Web タグの階層的クラスタリング手法の提案

竹内尚†* 鈴木育男† 山本雅人† 古川正志†
(北海道大学情報科学研究科)†

1 はじめに

Web 上のテキストや画像などのコンテンツに対して, エンドユーザがタグと呼ばれる分類情報を付加可能なフォークソノミーと呼ばれるサービスが増加している. サービスの例として動画共有サイトのニコニコ動画や, イラスト共有サイトの pixiv , はてなブックマークや del.icio.us などのソーシャルブックマークが挙げられる.

フォークソノミーは従来のディレクトリ検索やロボット検索とは異なり, サービスを利用するユーザーごとにアクセスしやすいと思えるタグを自由につけることが可能なため, さまざまな観点から情報を分類することが可能である. しかし, 問題点としてユーザーはディレクトリ検索のような複雑な階層構造を意識してタグを付けることがなく, また, メタノイズが存在することからタグの階層構造を容易に構築することは難しい.

タグの階層構造とは, 本論文ではタグに用いられている単語の抽象の度合いによる階層構造で抽象性が高い単語は幅広いタグに用いられていると考えられる. 逆に, 抽象性が低く具体性の高い単語はごく一部のタグに用いられると考えられる. そこで本稿では, タグの階層関係を抽出する手法を提案することを目的とする.

現在の google などに代表されるパターンマッチングによる検索では, ユーザーにボキャブラリーなどの知識が要求されるが, 階層構造が構築されれば 1 つのキーワードをきっかけとして次々とユーザの要求する情報を得ることができる.

本稿では, タグに基づく階層的クラスタリング手法を新たに提案し, 数値実験と実際にフォークソノミーとしてソーシャルブックマークを用いて, この手法の適用実験を行い, 検証を行う.

2 提案手法

2.1 概要

この節では,Self Organization Map(SOM) を用いた階層的手法の概要を簡潔に説明する. この手法のプロセスは 3 段階に分かれている.

1. 類似度の高いベクトルデータのマージを行う.
2. SOM による学習を行う.
3. 1,2 の手順を繰り返す.

この手法は, まずベクトル表現されたタグデータの入力をランダムに n 次元ユークリッド空間に与えて, 近傍の領域に重ならないように最大で半径 R の領域 D を作成する. 入力が D に入った場合, 領域を作成を行わない. 入力数があらかじめ指定された回数に達すると, ある入力の領域 D に含まれている入力のマージを行う. 次にマージされた入力を用いて SOM を学習し, クラスタリングを行う.

以上のプロセスの繰り返しによって階層的なクラスタリングを行う.

2.2 マージ方法

まず, 半径の最大値 R_0 の設定する. 次に入力をランダムに 1 回づつ選出し n 次元ユークリッド空間に与え, 他の入力の領域を重複させないように半径 r の領域 D を作成する. 領域の半径は階層ごとに指数的に増加させるために以下を定義した.

$$R_i = R_0 \times a^i \tag{1}$$

i 階層の最大半径 R_i, 底を $a = 1.25$ とした.

全入力が終わると, 入力 x_i のマージを行う. マージは領域 D に含まれる X_d に適用し, 重心を計算した. x_{i+1} を次の階層の入力として SOM の学習に用いた. 式を以下のように示す. ここで,X_d は領域 D に含まれる入力 x_i の集合を表す.

$$x_{i+1} = \frac{1}{|X_d|} \sum_{x_i \in X_d} x_i \tag{2}$$

2.3 Self-Organization Map(SOM)

SOM は,T.Kohonen[2] が提案した教師無しニューラルネットワークである. SOM の学習は, 参照ベクトルが初期化された状態から始まる. まず,n 次元の入力データからランダムに選出したデータを SOM に与える. ユニット m_i のうち, 入力データの距離が最も近いものを勝者 (BestMachingUnit,BMU) と呼び c で表す.n 次元の入力を $x \in \Re$ とすれば, 勝者の決定条件は以下で表される.

$$c = arg \lim_i \{\| x - m_i \|\} \tag{3}$$

$d(x, m_i)$ には, ユークリッド距離やマンハッタン距離などが目的によって使い分けられる. 勝者の参照ベクトルと, その近傍に位置するユニットは, 入力されたデータに近づくように更新される. 更新則は以下で表される.

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \tag{4}$$

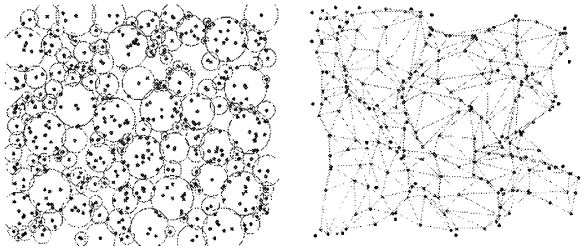
$$h_{c,i}(t) = \alpha(t) \cdot exp\left(-\frac{d_{c,i}^2}{2\sigma(t)^2}\right) \tag{5}$$

ここで t は学習時回数,$h_{ci}(t)$ は学習の条件付重み,$\alpha(t)$ は学習係数,σ_i は学習開始時の近傍半径であり, どちらも t の単調減少関数である.d_{ci} は勝者 c とユニット i 間の SOM のトポロジで定義されるユークリッド距離である. 近傍が大きくなるか,t が大きくなるほど近傍関数 $h_{ci}(t)$ は緩やかに減少する.

Propose the hierarchical clustering method use of Web tags
†Graduate school of information science and technology
Hokkaido University
*takeuchi@complex.eng.hokudai.ac.jp

	数値計算実験	実データ適用
入力信号の総数	500	7482
ユニット数	15 × 15	10 × 10
学習回数 T	5000	15000
初期学習係数 α_0	0.9	0.9
近傍半径 σ_0	7	5
初期最大半径 R_0	30	5

Table 1 実験設定条件



(a) 領域分布　(b) 位相マップ

Fig. 1 マージ領域の分布とマージ後の位相マップ

3 数値計算実験

入力を 2 次元の x,y 座標として $x = 0$〜400, $y = 0$〜300 のランダム値を用いて実験を行った. 実験では SOM の競合層に六角形のトポロジを採用し, 表 1 に示す実験条件を設定した.

入力のマージ領域とマージを行ったデータに対して SOM の学習を行ったときの位相マップを図 1 に示す. マージの領域は他の領域に重ならないように半径を広げており, マージした入力が参照ベクトルによって近似されていることがわかった.

4 実データへの適用

ソーシャルブックマークサイトによるブックマーク情報を用いて提案手法の適用を行った. livedoor クリップの提供している 2008 年 12 月にクローリングによって得られたデータセットを用いる. ブックマーク情報には, サービスの登録されたエントリ (URL) と, そのエントリに付加されたタグ情報を使用する. ブックマーク情報から, タグの付けられた回数が上位 100 個の単語と, タグの中で最大で 5 以上付加されているエントリを実験に用いた. エントリ数は 7482 である.

エントリとタグの関係を表現したベクトル行列を定義した.

$$entry = [tag1\ tag2\ tag3 \ldots tag100] \tag{6}$$

ベクトルの中には, そのエントリに付加されたタグの総数がタグ別に格納されている. これらを入力ベクトル集合 X として提案手法を適用し, マップを生成した. 実験設定は, 表 1 に示す条件を適用した.

5 結果

結果の考察のため, 各ユニットのエントリの数と各ユニットにおけるベクトルの値が高い上位 2 つのタグの表示を行った. 図 2 と 3 に結果を示す. 最下層のマップに見られる「windows」,「通販」タグなどの傾向が 6 層目

Fig. 2 最下層の SOM

Fig. 3 6 層目の SOM

のマップには見られなくなっており, 逆に 6 層目に見られる「社会」,「動画」などの傾向が最下層には見られなかった. また,6 層のユニットにある「まとめ」タグには最下層のユニットには「2ch」や「windows」などの具体性のあるタグと上位に存在している事から, 本手法によってタグの階層関係を抽出可能であることが確認された.

6 おわりに

本研究では, フォークソノミーで得られた情報の併合を繰り返し, その結果を 2 次元のマップ上に配置することで, 階層構造を抽出する手法を提案した. このことによって, 階層ごとにユニットの属性の変化が見られ, 階層関係を抽出することができた.

今後の研究として, フォークソノミーの性質による特徴量を明らかにすることやタグの階層化構造の可視化などが挙げられる.

参考文献

[1] PHILIP D.WASSERMAN: ニューラル・コンピューティング　上級編, 嘉数侑昇, 古川正志, 森川一共訳, 森北出版 (1998)

[2] T. コホネン: 自己組織化マップ 改訂版, シュプリンガージャパン (2005)