F-051

クラスタリング手法を用いたソーシャルブックマークからの時間的変化の抽出
Extraction of Temporal Cluster Changes from Social Bookmark Data

野口 裕貴† 山口 崇志† マッキン ケネス ジェームス† 永井 保夫†
Yuki Noguchi Takashi Yamaguchi Kenneth James Mackin Yasuo Nagai

1．はじめに

近年ニュースやブログ等を能動的な検索をせずに取得する情報推薦が注目されており、推薦手法としてクラスタリングを用いた方法がある[1]。これは個別のアイテムやユーザに対して推薦処理の計算量が増加する問題を解決する為や、対象の細かい差異を吸収する為に行われている。

その手法の一つとしてソーシャルブックマーク(SBM)に対してクラスタリングを用いるものがある。これは SBM を膨大な Web ページから有用なページを入手するフィルタリングとしてみなすものであり、SBM を用いることで無駄な対象を減らす事ができる。しかし、対象となる SBM は随時様々なページが追加されており、時間的に変化する特徴を持つのに対し、従来のクラスタリング手法では時間的変化を考慮していない。また、その変化を調べるには 全体を調べる必要があり、計算コストが高いという問題がある。

本稿では時間的変化を考慮したクラスタリング手法の提案と、そのクラスタリング結果からの時間的変化の抽出を行う。具体的には一定時間毎に設定した時間区間のデータを用いてクラスタリングを行い、時点毎のクラスタの変化を調べる事でデータ集合全体の時間的変化の抽出を行う。これにより各カテゴリの注目度の高いトピック等抽出し、推薦等への応用が期待できる。

2.提案手法

2.1 時間的変化の抽出手法

SBM のデータを特定の時間区間で区切り、区間毎にクラスタリングを行う。これによって作成される区間毎のクラスタの変化を調べる事で時間的変化の抽出を行う。本稿で提案する手法は 1～5 の 5 つのステップで構成されており、次の章で各ステップの詳細を説明する。

Step 1 データ収集
Step 2 ページとタグの親和度の計算
Step 3 ページのクラスタリング
Step 4 分類器の作成
Step 5 クラスタの変化の抽出

2.2 各ステップの詳細

2.2.1 データ収集

SBM データは SBM を利用している全てのユーザが登録した Web ページ、タグ、時間からなるデータである。本論文では図1の様にある時間 t を起点として、前後 s 時間分を1つのデータセットとして考える。これらデータセットを時間 Δt 毎に準備し、以下のように表す。

$$P = \{P_i\}_{i=1,2,\ldots,i\max}$$
$$T = \{T_j\}_{j=1,2,\ldots,j\max}$$
$$ST_t = \{T_j \mid T_j \in T\}$$
$$SP_t = \{P_i \mid P_i \in P\}, t < pt_u < t+s$$

†東京情報大学 総合情報学部 情報システム学科
Department of Information Systems,
Tokyo University of Information Sciences

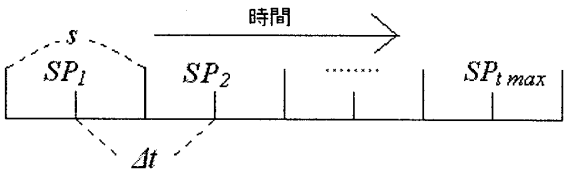

図 1 SBM データを時間で区切ったモデル図

P は全てのブックマークデータの集合であり、時間区間 s で区切った範囲を SP_t と呼ぶ。それぞれの SP_t にはその時間区間で登録された Web ページ P_i が登録されている。そして P_i には複数のタグ T が登録されている。例として表 1 に区間 SP_t における SBM データを示す。なお、Δt の値を s 以下にすることによってデータの重複をさせる事も可能である。

表 1 時間区間 SP_t で登録された SBM データの例

	タグ1	タグ2	タグ3	タグ4		タグ$\|ST_t\|$
url	java	oop	分散処理	並列処理		nio
page 1	9	0	6	7		0
page 2	11	14	0	0		0
page 3	7	0	11	0		0
page 4	4	0	0	10		8
page 5	0	0	5	7		0
⋮						
page$\|SP_t\|$	0	0	5	7		0

2.2.2 ページとタグの親和度の計算

ページとタグの親和度を計算する。親和度はページとタグの関連度を相対的に表す値である。ページ P とタグ T の親和度 $rel(P, T)$は次式により求められる[3]。

$$rel(P,T) = TF(P,T) \times IDF(T)$$

この場合、$TF(P,T)$はページ P に関連付けられている全てのタグに対してタグ T が占める割合を示し、$IDF(T)$ は全体のページにおけるタグ T の希少性を表す値である。

これにより以下の様なベクトル X_i^t が得られる。なお、スペルミス等で付けられたノイズ的なタグを除去する為、ベクトルのサイズは ST_t から IDF の値が 0.1 以上となる T の要素からなる集合を RT_t にした。

$$X_i^t = \{rel(P_i^t, T_y)|_{y=1,2,\ldots,|RT_t|}\}$$
$$RT_t = \{T_j \mid T_j \in T, \mathrm{IDF(Tj)} > 0.1\}$$

表 2 TF・IDF による親和度ベクトル

	タグ1	タグ2	タグ3	タグ4		タグ$\|RT_t\|$
url	java	oop	分散処理	並列処理		nio
page 1	0.2	0	0.3	0.4		0
⋮						
page$\|SP_t\|$	0	0	0	0		0

2.2.3 ページのクラスタリング

Step2 で求めたページとタグ間の親和度ベクトルをもとに，話題やテーマとして類似するページ同士をクラスタ化する。ここではクラスタリング手法として x-means[2]を用いる。x-means 法は、クラスタリングの代表的手法である k-means 法を、クラスタ数が未知である場合に拡張した手法である。

上記の x-means 用いて P に対してクラスタリングを行い、得られた m 個の SBM データクラスタ c_m^t クラスタ集合 C^t を各クラスタの重心や所属する要素数、分散と共に保存する。

$$C^t = \{c_m^t\}_{m=1,2,\ldots,m\max} \quad c_m^t = \{P_n\}_{n=1,2,\ldots,n\max}$$

2.2.4 分類器の作成

Step3 で作成したクラスタリング結果を基に入力された SBM データが結果と同様の分類を行える分類器を作成、保存する。作成した分類器では重心からの最近傍識別法を用いる。

2.2.5 クラスタの変化の抽出

クラスタが時間区画 *t-1* から *t* に変化した際のクラスタの変化を抽出する。例として図 2 を用いて説明をすると、「*t=1* から *t=2* に変化した際に、a が a1 と a2 へ分裂し、b が拡大する」といった変化を抽出する。

ステップは大きくわけて二つあり、関連するクラスタの特定と、変化の種類の特定である。関連するクラスタの特定には図 3 の様に Step3 で得られた時間区画 *t* のクラスタと、Step4 で作成した時間区間 *t-1* の分類器を用いる。結果、得られたデータから *t* のクラスタが *t-1* からどのような変化をしてきたかを抽出する。

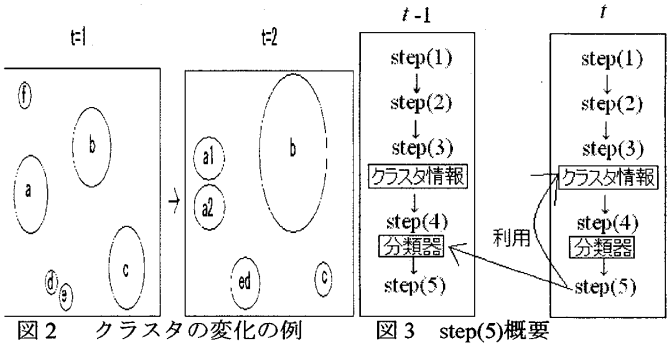

図 2　クラスタの変化の例

図 3　step(5)概要

変化のクラスは以下の 7 種類を定義した。

分裂、合併、発生、拡大、縮小、増加、減少

図4にフローチャートを示す。なお、以下に c_m^t と最も良く似ている *t-1* 時点でのクラスタとの尤度を求める計算式を次式に示す。

$$d = \frac{\max(A^{t-1})}{|c^t m|} \qquad \cdots \text{式(1)}$$

A^{t-1} は *t-1* での分類器に、c_m^t を入力して得られた分類結果のベクトルである。ベクトルとクラスタのインデックスが対応しており、ベクトルの要素は各クラスタに分類された要素数を表す。なお、ベクトルの長さは現在の *t* におけるクラスタ数 *m* ではなく、*t-1* 時点でのクラスタ数となる。

$$例: A^{t-1} = \{0,16,3,1,0\}$$

表 3　c_1^t の SBM データの各クラスタ分類例

	c_1^{t-1}	c_2^{t-1}	c_3^{t-1}	c_4^{t-1}		$c_{\|RT-1\|}^{t-1}$
c_1^t	0	16	3	1		0

図 4 によって以下のような *t* 時点でのクラスタと関連する *t-1* 時点でのクラスタ及び、その変化の種類が得られる。

$$\{c_m, 変化の種類, 変化の種類, c_1, c_2, \ldots, c_s\}$$

続いて得られたデータに対し以下の式を適応し、変化の種類を追加する。

$$V(c_m^t) > V(\Sigma c_s^{t-1})$$

これが真であれば、変化のクラスに拡大を追加し、偽であれば縮小を追加する。なお、*V* は分散を表す。

$$|c_m^t| > |c_s^{t-1}|$$

これが真であれば、変化のクラスに増加を追加し、偽であれば減少を追加する。

3 .実験

本手法の有効性を評価するために実際のブログデータを用

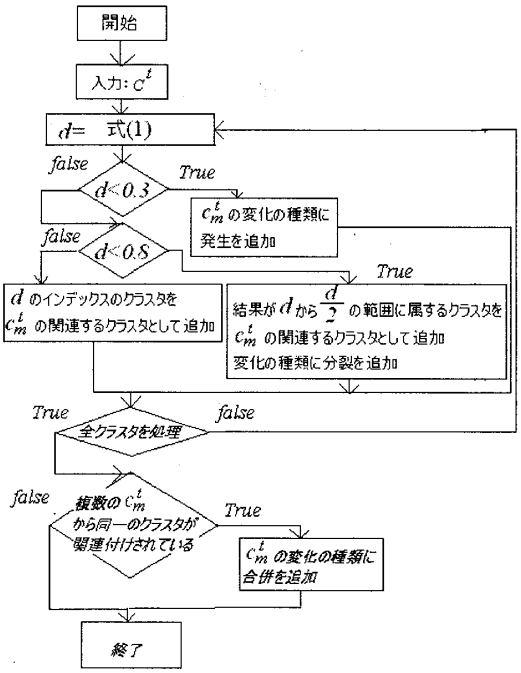

図 4　変化のクラスを特定するフローチャート

用いた実験を行った。実験データとして株式会社 liveDoor が提供する SBM サービスである liveDoor クリップ[4]の登録データを用いた。用いるデータは 2008 年 1 月～2008 年 12 月の期間のものである。これを 15 日毎に、+－15 日分のデータで分け、各区間の変化を抽出する

実験結果として変化の検出した際の精度を求めた。精度とは抽出された変化を判定したもののうち、実際にそのような変化が起ったものであった割合である。今回の実験では抽出された発生、拡大、縮小、増大、縮小の各変化のクラスが正しく抽出されたかどうかを人間によって判断した。この結果，精度が 73% であった。

4.おわりに

本研究の目的は、推薦や注目のトピックの発見に役立てる為、SBM 等の動的に変化する対象に対する時間的変化を抽出することである。本稿ではクラスタリングによってデータの時間的変化を抽出する手法を提案し、実際に SBM データから各種変化を抽出できる事を確認した。しかし、変化の種類がこれで十分か等、引き続き検討していく必要がある。また、今後はより長期の視点にたった変化の抽出や、各種応用について検討していく。

参考文献

[1] 丹羽 智史，土肥 拓生，本位田 真一, Folksonomy マイニングに基づく Web ページ推薦システム, Transactions of Information Processing Society of Japan 47(5) pp.1382-1392 (2006).

[2] Dan Pelleg, and Andrew W. Moore, X-means: Extending K-means with Efficient Estimation of the Number of Clusters, Proceedings of the Seventeenth International Conference on Machine Learning, (2000).

[3] Thorsten Joachims, A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization, Proceedings of the Seventeenth International Conference on Machine Learning, (2000)

[4] http://labs.edge.jp/datasets/