

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、ユーザーにポ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) を用いた階層的な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。

$$R_i = R_0 \times a^i \quad (1)$$

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。

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

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 (Best Matching Unit, BMU) と呼び c で表す。 n 次元の入力を $x \in \mathcal{R}$ とすれば、勝者の決定条件は以下で表される。

$$c = \arg \lim_i \{ \|x - m_i\| \} \quad (3)$$

$d(x, m_i)$ 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。

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

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

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

