4P-1

ソーシャルブックマークと Folksonomy を利用した Web ページ推薦システム

Web page recommendation system using by Socialbookmark and Folksonomy

菅野 直之‡ 大矢 健太‡ 小瀬木 浩昭‡ 武田 正之†

Naoyuki Kanno Kenta Ohya Hiroaki Ozeki Masayuki Takeda

1. はじめに

近年, Blog や Wiki の登場により個人が情報を発信する機会が増え, インターネット上の Web ページは急激に増加している. 2006 年 1 月の時点で Google に登録されている Web ページ数は 80 億を超える. このような状況の中で, ユーザにとって興味や価値のある情報を追い続けていくことは困難である. そのため, これらの情報を推薦してくれるシステムが実現すれば, その利用価値は高い. しかし, 従来の Web ページ推薦システムの多くは, 膨大な Web ページの中から協調フィルタリングを用いて推薦するために, 推薦精度が落ちるなどの問題があった.

本稿では, ソーシャルブックマーク(以下, SBM)上の Web ページを利用した Web ページの推薦システムを提案する. SBM では, 既にユーザによって読まれ価値があると思われるページがブックマークされている. そのため, あるページに対するブックマーク数はそのページの価値を示していると考えられる. また, SBM ではブックマークするページに対してユーザが「タグ」というメタデータを付与することができるが, そこではタグ表記の揺れという問題が発生する. 本稿では, この問題をタグ間の依存度を計算しクラスタ化することにより解決を試みる.

2. ソーシャルブックマークと Folksonomy

本提案システムでは, 推薦するページとして SBM 上のページデータを使用している. また, ページの分類に Folksonomy を利用している.

2.1. ソーシャルブックマーク

SBM とは, ブラウザで行うブックマークをインターネット上で行ったもので, 登録したページを公開・共有することができるサービスである. 単純なオンラインブックマークとは異なり, 多くのユーザでブックアークを共有し, 分類や人気度, コメント, 解説などの情報を付加していくことに主眼が置かれている.

2.2. Folksonomy

Folksonomy とは, folks(人々)と taxonomy(分類学)を組み合わせてできた造語である. 従来の Taxonomy のような階層構造を持たず, 完全にフラットなネームスペースを持つ. 基本的に無秩序で有機的な構造だが, 逆にユーザは自分ならではの方法を発見しながら情報を整理することができる.

2.3. Folksonomy の問題点

Folksonomy では, ユーザが個別にタグ付けを行うため, タグの綴り間違えや表記の揺れ(book, books, 本など), タグの正当性をどのように評価するのかという問題がある.

†東京理科大学 理工学部 情報科学科,
Dept. of Information Science, Tokyo University of Science
‡東京理科大学大学院 理工学研究科 情報科学専攻,
Graduate School of Science and Technology,
Tokyo University of Science

3. 提案システム

本提案システムは, タグをクラスタ化するフェーズと, ページを推薦するフェーズの 2 つのフェーズによって構成される. **図 1** に提案システムの概要を示す.

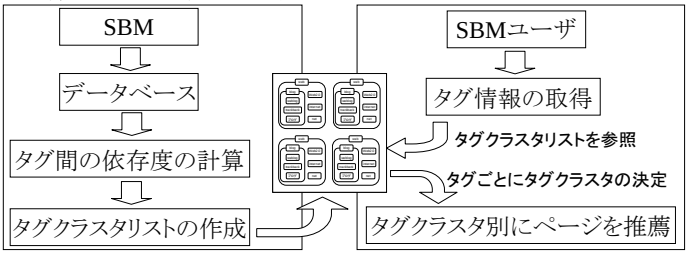


図 1・提案システムの概要

3.1. タグのクラスタ化フェーズ

3.1.1. SBM よりデータを収集

本提案システムは, 主に次の 3 つのデータを SBM より収集し利用している. 1. ブックマークされているページ群, 2. ページに付けられたタグ群, 3. ページに同じタグを付けた人数. 3. のデータはタグによるページの人気度を示している. 例えば, ページ P にタグ T を付けたユーザ数が 10 の場合, タグ T によるページ P の人気度は 10 となる.

3.1.2. タグ間の依存度の計算

タグ間の依存度は, 以下の式で定義される.

$$dep(T_1,T_2)=\sum_{P_i\in Pages} w(P_i,T_1)\times tfidf(P_i,T_2)$$

$w(P,T_1)$ は, ページ P にタグ T_1 を付けた人数を示している. $tfidf(P,T_2)$ は, $tfidf$ 法[1]を用いておりページ P に付けられたタグ群の中でのタグ T_2 の特徴度を示している.

$$tfidf(P,T)=tf(P,T)\times idf(T)$$

$tf(P,T)$ は, タグ T に関連のあるページに付けられているタグ群の中での, タグ T_2 の割合を示している.

$$tf(P,T)=\frac{w(P,T)}{\sum_{T_i\in Tags} w(P,T_i)}$$

$idf(T)$ は, 全タグの中でのタグ T の希少性を示している.

$$idf(T)=\log\frac{\sum_{P_j\in Pages}\sum_{T_i\in Tags} w(P_j,T_i)}{\sum_{P_j\in Pages} w(P_j,T)}$$

3.1.3. タグのクラスタ化

求めたタグ間の依存度を用いて, タグをクラスタ化する. クラスタ化する際, クラスタの中で最も dep が大きいタグを親タグと呼ぶことにする. つまり, 親タグはクラスタの中で最も重要度の

高いタグであるので，親タグを決定することによりタグをクラスタ化する．以下に，親タグの決定方法を示す．

タグ T_1 に対して最も dep の大きいタグ(T_1 自身は除く)を T_2 とし，$dep(T_1,T_2)$ の値が予め決めた閾値を超えた場合は，T_1 の親タグを T_2 とし，超えない場合は T_1 自身を T_1 の親タグとする．この場合，タグ T_1 はサイズ 1 のタグクラスタを作ったことになる．

T_1 と T_2 がお互いにとって最も依存度の高いタグである場合，$dep(T_1,T_2)$ と $dep(T_2,T_1)$ を比較して，前者が大きければ T_2 を，後者が大きければ T_1 を双方の親タグとする．以上の操作を繰り返し，タグクラスタを作成する．

3.2. ページ推薦フェーズ

ユーザが SBM で付けたタグ情報を取得し，それぞれのタグが所属するタグクラスタを決定する．推薦するページは，タグクラスタ内で，$w(P,T)$ の大きい順に選ばれる．ユーザはブラウザを通してクラスタ内でのタグの選択や閾値の選択ができる．

4. 実験結果と評価

4.1. 収集したデータについて

今回の実験は，はてなブックマーク[2]で 10 人以上がブックマークした34478ページに関するデータを収集して行った．そこに付けられていたタグの種類は全部で 24880 個であった．最も多くのページに付けられていたタグは「web」で 8741 ページ，二番目に多かったのは「ネタ」で 5946 ページであった．

図 2 は，ページ数とタグ数の関係を示している．全体の約84%のタグは，10 ページ未満にしか付けられていないことが分かる．これは，ユーザのタグの付け方にばらつきがあることを示している．よって，類似するタグをクラスタ化することは，Folksonomy 一般にとって有益なことと考えられる．

4.2. タグのクラスタ化について

図 3 に，タグをクラスタ化する際に使用した閾値と，各タグのクラスタサイズの関係を示す．また，クラスタサイズが 10 以上であった 26 個のクラスタの平均についても示す．平均の値を見ると閾値 50 あたりから，クラスタサイズの減少が緩やかになっていることが分かる．

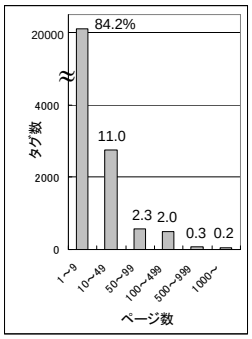


図 2・ページ数とタグ数の関係

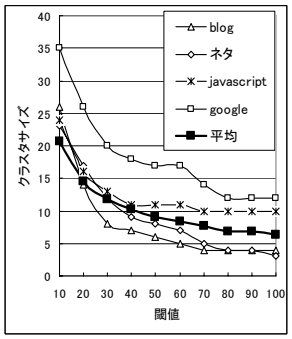


図 3・閾値とクラスタサイズの関係

図 4 に，「book」を親タグとするクラスタの内容を示す．この結果より，タグのクラスタ化には 2 つの意味があることが分かる．1つは，等価関係であり，もう1つは包含関係である．等価的内容のタグをクラスタ化することにより，「blog」や「ブログ」など同じ意味で表記の違うタグを 1 つとみなすことができる．また，包含的内容のタグをクラスタ化することにより，ユーザの潜在的ニーズを満たすページを推薦することができると考えられる．

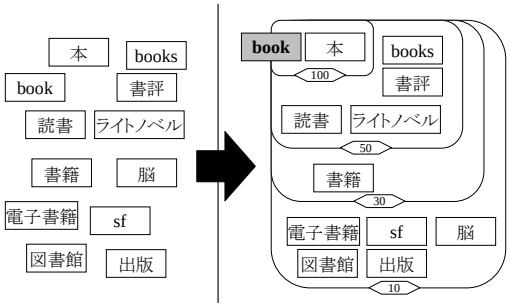


図 4・閾値とクラスタ化の様子

4.3. 推薦ページの人気度について

図 5 は，タグをクラスタ化して推薦した場合と，クラスタ化しないで推薦した場合とで，ページ人気度の比較を行っている．クラスタ S とは，クラスタサイズが 4 以上で閾値を 50 と設定した場合のクラスタの集合である．タグ T_1 はクラスタ S の中の親タグの集合，タグ T_2 は T_1 の各親タグに対して最も dep の大きいタグの集合である．グラフの上の数値は，クラスタ S の T_1，T_2 に対しての上昇率を示している．**図 5** の結果より，クラスタ S はクラスタ化前のタグ単体での推薦方式と比較して，人気度の高いページを推薦できていることが分かる．親タグはクラスタ内で最も重要なタグであるため，親タグで推薦されるページと，クラスタで推薦されるページとではそれほど差異はない．しかし，T_2 のように少しでも親タグと表記が違うと大幅に人気度が落ちる．

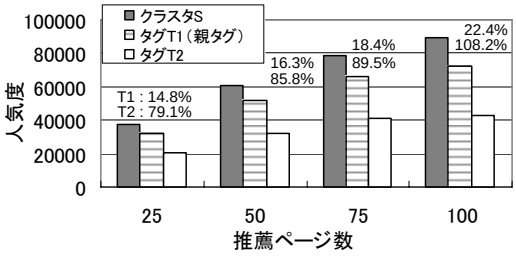


図 5・クラスタ化によるページ人気度と上昇率

5. 関連研究

[3]では，ユーザの興味分野を共有する他のユーザがいなければ，そのユーザには有益な情報を提供できないという問題点があるが，本提案方式では既に構築された SBM を利用することで，この問題を解決している．

6. まとめと今後の課題

本稿では，SBM 上のデータを利用した，Web ページ推薦システムについて述べた．Folksonomy ではタグの表記の揺れが問題になるが，タグ間の依存度を求めることによりこの問題の解決を試みた．今後は，推薦対象を SBM ユーザに限定せず，登録 RSS や現在見ているページなど，様々な嗜好データを利用してページ推薦が行えるよう検討していきたい．

参考文献

[1] Salton, G.and McGill, M.J.:*Introduction to Modern Information Retrieval,* McGraw-Hill Book Company, 1983

[2] はてなブックマーク:http://b.hatena.ne.jp/

[3] 佐野浩司，佐山弘樹:Web ブラウザのブックマーク機能を利用した情報共有システム，人工知能学会論文誌 20 巻 4 号 B 281-288 (2005)