行う行動を示している動詞と,「参入する」や「発表する」などのキーワードが示す機器が行われる行動を示している動詞に応じてトピックが分割されることがわかった．記事の投稿時間などを用いることにより，“いつ” の時期に “何を行ったか” がわかると考えられる．表 9 の形容詞に重み付けを行った場合では，“何が”，“どうなのか” などが捉えやすいトピックが抽出されることがわかった．これは,“どうなのか” という感想や使い勝手に関する記述は，それ単体ではなく “何が” や “何の” といった事柄と結びついていることに起因していると考えられる．“どうなっている

か” ということを表す語は，名詞に重み付けを行った場合や動詞に重み付けを行った場合にはあまり上位の語として抽出されていなかった．

各重み付けで得られたクラスタのうち，記事数の多い上位 3 つのクラスタについて，出現語句について分類を行ったものを図 1 に示す．得られたクラスタにおいて，代表的なブログ記事の文章を表 10 に示す．図 1 によると，名詞に重み付けを行った場合には，全体的に機種やスペックに関することが重みの高い上位の語句として出現していることが分かる．それに対して，動詞に重み付けを行った場合には，機種やスペックに関することはあまり多くは出現しておらず，代わりにユーザなどが行う動作に関することが出現していることが分かる．また，形容詞に重み付けを行った場合では，機種やスペックに関することとともに，感想や評価に関する語も多く出現していることが分かる．

3.4 K-means 法との比較

実際のブログ記事に対して，K-means 法を用いてクラスタリングした結果と LWC 法を用いてクラスタリングした結果の比較を行う．クラスタリングで得られた結果は複雑であり，一概にどちらが良いものか判定しにくいことが多い．事前にクラスタリング結果の正解がわかっているような場合，F-measure [18] [19] や正規化相互情報量 (NMI:Normalized Mutual Information) [15] [20] などを評価指標として利用することが可能である．しかし，ブログ記事の実データの多様性により，事前に正解となるようなクラスタリング結果を決定することは困難であると考えられる．本実験ではクラスタの評価指標として，K-means 法の評価関数を用いた．K-means 法はこの評価関数の値を最小化するようにクラスタを分割していく手法であるので，評価関数によって算出される値が小さいものが，より適切なクラスタであると考えられる．本実験で用いた評価関数を式 9 に示す．

$$Eval = \sum_{i=1}^{k} \sum_{x \in C_i} D(x, c_i)$$
$$\text{ここで, } D(x, c_i) = \sqrt{\sum_{j=1}^{m} |x_j - c_{kj}|^2} \tag{9}$$

ここで，k はクラスタ数，m はデータの次元数を示す．c_i は i 番目のクラスタの重心を示す．

「iPhone」の語句を含むブログ記事 174 件に対して，K-means 法，LWC 法でクラスタリングした結果から算出した評価値を表 11 に示す．初期値による影響を考慮し，5 回試行した結果の平均を算出した．

表 11 によると，どの重み付けの場合でも，比較的 LWC の方が値が小さくなっていることが分かる．このことから，実際のブログ記事に適用した際にも，K-means 法よりはある程度適切なクラスタが得られる場合があることがわかった．値の差の大きさが必ずしもクラスタリング結果の質の差を表しているわけではないが，表 11 で得られた差はそれほど大きいものではないため，クラスタリング結果のさらに精密な評価は今後の課題である．

表 11　ブログ記事に対する K-means 法と LWC 法の比較結果

	LWC			K-means		
	名詞	動詞	形容詞	名詞	動詞	形容詞
1 回目	74.237	60.995	57.917	75.729	62.587	57.483
2 回目	74.595	61.642	54.694	76.456	63.014	60.249
3 回目	74.508	62.356	57.119	76.367	63.645	58.973
4 回目	76.822	62.356	57.119	77.090	63.014	60.250
5 回目	75.948	62.356	56.174	77.141	63.718	58.174
平均	75.222	61.941	56.603	76.556	63.196	59.026

4. 関 連 研 究

トピック抽出に関する研究としては, 以下のようなものが挙げられる．Allan ら [21] は, ニュースから話題を自動的に抽出, 追跡することを目的としている．本研究は，ブログ記事を対象とすることで，ニュースとは異なり，人の意見や感情が多く含まれていると考えられる．そのため，ニュース記事では一様に扱えたものが，対象をブログ記事にすることにより多様性が含まれるため，本研究とは異なっている．Shirberg ら [22] は, 人の発話を，意味のあるまとまりごとに分割するために，音声からのトピック抽出，追跡を行う手法を提案している．本研究では，意味のあるまとまりごとに分割することを目的としているのではなく，分割の方法を複数にすることにより，多様な分割を行うことを目的としている点で異なっている．Castellanos ら [23] は, カスタマーサポートセンターのログのようなタイプミスや省略, 特殊記号などの含まれている, ノイズの多いデータよりトピック検出を行う試みを行っている．Castellanos ら [23] は，話題となっているトピック (Hot Topic) を検出することを目的としており，ユーザの目的が多様であることを想定していないという点で本研究とは異なる．

トピック抽出手法としては，burst の検出によるものが挙げられる．手法 [24] 及び [25] では，ある語に対し，その語が出現する時間間隔の定常状態を求めておき，その時間間隔よりも短い間隔で語が出現しているとき，その語をトピックに関連する語として抽出する．手法 [24] 及び [25] では，急激に話題になったようなトピックの抽出を目的としたものであり，ほどんど変化なく取り扱われているトピックに対しては，うまく抽出できない．本研究では，急激に話題になったようなトピックだけではなく，ほとんど変化なく扱われているようなトピックに対しても，多視点からのトピックを抽出することを目標としているため，手法 [24] 及び [25] では本研究の目的に合わない．

5. お わ り に

本論文では，ブログ記事における品詞の出現頻度差異と，Locally Weighted Clustering を利用した，ユーザの意図を意識した多様な視点からのトピック抽出手法を提案した．予備的実験により，ノイズになる記事を除去した際に，上手く働くことを示した．また，K-means 法との簡単な比較により，LWC 法の方が本研究の目的に沿ったクラスタが生成されることを示

表 7　名詞に重み付け (iPhone)

クラスタ	記事数	重みの高い上位 5 語
クラスタ A	28	iPhone，機能，電話，iPod，iPodtouch，携帯，画面，ソフト，世代，発表
クラスタ B	11	携帯，iPhone，端末，回線，情報端末，アプリケーション，アップル，機能，スマートフォン，参入
クラスタ C	7	iPhone，日本語，動画，電話，機種，セッティング，アドレス，文字化け，網膜，表示
クラスタ D	5	iPhone，音楽，メール，アプリ，本体スイッチ，タン，バッテリー，感動，スピーカー，経由
クラスタ E	5	アップル，iPhone，製品，電話，携帯，発売，参入，通信，機種，操作

表 8　動詞に重み付け (iPhone)

クラスタ	記事数	重みの高い上位 5 語
クラスタ F	15	iPhone，携帯，間借りする，参入する，参入，触る，回線，設定する，云う，感動する
クラスタ G	12	iPhone，iPodtouch，触る，使う，付く，機能，動画，iPod，縛る，電話
クラスタ H	12	iPhone，iPod，使う，買う，使える，聴く，いう，機能，電話，Apple
クラスタ I	12	iPhone，iPod，持つ，BlogPeople，機能，ズワイガニズワイガニプラン，発表する，電話，要る，ケータイ
クラスタ J	5	iPhone，使える，売る，狙う，モバイル，電話，誘いあう，登場する，携帯，参照

表 9　形容詞に重み付け (iPhone)

クラスタ	記事数	重みの高い上位 5 語
クラスタ K	21	iPhone，携帯，新しい，Apple，使える，日本語，電話，表示，画面，機種
クラスタ L	20	iPhone，iPod，機能，電話，欲しい，新しい，音楽，携帯，メール，BlogPeople
クラスタ M	10	iPhone，iPodtouch，機能，ケータイビジネス，触る，Apple，動画，電話，ポータブルプレイヤー，クール
クラスタ N	3	iPhone，回線，アップル，参入，イーモバイル，間借りする，間借り，ない，手口，借り
クラスタ O	2	iPhone，参照，Gphone，モバイル，市場，Google，さみしい，ビジネススタイル，プラットフォーム，電話

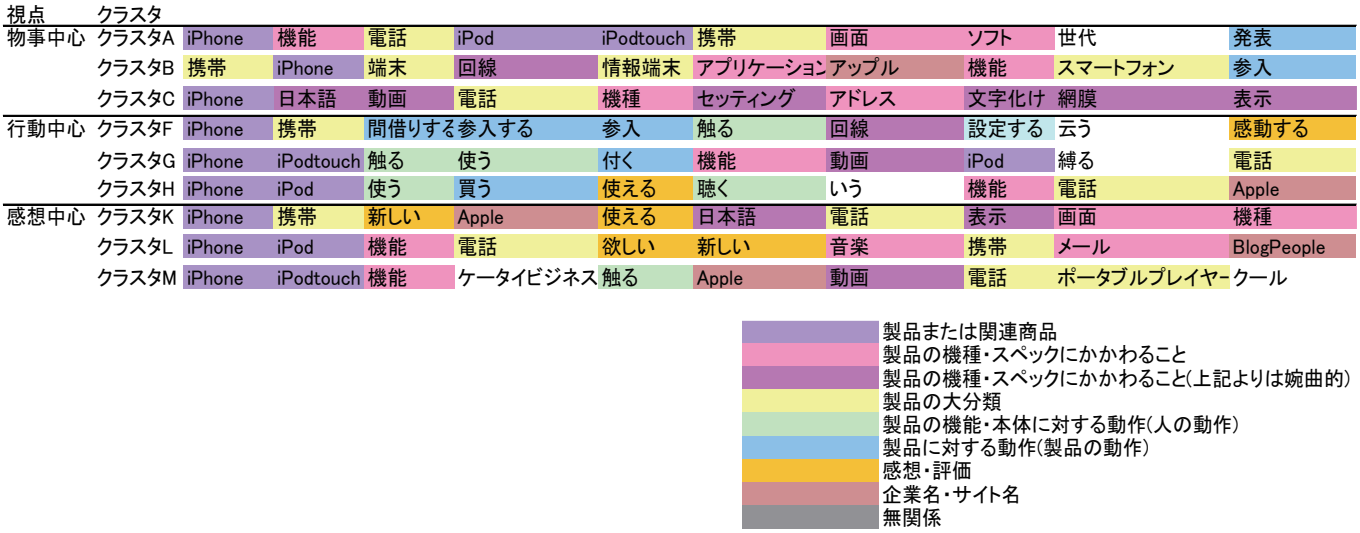

図 1　各重み付けごとの上位語句に関する分析

した．

今後の課題としては，品詞情報に直結しないような視点の検討などが挙げられる．新しい視点の導入やユーザによる視点の調整・特殊化を実現する方法についても，検討する必要がある．また，K-means 法のクラスタ数を自動で決定する手法などを導入し，利用時のユーザビリティを向上させることも重要である．

そのほかの課題としては，ある重み付けして抽出したトピックに対し，別の重み付けを行ったトピック抽出を行い，サブトピックを抽出することなどが考えられる．例えば，物事・属性中心に抽出したあるトピックに対して，感想中心のトピック抽出を行うことにより，どの属性に対して，どのような感想が含まれるのかを，サブトピックとして抽出できる可能性がある．

また，アプリケーション化への課題の 1 つとして，クラスタリング結果をどのように提示するかなども検討していく必要がある．

また，ユーザの多様な視点を意識して抽出したトピックに対して，その変遷を抽出し，実世界との関連性を抽出することも挙げられる．実世界をどのように捉えるかなどは今後の検討事項である．

謝辞　本研究の一部は科学研究費補助金基盤研究（B）（課題番号 19300026）及び，科学研究費補助金特定領域研究（課題番号 19024035）の助成による．

文　献

[1]　Nifty. ブログサイトに関する共同研究調査. http://www.nifty.

表 10　代表的なブログ記事例 (一部抜粋)

視点	代表的なブログ記事例 (一部抜粋)
物事中心	...3.5 インチの画面サイズだと十分動画を鑑賞することが可能だ。 基本的に自分の通勤環境が、電車で約 30 分なので 30 分番組を 1 本見るのにちょうどいい。 16GB だと 30 分のテレビ番組のシリーズを 2 セット（50 話分）ぐらい入れて、まだ、余裕だ。... ... オペレーティングシステム、ミドルウエア、主要なモバイルアプリケーションを含む、 モバイルデバイス向けの完全なソフトウエアセットです。...
行動中心 、	... それでも新規参入というのか？ ディズニーがソフトバンクの回線で携帯事業に参入するとなると... ...11 月 9 日にドイツとイギリスで iPhone が発売されたそうです 特にドイツでは風雨の中数百人の Apple ファンが行列を作るほどの人気だったそうです... ... 用も無いのに電気屋さんに行って、展示されてた iPodtouch を触ってみた！ にひひめっちゃクール！星空...
感想中心	...Apple が iPhone を発表したときにコレ欲しいぃ〜DASH!と思ったはちですが iPod touch の方がいいかしらん☆なんて心変わりしちゃったりしてにひひ... ... 僕的には "iPhone" が欲しいのだけど （どっちにしてもまだ電話機能が日本では使えないのだけど）... ...「今度 Apple から新しい iPod が出たんだけどこれにしていい？」と聞いてみた。 「ふーん、何で前面は黒なんだろう」と良いのか悪いのかわからないような反応...

co.jp/tenpu/080403-tenpu.pdf.

[2] kizasi.jp. http://kizasi.jp.

[3] goo 評判分析. http://blog.search.goo.ne.jp/wpa/guide/index.html.

[4] Yahoo!ブログ検索. http://blog-search.yahoo.co.jp/.

[5] Clusty. http://clusty.com/.

[6] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pp. 171–180, New York, NY, USA, 2007. ACM.

[7] G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[8] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. Arsa: a sentiment-aware model for predicting sales performance using blogs. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 607–614, New York, NY, USA, 2007. ACM.

[9] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 78–87, New York, NY, USA, 2005. ACM.

[10] Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pp. 533–542, New York, NY, USA, 2006. ACM.

[11] 奥村学. ブログマイニング技術の最新動向. 電子情報通信学会誌, Vol. 91, No. 12, pp. 1054–1059, 2008.

[12] 戸田智子, 福田直樹, 石川博. Blog 記事のクラスタリングによるカテゴリ別話題変遷パタンの抽出. 電子情報通信学会データ工学ワークショップ DEWS2007, 2007.

[13] 戸田智子, 鎌田基之, 黒田晋矢, 福田直樹, 石川博. ブログ記事からのトピック別評判情報変遷パタンの抽出手法について (sns・blog, 夏のデータベースワークショップ 2007(データ工学, 一般)). 電子情報通信学会技術研究報告. DE, データ工学, Vol. 107, No. 131, pp. 201–206, 20070625.

[14] 戸田智子, 黒田晋矢, 福田直樹, 石川博. ブログにおける多視点からのトピック抽出手法の提案. 電子情報通信学会第 19 回データ工学ワークショップ DEWS2008, 2008.

[15] Hao Cheng, Kien A. Hua, and Khanh Vu. Constrained locally weighted clustering. *Proc. VLDB Endow.*, Vol. 1, No. 1, pp. 90–101, 2008.

[16] 形態素解析システム sen. http://ultimania.org/sen/.

[17] Douglas Turnbull and Charles Elkan. Fast recognition of musical genres using rbf networks. *IEEE Trans. on Knowl. and Data Eng.*, Vol. 17, No. 4, pp. 580–584, 2005.

[18] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 16–22, New York, NY, USA, 1999. ACM.

[19] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. KDD Workshop on Text Mining '00, 2000.

[20] Xin Zheng, Deng Cai, Xiaofei He, Wei-Ying Ma, and Xueyin Lin. Locality preserving clustering for image database. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 885–891, New York, NY, USA, 2004. ACM.

[21] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study: Final report. In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, 1998.

[22] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gükhan Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Commun.*, Vol. 32, No. 1-2, pp. 127–154, 2000.

[23] Malu Castellanos. *Survey of Text Mining*, chapter 6. HotMiner: Discovering Hot Topics from Dirty Text. Springer, 2003.

[24] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proc. the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 91–101, New York, NY, USA, 2002. ACM.

[25] 藤木稔明, 南野朋之, 鈴木泰裕, 奥村学. document stream における burst の発見 (情報抽出・データマイニング). 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2004, No. 23, pp. 85–92, 20040304.