Vol.2009-DBS-148 No.8
Vol.2009-FI-95 No.8
2009/7/28

多様な視点からのブログ記事マイニングへの制約付きクラスタリングの適用

青島傳隼[†1] 戸田智子[†1,*1] 福田直樹[†1]
横山昌平[†1] 石川博[†1]

ブログマイニングに関するアプローチの多くでは，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，ブログで注目されるトピックの変遷の抽出などの試みがされている．ここでの課題の 1 つは，ブログのマイニング時に，必要とされるトピックの切り出し方を，こうした多様な目的の違いに応じて変えなければならない点である．本論文では，ブログ記事よりトピックを抽出する際に，利用者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．視点の違いに合わせて適切にトピックの抽出を行うためには，トピック抽出過程で視点の違いを何らかの方法で反映させる必要がある．本論文では，利用者がブログを分析したいと思う際の視点の違いを，文書中に出現する特徴語となるべき語の品詞の差異，および，背景知識に基づいて明示的に指示されたブログ記事間の関係の 2 つから表現することを考える．トピック抽出のためのクラスタリング手法に，要素内の局所的な構造を反映させやすく，要素間での制約を記述できる非階層的クラスタリング手法である Constrained Locally Weighted Clustering 法を用いる．本論文では，収集したブログ記事集合に対する実験的評価，および，試作システムの実装について述べる．

Constrained Clustering for Blog Articles with Diverse Views

Tsugutoshi Aoshima,[†1] Tomoko Toda,[†1]
Naoki Fukuta,[†1] Shohei Yokoyama[†1]
and Hiroshi Ishikawa[†1]

There are many approaches about mining blogs that are focusing on characteristics that the blogs often reveal various communications and personal opinions about timely events within chronological line. Some approaches are trying to analyze shifts and moves about interests on blogs by extracted topics from them. 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. Furthermore, we try to reflect implicit relations among articles by clustering constraints. We propose a prototype system for adaptive topic extraction in various aspects from the specified blog articles by using grammatical characteristics of words that are related to the aspects. Constrained Locally Weighted Clustering is applied for better clustering with associated constraints among articles referred from certain background knowledge.

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．

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．このようなものは，ブログの分析を目的として行われている．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[1)2)]．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，Yahoo!ブログ検索[3)] や Clusty[4)] 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[5)] 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

†1 静岡大学情報学部情報科学科
Department of Computer Science, Faculty of Informatics,Shizuoka University
*1 現在，富士通テン株式会社
Presently with Institute for FUJITSU TEN LIMITED

Vol.2009-DBS-148 No.8
Vol.2009-FI-95 No.8
2009/7/28

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[6] 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[7] 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[8] 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[9] 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 10) 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によって，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，ブログの分析を行う目的が多様になってきている．ある製品について調べたいという目的に対しても，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，その製品の詳細を知りたい場合，あるいは，その製品を実際に購入した人や購入を控えた人はどの程度いるかを知りたい場合など，その意図が多様性を持つと予測される．

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．このようなトピックは，捉えたい話題について記述されているブログ記事の集合として抽出することとする．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，ブログ分析の目的の多様性にかかわらず一様なトピック抽出のみを用いた場合では，ユーザにとって有益なトピック抽出を行えない場合があると考えられる．

我々は，文献 11) において，局所性を利用したクラスタリング手法を用いて多様な視点を考慮したトピック抽出を試みた．本論文は，文献 11) の改良として，制約を利用したクラスタリング手法である Constrained Locally Weighted Clustering（CLWC）法[12] を用いて多様な視点からトピック抽出を行うことができるようなシステムを試作する．

本論文は，次のような構成になっている．2 節では，CLWC 法を用いた多様な視点からのトピック抽出手法のアプローチについて述べる．3 節では，制約付きクラスタリングの検証実験について述べる．4 節では，試作したアプリケーションについて，その概要と適用例を示す．5 節では，関連研究について述べる．最後に 6 節でまとめを述べる．

2. 多様な視点を考慮したトピック抽出

2.1 研究の目的

本節では，本研究の目的について述べる．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, 視点をそれぞれの品詞に対応付け，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．

我々は，この目的のために，文献 13) において多様な視点を考慮したトピック抽出手法を提案した．そこでは，ノイズの問題や，ユーザの背景知識の反映などが課題であった．そこで，ノイズの排除，および，計算量や処理時間改善を狙い，文献 11) において非階層型クラスタリング手法である Locally Weighted Clustering（LWC）法[12] を用いたトピック抽出手法を提案した．

本研究では，文献 11) の方法をさらに拡張し，ブログを閲覧するユーザの背景知識を反映するために，クラスタリング時に制約を用いることができるように拡張する．クラスタリングには，LWC 法の拡張であり制約を考慮したクラスタリングが可能な Constrained Locally Weighted Clustering（CLWC）法[12] を用いたトピック抽出を試みる．

LWC 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たものであり，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文献 12) で示されている．

CLWC 法では，制約を用いた場合にも効率的に LWC 法に基づくクラスタリングが行えることが同文献[12] で示されている

2.2 トピックの定義

本節では，本論文で用いるテーマ，トピック，視点の定義について述べる．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

Vol.2009-DBS-148 No.8
Vol.2009-FI-95 No.8
2009/7/28

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．トピックとは，その視点による集合の分割を示す．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．$tp_1^a, \ldots, tp_n^a$ は，視点 a において抽出されたトピック，$tp_1^b, \ldots, tp_m^b$ は，視点 b において抽出されたトピックを示す．

2.3 CLWC（Constrained Locally Weighted Clustering）法

2.3.1 アルゴリズム

$\mathcal{R}^m$ 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．

$$\mathcal{L}_{2,\vec{w_k}}(\vec{x}, \vec{c_k}) = \frac{w_{k1}x_1 \cdot w_{k1}c_{k1} + \ldots + w_{km}x_m \cdot w_{km}c_{km}}{|\vec{x}| \cdot |\vec{c}_k|} \tag{2}$$

$$\begin{aligned} \text{ここで，} |\vec{x}| &= \sqrt{(w_{k1}x_1)^2 + \ldots + (w_{km}x_m)^2} \\ |\vec{c}_k| &= \sqrt{(w_{k1}c_{k1})^2 + \ldots + (w_{km}c_{km})^2} \end{aligned}$$

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．

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

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．

$$\sum_{i=1}^{N} \mathcal{L}^2_{2,\vec{w}_{\phi_c(\vec{x}_i)}}(\vec{x}_i, \vec{c}_{\phi_c(\vec{x}_i)}) \tag{5}$$

$$\text{このとき,} \forall k \prod_{j=1}^{m} w_{kj} = 1$$

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．$1 \le k \le K$，$1 \le j \le m$ とする．

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

$$w_{kj} = \frac{\lambda_k}{\sum_{\vec{x} \in C_k} |x_j - c_{kj}|^2} \tag{7}$$

$$\text{このとき,} \lambda_k = \left(\prod_{j=1}^{m} \left(\sum_{\vec{x} \in C_k} |x_j - c_{kj}|^2\right)\right)^{\frac{1}{m}}$$

CLWC 法では，制約を付与した後に，must-link を用いてチャンクレットと呼ばれる同じクラスタに属することを保証された小文書群を用いてクラスタリングを行う．全てのデータは，クラスタリングの処理に入る前の段階で全て大きさが 1 のチャンクレットとし，must-link で繋がれたチャンクレット同士を，全ての must-link を満たすまで合わせ続ける．ただし，このとき必ず cannot-link を考慮し，cannot-link で繋がれたデータペアを同じチャンクレットに属させないようにする．制約についての具体的な処理は以下のように行う．

- `must-link`：この制約で結ばれた 2 つの記事は同じクラスタに属させる．
- `cannot-link`：この制約で結ばれた 2 つの記事は異なるクラスタに属させる．

CLWC 法のアルゴリズムを**表 1** にまとめる．

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[14] 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

Vol.2009-DBS-148 No.8
Vol.2009-FI-95 No.8
2009/7/28

表 1 CLWC アルゴリズム
Table 1 CLWC algorithm

アルゴリズム Constrained Locally Weighted Clustering(CLWC)
Require： $\vec{x}_i \in \mathcal{R}^m$，クラスタ数 K
Ensure： クラスタ K の重心 $\vec{c_k}$，クラスタ内の語 t_j の重み $\vec{w_{kj}}$
1:全てのデータをサイズ 1 のチャンクレットとし， 制約をもとにチャンクレット同士を合わせる．
2: クラスタの初期重心を決定する． クラスタ内のすべての語の重みを 1 にする．
3: E-Step1: 制約を全く持たないチャンクレットを一番近いクラスタに割り当てる．
4: E-Step2: 次に大きいチャンクレットから順に同様にクラスタに割り当て， もしそのチャンクレットの近傍にチャンクレットがあった場合は同時に割り当てる．
5: M-Step: それぞれのクラスタに対して式（4）により，重心を再計算し， 重み $\vec{wk}$ を更新する．
6: 収束するまで 3〜5 を繰り返す．

加えて，名詞による複合語・未登録語を抽出する．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し，複合語および未登録語の抽出は文献 11) の手法と同様とする．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て抽出する．

2.4.1 多様な視点を考慮したトピック抽出

本節では，多様な視点を考慮したトピックの抽出のために，品詞ごとに重み付けを行った文書ベクトルを生成する．生成した文書ベクトルに対し，CLWC 法を用いてクラスタリングを行い，トピックを抽出する．ブログ記事ごとの文書ベクトルの各要素には，一般的な TFIDF 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付けの値を表す．$tf(t,E)$ 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．

本手法では，ブログ記事群へ多様な視点からの制約付きクラスタリングを行うにあたって，CLWC 法を用いる．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）[15] 法を用いる．作成した文書ベクトル群に対して，CLWC 法を用いて行う．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 $w_{TFIDF_{min}}$ 以上のもののみとし，この閾値は分布によって決定する．

3. 実　　験

3.1 本実験の目的

本手法によって得られたトピックの詳細を調べ，トピック抽出結果の妥当性の評価を行う．

3.2 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を 0.1 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．CLWC 法を用いる際のクラスタ数 k は，$k = 5$ とした．

Wagstaff らは，文献 16) において，あらかじめデータに付与されたラベルをランダムに比較し制約を与えることで，クラスタリング精度が向上することを示している．今回用いるブログ記事にはラベルが前もって付与された状態ではないため，今回の実験では，あからじめ重みの高い語を上位 5 つまで各文書から抽出し，これを疑似ラベル群として考え Wagstaff らと同様の手法を用いることにした．本実験では文書ペア内で複数のラベルを比較した結果を用いて制約を付与した．文書ペアをランダムに x 回選択し，それらの抽出語を比較し，1 つ以上同じ抽出語を持つ場合は must-link の制約を与え，1 つも同じ抽出語を持たない場合には cannot-link の制約を与えた．文書ペアの比較回数 x の値は $x = 5$ とした．

チャンクレットの近傍の定義は，本実験では，チャンクレット間の類似度が閾値 0.3 以上の場合とした．

Vol.2009-DBS-148 No.8
Vol.2009-FI-95 No.8
2009/7/28

表 2 「iPhone」を含むブログ記事（174 件）の内訳
Table 2 Summary of obtained blog contents for iPhone.

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	10
	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 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 件のうち，スパム記事等を除いた 56 件で実験を行った．174 件の内訳を**表 2** に示す．表 2 によると，スパム記事と判定されるものが 108 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，人手で行った．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，新型 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 10 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 56 件について，トピック抽出を行った．

3.4 トピック表現語の抽出

クラスタ C_i の記事群 E の中に含まれる語句 t の重み $weight(t)$ を式（9）に従って計算し，上位 5 語をそのクラスタのトピック表現語として抽出した．

$$weight(t) = \frac{1}{|C_i|} \sum_{E_j \in C_i} w^t_{E_j} \tag{9}$$

また，同時に記事群 E の中で，出現記事数が閾値以上の，トピック表現語として抽出されていない名詞を抽出した．本手法では，出現記事数がクラスタ C_i に属する記事数の 3 分の 1 以上の名詞を最大 5 個まで抽出し，そのクラスタのトピックとの関連性を調べる．

3.5 抽出されたトピックの考察

実験によって得られたトピックとその表現語について**表 3**，**表 4**，**表 5** に示す．

クラスタに含まれる記事数は，視点によって若干差はあるが，7 件から 19 件の間で分布した．語数の少ない記事の語句がどの視点においてもトピック表現語として抽出される傾向にあり，必ずしもそのクラスタのトピックとはいえないようなものも抽出された．

物事視点では，キーワード (本手法では「iPhone」) が，“何なのか” ということや，“どんな機能を携えているのか” ということを捉えやすいトピックが抽出された．また行動視点では，キーワードが，“何ができるのか”，“何をしたのか，されたのか” を捉えやすいトピックが抽出され，感想視点ではキーワードが，“どうなのか” といった感想を捉えることが可能なトピックが抽出された．その一方で，行動視点や感想視点のみのトピックだけでは，主語となるような名詞がうまく抽出されないため，クラスタ内の記事を具体的に見なければ，トピックの対象が理解できない場合があった．

3.5.1 物 事 視 点

表 3 は物事視点からクラスタリングを行った結果である．抽出されたトピックやその表現語の傾向としては,「携帯」や「携帯電話」といった iPhone の大まかな分類を指すようなトピック表現語や,「iPodtouch」や「ソフト」「動画」といった関連商品であったり，iPhone の持つ機能を捉えることが可能なトピック表現語が抽出された．

また，クラスタ E の結果からは，iPhone 登場後に一定期間過ぎてから関連するようになった企業名がトピック表現語として抽出できたが，Apple のように関連の密な企業はトピック表現語としてはどのクラスタにおいても抽出されなかった．これは関連性が強すぎるためにどの記事でも頻出し，重みが相対的に低くなっていることが原因として考えられる．

3.5.2 行 動 視 点

表 4 に行動視点からクラスタリングを行った結果を示す．トピック表現語の傾向として,「触る」,「設定する」といったユーザ側が iPhone に対して行う動詞と,「でる」,「売る」といった iPhone が提供側から行われたような動詞の 2 種類のトピック表現語が抽出される傾向にあった．

行動視点では，名詞がほとんどトピック表現語として抽出されないため，クラスタ内で頻度の高い名詞を用いて，抽出されたトピック表現語と関連付ければ，トピックの分析にかかる負担をある程度軽減できると考えられる．

Vol.2009-DBS-148 No.8
Vol.2009-FI-95 No.8
2009/7/28

表 3 物事視点から見た場合のクラスタ
Table 3 Clusters applies focus to noun.

クラスタ	記事数	トピック表現語 (頻度の高い名詞)
クラスタ A	18	ソフト，iPodtouch，事，海外，iPhone (日本，Apple，iPod)
クラスタ B	14	動画，携帯，今日，ここ，iPodtouch (iPhone，日本)
クラスタ C	9	ケータイ，表示，天気，株，地図 (iPhone，iPod，メール，iPodtouch，画面)
クラスタ D	8	日本語，イエス，ノー，質問，BlackBerry (iPhone，OK)
クラスタ E	8	ディズニー，回線，ディズニーキャラクター，携帯電話，独自 (iPhone，日本，今，ソフトバンク，ドコモ)

表 4 行動視点から見た場合のクラスタ
Table 4 Clusters applies focus to verb.

クラスタ	記事数	トピック表現語 (頻度の高い名詞)
クラスタ F	19	できる，使う，出る，入れる，売る (iPhone，日本，iPodtouch，iPod)
クラスタ G	10	でる，ついてる，使う，回線，十分元取れる (iPhone，ソフトバンク，Apple，iPod)
クラスタ H	9	設定する，iPhone 用アプリ入る，云う，感動する，ソフト (iPhone，日本，iPodtouch，今)
クラスタ I	9	触る，もらう，帰る，コメントする，iPhone 触る (iPhone，iPodtouch，ブログ，何)
クラスタ J	9	持つ，ケータイ，見る，言う，wi-fi (iPhone，iPod，日本，目，携帯電話)

表 5 感想視点から見た場合のクラスタ
Table 5 Clusters applies focus to adjective.

クラスタ	記事数	トピック表現語 (頻度の高い名詞)
クラスタ K	18	感動する、回線、ディズニーキャラクター、ディズニー、表示 (iPhone、今、iPodtouch、日本)
クラスタ L	13	ない、いい、コメントする、イエス、ノー (iPhone、iPod)
クラスタ M	9	欲しい、早い、楽しい、iPodtouch、触る (iPhone、日本、Apple、話、発売)
クラスタ N	8	日本語、速い、ほしい、賢い、良い (iPhone、機能、アメリカ、日本)
クラスタ O	7	楽しい、新しい、厳しい、ディズニー、携帯電話事業 (iPhone、音楽、日本、iPodtouch)

3.5.3 感 想 視 点

表 5 に感想視点からクラスタリングを行った結果を示す．感想視点でトピック表現語を抽出した場合には,「楽しい」や「良い」といった使い勝手や感想の形容詞が抽出された一方で,「早い」のような，それ単独では何に関しての感想なのか判断の難しいものも抽出された．行動視点と同様に名詞はトピック表現語としてほとんど抽出されないため，トピック表現語のみでは何の感想であるのかを概観しづらいトピックも存在した．

4. アプリケーションの試作

本手法を用い，クラスタリング結果を確認しながら，ユーザが制約を与えたり，視点ごとの重みを動的に変えられるようにしたアプリケーションを試作した．試作アプリケーションの実装には，Java 言語を用いた．

4.1 制約の生成

記事間への制約の付与は，クラスタリングを行う前，クラスタリングの実行中，クラスタリング結果の表示中のいずれでも行えるようになっている．ユーザは，記事を直接選択してその間に制約を付与することができ，ブログ記事のタイトル，内容，およびクラスタリング結果を用いて，特定の条件を満たすもの同士にまとめて制約を付与することが可能である．

制約は，クラスタ表示ウィンドウ上に表示された記事を表すアイコンから他のアイコンへ線を引くことで付与できる．制約で繋がれた記事間には，must-link である場合には緑のリンクが，cannot-link の場合には赤いリンクが表示される．

制約を削除するためには，再び同頂点間に線を引くことで削除することが可能である．

また，ユーザによって制約が付与される場合，**図 1** のように暗黙的な制約関係との矛盾を含む制約が生成される場合がある．具体的には，以下のような場合である．

- `Case 1`：暗黙的な must-link の関係にあるデータ間に cannot-link を引く場合
- `Case 2`：暗黙的な cannot-link の関係にあるデータ間に must-link を引く場合

本試作システムのユーザインタフェース上では，これらのような矛盾を含む制約を付与できないようにしていてる．

4.2 動的な視点の変更

我々は，文献 13) および 11) において，ユーザの多様な視点を考慮するために，前もってブログ記事群で登場する語句の対応した品詞に重みを加えてクラスタリングを実行し，トピック抽出を試みた．本論文では，クラスタリングの実行中および実行後での結果を見てからの視点の重み付けを変更する機能を実装した．動的な視点の変更を可能にすることで，クラスタリング結果をユーザの望む視点を考慮したものにすることができ，クラスタリング結果からよりキーワードに関連する内容を得ることができるようになると考えられる．文献

Vol.2009-DBS-148 No.8
Vol.2009-FI-95 No.8
2009/7/28

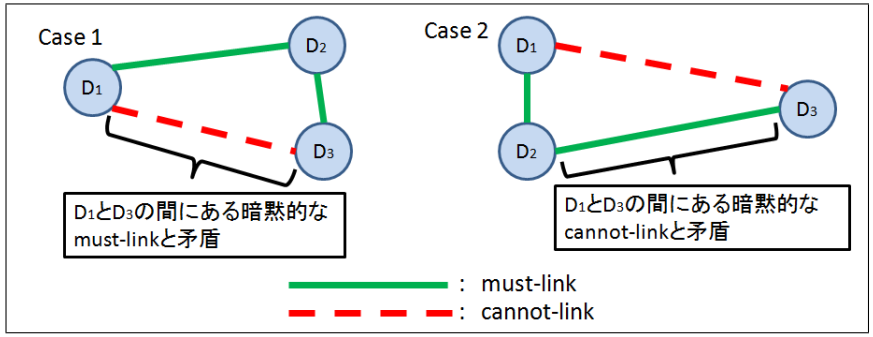

図 1 制約間の矛盾の例
Fig. 1 Conflicts among constraints.

11) では，特に 1 つの品詞に対して多くの重み付けを行うことを中心に扱ったが，本論文では，それらの重み付けを，対象とする記事集合の特性に合わせて，一定の範囲内で動的に変更できるようにした．

本実装では，**図 2** のようにクラスタリング時におけるそれぞれの視点の重み付けをスライドバーを用いて変更できるようにした．クラスタリング結果を提示する画面のスクリーンショットを**図 3** に示す．図 3 左は制約を与えずにクラスタリングを行った結果で，その結果に制約を動的に与えてクラスタリングすると図 3 右のようになった．

5. 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[17] 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[18] 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[19] 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

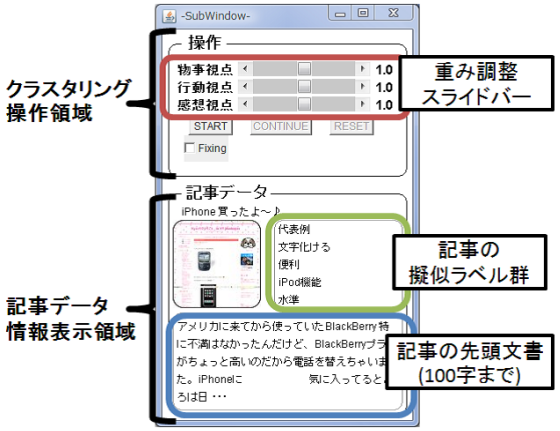

図 2 記事詳細表示と視点の重み付け調整
Fig. 2 A property window with weight-adjustments for each viewpoint.

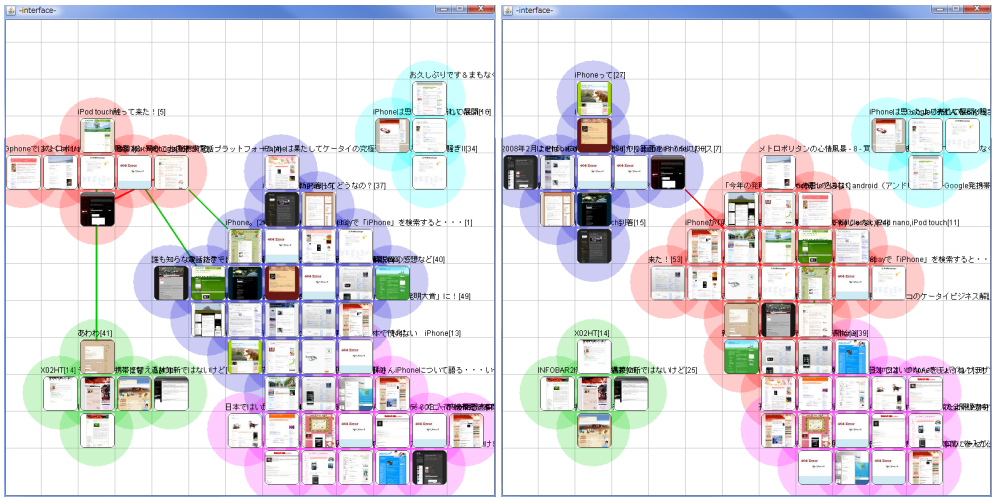

図 3 制約の追加によるクラスタリング結果の変化
Fig. 3 Dynamic constraints and clustering results.

Vol.2009-DBS-148 No.8
Vol.2009-FI-95 No.8
2009/7/28

る．Castellanos ら[19] 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 20) 及び 21) 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 20) 及び 21) 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 20) 及び 21) では本研究の目的に合わない．

6. 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，Constrained Locally Weighted Clustering を利用したトピック抽出を行い．ユーザの意図をクラスタリングに反映するためのアプリケーションの試作を行った．クラスタリング処理の開始時に制約を与えることで，制約を反映したクラスタリングを行うことができることを確認した．

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）の助成による．

参考文献

1) : kizasi.jp, http://kizasi.jp.
2) : goo 評判分析，http://blog.search.goo.ne.jp/wpa/guide/index.html.
3) : Yahoo!ブログ検索，http://blog-search.yahoo.co.jp/.
4) : Clusty, http://clusty.com/.
5) Mei, Q., Ling, X., Wondra, M., Su, H. and Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs, *Proc. the 16th international conference on World Wide Web(WWW '07)*, New York, NY, USA, ACM, pp.171–180 (2007).
6) Mishne, G. and Glance, N.: Predicting Movie Sales from Blogger Sentiment, *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs* (2006).
7) Liu, Y., Huang, X., An, A. and Yu, X.: ARSA: a sentiment-aware model for predicting sales performance using blogs, *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, ACM, pp.607–614 (2007).
8) Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A.: The predictive power of online chatter, *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, New York, NY, USA, ACM, pp. 78–87 (2005).
9) Mei, Q., Liu, C., Su, H. and Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs, *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, ACM, pp.533–542 (2006).
10) 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).
11) 戸田智子，横山昌平，福田直樹，　石川博：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，第 1 回データ工学と情報マネジメントに関するフォーラム DEIM2009 (2009).
12) Cheng, H., Hua, K.A. and Vu, K.: Constrained locally weighted clustering, *Proc. VLDB Endow.*, Vol.1, No.1, pp.90–101 (2008).
13) 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).
14) : 形態素解析システム sen，http://ultimania.org/sen/.
15) Turnbull, D. and Elkan, C.: Fast Recognition of Musical Genres Using RBF Networks, *IEEE Trans. on Knowl. and Data Eng.*, Vol.17, No.4, pp.580–584 (2005).
16) Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S.: Constrained k-means clustering with background knowledge, *In ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, USA, pp.577–584 (2001).
17) Allan, J., Carbonell, J., Doddington, G., Yamron, J. and Yang, Y.: Topic detection and tracking pilot study: Final report, *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp.194–218 (1998).
18) Shriberg, E., Stolcke, A., Hakkani-Tür, D. and Tür, G.: Prosody-based automatic segmentation of speech into sentences and topics, *Speech Commun.*, Vol.32, No.1-2, pp.127–154 (2000).
19) Castellanos, M.: *Survey of Text Mining*, chapter 6. HotMiner: Discovering Hot Topics from Dirty Text, Springer (2003).
20) Kleinberg, J.: Bursty and hierarchical structure in streams, *Proc. the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, ACM, pp.91–101 (2002).
21) 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).