Web の閲覧履歴を情報源としたソーシャルブックマークにおけるタグ推薦の提案

阿部　佑樹† 糸川　剛† 北須賀輝明† 有次　正義†

† 熊本大学大学院自然科学研究科 〒 860-8555 熊本県熊本市黒髪 2-39-1
E-mail: †abec@dbms.cs.kumamoto-u.ac.jp, ††{itokawa,kitasuka,aritsugi}@cs.kumamoto-u.ac.jp

あらまし　ブラウザで管理していたブックマークをインターネット上で管理するソーシャルブックマークが注目されている．ソーシャルブックマークの特徴の一つにタグがあり，タグを付けることでユーザ自身が再度ブックマークを閲覧するときや, 他人のブックマークを閲覧するときに参考になる．タグ付けをするには, ユーザ自身が対象となるページの内容を表している単語を考える必要があり， 現状ではその補助として，タグ付け履歴を用いたタグ推薦が主に利用されている． 本研究では従来手法を補完するものとして対象ページに付与するタグや コンテンツ内の単語と Web の閲覧履歴との関連度を計算することで, 閲覧履歴からユーザの興味を反映したタグ推薦を行う手法を提案し, 有用性を確認する．

キーワード　ソーシャルブックマーク, タグ推薦

A Proposal of Tag Recommendation Based on User's Web Browsing History

Yuki ABE†, Tsuyoshi ITOKAWA†, Teruaki KITASUKA†, and Masayoshi ARITSUGI†

† Graduate School of Science and Technology, Kumamoto University
2-39-1 Kurokami, Kumamoto, Kumamoto 860-8555, Japan
E-mail: †abec@dbms.cs.kumamoto-u.ac.jp, ††{itokawa,kitasuka,aritsugi}@cs.kumamoto-u.ac.jp

1 はじめに

近年，ブログやミニブログと呼ばれる twitter のようなユーザが容易にコンテンツを生成できるサービスが普及し，Web 上に大量に情報が発信されるようになった．膨大なコンテンツの中で偶然に興味のあるコンテンツにアクセスする事は難しく，ユーザは一般的に Yahoo! や Google などの検索サイトなどで興味のあるキーワードを入力し，検索結果を元にコンテンツへアクセスする．その閲覧履歴を収集，分析し，Amazon や Google AdSense などは，ユーザ毎に異なる商品の推薦やユーザが日々閲覧しているサイトのカテゴリを考慮した広告の配信を行っている．ユーザの興味を反映した閲覧履歴は，インターネット上でサイトを運営し，オンラインショップや検索エンジンに対して広告を出稿し宣伝を行っている企業だけでなく，ユーザにとっても有益な情報源となっている．

ユーザの閲覧履歴を用いて，解析を行うことでユーザに対して有益な情報を提示する研究も行われており，閲覧履歴を元にブラウザの閲覧支援を行う研究 [1] やユーザへの情報推薦や情報フィルタリングを行う研究も行われている [2,3].

ユーザはサイトを閲覧し，再閲覧する可能性があれば，ブラウザの機能であるブックマークを用いることで，検索の手間を省くことができる．ブックマークを他人と共有可能にしたものがソーシャルブックマークである．ソーシャルブックマークはサービスを提供しているサーバで管理でき，他人のブックマークを閲覧したり，自分のブックマークを公開する機能を持つ．また，タグと呼ばれる登録時にユーザが付与する語句により分類でき， 同一のタグを付けているユーザやウェブページをたどることで，興味の似通ったコンテンツを見つけやすくなる．ソーシャルブックマークを情報源として用いた情報検索や情報推薦の研究 [4–6] が盛んに行われている．

現在，はてなブックマーク [7] などのソーシャルブックマークサービスでは，既に他のコンテンツに付けていたタグを提示することでタグ付けを行う手法や，同一のページをブックマークした他のユーザが付与したタグを参考にし，タグ推薦を行う手法を用いている．しかし，既に他のコンテンツに付けていたタグを提示する手法はソーシャルブックマークを使って間もな

いユーザにとっては，対象となる興味分野に対する単語の語彙数が少ない．また，同一のページをブックマークした他のユーザが付与したタグを提示する手法では，誰かが先にブックマークしていないとタグを推薦しない．また，コンテンツと関係ないタグ付けを行うスパム行為があるため，提示されるタグにノイズが混ざることもある．

[8] で議論されているように，ユーザがタグをつける動機はいくつか考えられる．本研究では，他人と情報を共有することを動機としたタグ付けを対象とする．つまり，個人的に再利用することを目的にしたタグにみられるような，一般的すぎるタグや，共有するためには特殊すぎるタグは対象とせず，適度に情報を表現する適度に一般的と考えられるタグを対象とする．

本研究では，ユーザが閲覧した Web コンテンツの履歴を情報源としたソーシャルブックマークのタグ推薦システムを提案する．提案手法の目的は，はてなブックマークなどで用いられている既に他のコンテンツに付けていたタグを元にタグ推薦を行う従来手法の補助を行うことである．ここで補助とは，従来手法で推薦されなかった Web ページに適したタグが提案手法で推薦されることを示す．

提案手法で閲覧履歴を用いた理由を述べる．ユーザは興味を持ったコンテンツを再閲覧すると考え，ブックマークを行う．ブックマークを行わないコンテンツでも興味を持ったコンテンツに頻繁にアクセスしていると仮定し，ブラウザの閲覧履歴を情報源とすることで，興味を持ったコンテンツをもとにタグ推薦が行えるのではないかと考えた．閲覧履歴には，ユーザが興味を持っている分野の単語が出現していると仮定し，閲覧履歴を解析することでユーザへ推薦を行うための単語を抽出できるのではないかと考えた．

本論文の以降の章は以下の通りである． 2 章では，本研究で扱うソーシャルブックマークの概要について述べる． 3 章では，本研究と関連する既存の研究に関して述べる． 4 章では，本研究で提案する手法に関して述べる． 5 章では，今回行った実験に関して述べる．最後に，6 章で本論文のまとめと今後の課題に関して述べる．

2 ソーシャルブックマーク

ソーシャルブックマークは，ブログや SNS (Social Networking Service) など今まで情報の受け手であったユーザが情報を発信できるようになったメディアの形態であるソーシャルメディアの一つである．国外の del.icio.us [9] から人気が出始め，国内でははてなブックマークや livedoor クリップ [10] などがある．これらのサービスは，ユーザにブックマーク機能を提供し，サービスを提供している企業のサーバにブックマーク情報を格納する．ユーザ自身の端末にブックマークするのではなくサーバにブックマークするため，インターネットに接続されてさえいれば，どこからでもブックマークにアクセス可能である．また公開されている他人のブックマークを閲覧でき，興味の分野が似ている他人とブックマークを共有することやコメントを残すことが可能で，同じコンテンツを見たユーザの感想を確認することができる．コンテンツごとにブックマークされた数も確認できるため，閲覧時の参考にすることもできる．また，ソーシャルブックマークサービスを提供している企業は，ユーザのブックマークを解析し，話題になっているコンテンツの情報を提供することができる．

ソーシャルブックマークは， 民衆を意味する folk と分類法を意味する taxonomy からなる造語であるフォークソノミー [11] と呼ばれるデータ構造となっている．フォークソノミーはユーザ，タグ，リソースとなるコンテンツ，そして，ユーザがあるリソースにつけたタグで構成される TAS と呼ばれるタグの割り当ての 四つの要素からなる．フォークソノミーの特徴であるタグは，画像や動画，音楽に対しても付けることができる文字情報で，その文字情報を元に 検索を行い，利用者が目標としている情報へとアクセスすることができるメタデータである．また，一つのコンテンツに対して複数のタグを付けることにより柔軟な分類ができるといった面もある．

3 関 連 研 究

ソーシャルブックマークは，ユーザの嗜好を反映しているため，有益な情報源として，研究が盛んになっている．

多くの Web ページ推薦システムで用いられている協調フィルタリングでは，ユーザ数に比べ Web ページ数が圧倒的に多い．そのため， ソーシャルブックマークをコンテンツ推薦の情報源とする研究として，丹羽ら [12] はソーシャルブックマークのデータをユーザの Web ページ嗜好データとして利用することを提案した．ソーシャルブックマークのユーザ数に比べ，Web ページの数が多いため，ユーザ同士の嗜好の類似度を比較する際に，ブックマークページの共有数を参考にするのではなく，ユーザの嗜好をユーザと各タグとの親和度とすることで，ユーザの嗜好表現が抽象化され，直接比較するよりもユーザ間の嗜好類似度が計算しやすくなると述べた．また，佐々木ら [13] はタグに着目するのではなく，タグが付与された Web ページのクラスタに注目し，クラスタ間の類似度を仮説検定問題として求め，得られた類似度に基づき，コンテンツを推薦することで付与するタグが他ユーザと違っていても有効に推薦できることを示した．

また，付与するタグを推薦する研究も行われている．小野ら [14] はタグ付けを自動化を実現する方法として機械学習を用いた．タグ付けを行うか行わないかという 2 クラスの分類問題として定義し，タグごとに分類器を生成した．

フォークソノミーのひとつであり，画像を Web 上にアップロードして共有できる Flickr [15] において，手動ではなく自動でタグを付与しようとする研究が行われている．新しくアップロードした画像が，すでにタグが付与されている画像と類似している場合，そのタグを用いるという手法 [16–18] があるが，画像の特徴と語句の種類が膨大な Flickr などのような Web 上の画像に対しては，適切な学習用データセットを作成することが難しいため，この手法を用いるのは困難である．また，Garg ら [19] は，ユーザのタグ付けの履歴や Flickr 全体でのタグ付けのデータからタグの共起を計算しタグ推薦を行った．Sigurbjörnsson ら [20] はユーザのタグ付けの履歴に対す

る naive Bayes と Flickr 全体での TF-IDF を用いてタグの推薦を行った． また， Takashita ら [21,22] はユーザのタグ付け履歴を用いた手法では推薦するタグのリストの精度がユーザの過去のタグ付けに依存しすぎているとし，ユーザの Web 閲覧情報を用いて，タグを推薦する手法を提案した．

本論文では，Takashita らと同様，ユーザ閲覧履歴を用いて，ソーシャルブックマークのタグ推薦を行う．

4 提 案 手 法

本研究では，ユーザの Web ページの閲覧履歴を用いてソーシャルブックマークのタグ推薦システムを提案する．提案手法の目的は，従来用いられている過去にユーザが付けたタグを参考にタグ推薦を行う手法の補助を行うことである．従来の手法では推薦されなかった Web ページに適したタグを推薦することで，従来手法に比べ，より適切なタグ付けを支援することができ，ユーザが再閲覧する際や，他者が閲覧した際にコンテンツの内容を明確に提示することができる．

提案手法で閲覧履歴を用いる理由を述べる．ユーザは興味を持ったコンテンツをブックマークすることで再閲覧すると考えられる．また，ブックマークを行わないコンテンツでも興味を持ったコンテンツに頻繁にアクセスしていると考えられる．このような閲覧履歴には，ユーザが興味を持っている分野の単語が出現していると考え，閲覧履歴を解析することでユーザへ推薦を行うための単語を抽出できるのではないかと考えた．

ユーザがブックマークしようとする Web ページを対象ページとする．対象ページに TF-IDF を用いて特徴語群を作成する．対象ページの特徴語群と付与しようとしているタグを用いて，閲覧履歴から抽出した特徴語群との関連度を計算する．対象ページの特徴語群との関連度が高かった閲覧履歴の単語，対象ページのタグとの関連度が高かった閲覧履歴の単語を推薦する．

本研究では，ユーザがブラウザで閲覧したページの履歴を情報源として用いてソーシャルブックマークのタグ推薦を行う手法を提案する． 図 1 は提案手法の流れである．図内の四角の中に行われる処理が書かれており，矢印上には処理に必要なものが書かれている．

図 1 の (1) を 4.1 節で説明する．この区間では Web 閲覧情報から推薦する単語群を抽出する．図 1 の (2) を 4.2 節で説明する．この区間ではタグ付け対象を解析し，推薦する単語を決定する．

ソーシャルブックマークを利用する際には，対象ページの URL と対象ページに付けたいタグを入力する．本研究では，図 1 の (1) と (2) の双方に含まれる Web ページ前処理の箇所を容易にするため，音楽，画像などのリッチコンテンツは除き，文字情報で構成されているコンテンツのみを対象とする．

4.1 Web 閲覧履歴からの語句抽出

（1） Web ページ閲覧

ユーザがブラウザを使い，Web 上のコンテンツを閲覧するとブラウザ内のキャッシュに Web の閲覧履歴が蓄積される．

（2） Web ページに対する前処理

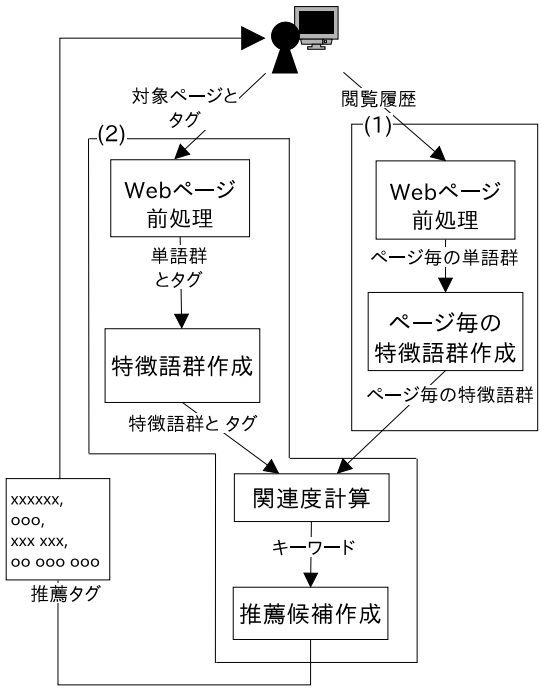

図 1 提案手法の流れ

キャッシュされた Web ページには HTML タグが含まれているため，そのままでは形態素解析がうまく行えない．HTML タグを除去し，文章のみの状態にする．文章のみの状態になった Web ページは形態素解析器にかけられ，形態素ごとに分割される．本研究では名詞を推薦するタグとするため，名詞のみを抽出する．このとき，頻出するがタグとして推薦されても役に立たないと判断したものは抽出しない．

（3） ページ毎に特徴語作成

閲覧したページごとに特徴語となる単語を抽出する．よく特徴語を抽出するときに用いられる手法である TF-IDF を本研究でも用いる． まず，TF を計算するため，ページごとに名詞の単語ごとの出現回数を調べ，ページ内での総単語数で割ることで正規化を行う．正規化を行うことで，文章の長い短いにかかわらず，割合としてみることができる．

また，IDF の計算には，検索エンジンを用いて，その検索エンジンでインデックスされているコンテンツの数，つまり，対象となる単語の検索結果の件数を DF とする．

ページごとに得られた単語の TF と単語に対する IDF を元に TF-IDF を計算し，ページ内で TF-IDF 値が大きい単語をその Web ページでの特徴語とする．本論文では，各 Web ページの TF-IDF 値が上位 10 % の単語をその閲覧履歴ページの特徴語とし，閲覧履歴全てのページでの特徴語を併合し，推薦候補群とする．

4.2 タ グ 推 薦

本研究では，ユーザがブックマークを行いたいページの URL とそのページにタグ付けしたいタグを入力ソースとし，そのページに対してタグを推薦する．

以下より，図 1 内の (2) の区間，タグ推薦の説明を提案手法の流れに沿って行う．

（1） ページ取得

ユーザがブックマークを行う対象ページの URL と対象ページに付与したいタグを入力すると，システムが対象ページを取得する．

（2） 前処理

HTML ソースから HTML タグを除去し，形態素解析を行い，名詞のみを抽出する

（3） 特徴語作成

抽出した名詞それぞれの TF-IDF を計算する．対象ページでの特徴語は，TF-IDF の値上位 5 % の名詞とする．

（4） 関連度計算

対象ページの特徴語，付与しようとしているタグと閲覧履歴の推薦候補群との関連度を計算する．本研究では [21, 22] と同様に式 (1) の NGD を用いて式 (2) を適用し，0 から 1 の範囲で関連度を計算する．式 (2) のパラメタ α は本研究では 1 とする．

$$NGD(v,w) = \frac{\max\{\log f(v), \log f(w)\} - \log f(v,w)}{\log G - \min\{\log f(v), \log f(w)\}} \quad (1)$$

$$Relevancy(v,w) = \exp[-\alpha \cdot NGD(v,w)] \quad (2)$$

NGD を提案手法で用いるため，5 章の実験では，5.1.2 節で述べるデータセットを母集団とし，$f(v)$ は v をタグにもつブックマークの数，$f(v,w)$ は v と w の両方をタグにもつブックマークの数とする．また， G はデータセットので 1 個以上タグが登録されているブックマークの総数を用いる．

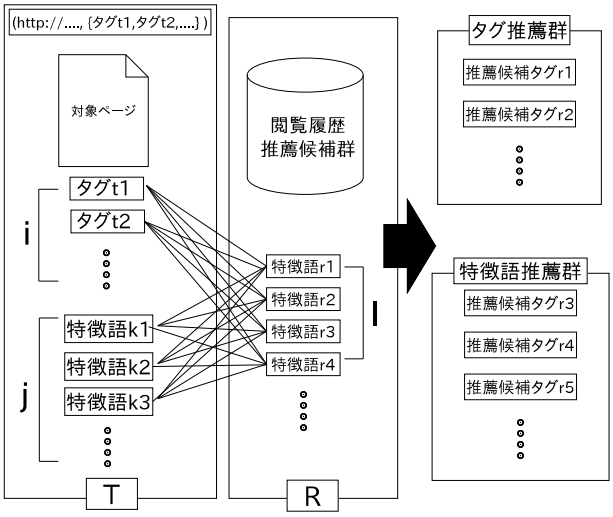

図 2 対象ページへのタグ推薦

（5） 推薦候補作成

図 2 は対象ページに対してタグ推薦を行っている図である．付与したいタグ i 個の集合 $\{t_1, t_2 \ldots, t_i\}$ ，また対象ページの TF-IDF の値，上位 5%までの特徴語 j 個の集合 $\{k_1, k_2 \ldots, k_j\}$ からなる対象単語群を T とし，4.1 節で行った閲覧履歴から抽出された推薦候補の単語 l 個からなる推薦候補群を R とし，対象単語群 T と 推薦候補群 R を用いて，図 2 のように関連度を計算する．なお，T と R は式 (3)，式 (4) で表される．

$$T = \{t_1, t_2, \ldots, t_j\} \cup \{k_1, k_2, \ldots, k_j\} \quad (3)$$

$$R = \{r_1, r_2, \ldots, r_l\} \quad (4)$$

推薦候補群 R 内の単語で，付与したいタグとの関連度，対象ページ内の特徴語との関連度，それぞれの値が高かった推薦候補群の単語 10 個をそれぞれに対する推薦するタグとする．

5 実　　験

5.1 準　　備

5.1.1 ページ取得

[21, 22] と同様に，提案手法で用いるユーザの閲覧履歴については，ユーザは検索結果の上位の Web ページを閲覧してると仮定し，Yahoo!デベロッパーネットワーク [23] の検索 API を用いて，興味のあるキーワードを含む Web ページを収集し，それを閲覧履歴として代用する．実験では，“ruby”，“東京”，“研究” の単語それぞれを興味の分野とし検索を行い，それぞれ 220 件ずつ，計 660 件のページをインターネットから収集した．収集したコンテンツにはブログが多く含まれており，ブログにはヘッダやサイドバーなど，本研究で注目したい記事内容がある本文部以外の文字情報も含まれている．そこで，Ruby ライブラリの ExtractContent [24] を用いて，対象コンテンツの本文を取得した．ExtractContent を用いることで，HTML ソースからヘッダやサイドバーなど，形態素解析を行う記事内容がある本文のみを抽出することができる．

本研究では，形態素解析器 MeCab [25] で形態素解析を行い名詞のみを抽出する．また，IDF を計算するために，本研究では Takashita ら [21] と同様に Yahoo!デベロッパーネットワークの Web 検索 API を用いる．対象の単語を含むドキュメント数である DF はその単語で Web 検索を行った検索結果の件数とし，総ドキュメント数を Takashita ら同様，Yahoo!に登録された総サイト数 190 億とした．本研究で用いた Web 検索 API は，日毎のクエリ上限が決まっているため，クエリ数を削減するため取得したページ全ての単語から重複を取り除き，API を実行した．

5.1.2 データセット

本研究で用いたソーシャルブックマークのデータは，livedoor labs が提供している EDGE Dataset [26] である．このデータは livedoor クリップのデータを CSV ファイル形式で 6ヶ月ごとに書き出したもので，本研究では，2006 年 6 月から 2010 年 6 月までにブックマークされたデータを用いる．

データの詳細を表 1 に示す．また，同様のクエリでデータセットから取得し，96 件の Web ページを対象とした．また，式 (1) の NGD で用いる G は表 1 のブックマーク数からタグを含んでいないブックマークを除いた 2,285,878 件とする．

5.1.3 従来手法として用いるデータ

従来の手法であるユーザが過去につけたタグの履歴を用いた推薦に関しても実験を行った．サイト登録数，タグ登録数に着目した．着目した理由として，サイト登録数をみることでソーシャルブックマークをよく利用しているユーザを確認すること

表 1　データセットの詳細

項目	個数
ユーザ	55,278
URL	410,059
タグ	189,239
ブックマーク	3,005,129

ができ，タグ登録数をみることで，ユーザの語彙力を確認できると考えたためである．

サイト登録数が多くタグ登録数が多いユーザ，サイト登録が多くタグ登録数が平均的なユーザ，サイト登録数が平均的でタグ登録数が多いユーザを想定して，これらのユーザが登録しているタグ全てとユーザが対象ページに付けたいタグとの関連度を計算し，提案手法同様関連度の高いタグを推薦する．以降，サイト登録が多くタグ登録数が多いユーザを A，サイト登録が多くタグ登録数が平均的なユーザを B，サイト登録数が平均的でタグ登録が多いユーザを C とする．

表 2 はユーザを想定するときに用いた値であり，平均値と平均値以上の範囲での平均値，平均値以上の範囲での平均値と最大値の間での平均値前後 20 %の値から二人ずつユーザを抽出した．

表 2　サイト登録数，タグ登録数の平均値

	平均値	平均値と平均値以上の範囲での平均値	平均値以上の範囲での平均値	平均値以上の範囲での平均値と最大値の間での平均
サイト登録数	54.36	186.56	542.16	2005.80
タグ総数	45.70	104.10	310.52	955.26

5.2 結　　果

以降，クエリ毎に例を上げて結果を記す．各推薦候補がサイトの内容にマッチしているかは，実際に目視し確認した．

5.2.1 東京に関するページへのタグ推薦結果

表 3 は付与したいタグと各手法で推薦されたタグのリストである．サイトの内容としては，東京の飲食店の紹介であった．店舗の住所が記載されており，写真が多用されていた．結果に対する考察として，実際にページを確認したところ，東京のクエリで取得した 36 件中従来手法では，各推薦候補で平均 0.5 個マッチしており，提案手法では，3.4 個マッチしていた．地名が推薦されている候補が多いが，これは従来手法では，店舗やイベントに関するコンテンツをブックマークした際に，タグにコンテンツの内容と共に，その店舗やイベントが開催される地域を入力することが多く，推薦されたと考えられる．また，提案手法で閲覧履歴とタグの関連度の候補について地名が多く推薦されたが，データセットで “東京” が付けられたブックマークを確認したところ，“東京” というタグと共に他の地名や店舗名をタグとして付与し，“東京” のタグに興味を持ったユーザを誘導しようとしているユーザが多く確認できた．これは，データセットを提供している Livedoor labs がスパム活動を行っているユーザは削除せずにデータセットとして提供していたためだと考えられる．スパム活動を行っているユーザをデータセットから除外することで，純粋にブックマーク対象に興味を持っているユーザの評価を用いて推薦が行うことができ，タグ推薦の質を向上することができると考えられる．

表 3　“東京” をクエリとしてデータセットから抽出されたサイトへの従来手法と提案手法のタグ推薦結果

付与したいタグ	東京, 東京グルメ, グルメ, 東京居酒屋
タグ履歴と付与したいタグの関連度を用いた推薦:A	名古屋, 格闘技,jr, 料理, 地下鉄, 食, イベント, バイク, 野球, ローカライズ
〃	スター, 大阪, 福岡, 銀行, 名古屋, 横浜, 店, 東京スター銀行, 料理, 東京
タグ履歴と付与したいタグの関連度を用いた推薦:B	food, まとめ,shopping,ide,news, ネタ ,api,tmp, 資料, デザイン
〃	家電, 周辺機器, ランキング, 野球, 学校, スポーツ, 生活, 健康, トリビア, メール
タグ履歴と付与したいタグの関連度を用いた推薦:C	食べ歩き, おかず, グルメ, 天神, 埼玉, ガイドブック, 居酒屋, ワイン, 食材, 格闘技
〃	エステ, 天神, 居酒屋, 整体, 食材, 料理, 位置, 勉強, 観光, 宿
閲覧履歴と付与したいタグの関連度を用いた推薦	東京, 場所, 銀座, 渋谷, 大阪, 福岡, 博多, 名古屋, カフェ, 札幌
閲覧履歴とページ内特徴語の関連度を用いた推薦	表示, 料理

5.2.2 Ruby に関するページへのタグ推薦結果

表 4 は付与したいタグと各手法で推薦されたタグのリストである．サイトの内容としては，プログラミング言語 Ruby で対話的に操作が行うことができる irb (interactive Ruby) コマンドの設定の紹介であった．結果に対する考察として，実際にページを確認したところ，Ruby のクエリで取得した 31 件中，従来手法では，各推薦候補で平均 0.5 個マッチしており，提案手法では，2.7 個マッチしていた．タグに注目すると従来手法では，プログラミング技術に関する単語を推薦するユーザと勉学に関する単語を推薦しているユーザに分かれている．従来手法で用いたユーザはブックマークの登録数とタグの総登録数で分類し，ユーザの興味関係なしに抽出した．その結果，プログラミング技術に興味を持つユーザは，付与したいタグである “Ruby” に関連した単語を推薦できたが，興味を持っていないユーザは語彙がないため関連した単語を推薦できなかった．提案手法では，“Ruby” に関する Web ページを閲覧しているユーザを想定して，推薦候補群を作成したので “Ruby” に関連した単語を推薦できた．ユーザが興味を持つためには閲覧をしていき，興味分野の知識が高まったある段階でブックマークをしていくと考えられる．そのため，ある分野に対して興味を持ったユーザが初めて，その分野に対してブックマークを行う際，そ

表 4 “Ruby” をクエリとしてデータセットから抽出されたサイトへの従来手法と提案手法のタグ推薦結果

付与したいタグ	ruby
タグ履歴と付与したいタグの関連度を用いた推薦:A	pragger,rubygems,sinatra,irb,rspec, merb,rubycocoa,nokogiri,python,cocoa
〃	試験, 翻訳, サーバー, ホスティング,web, レンタルサーバー, キャッシュ,3, 一覧,mp3
タグ履歴と付与したいタグの関連度を用いた推薦:B	rails,framework,netbeans,ide,test, vim,c,c++, コーディング規約,php
〃	tips,html, プログラム,web アプリ,twitter,amazon, マリオ, サーバー, windows,linux
タグ履歴と付与したいタグの関連度を用いた推薦:C	一覧, 勉強, 説明, ブラウザ, 学習, 本, 便利, 人生, 教育, 音楽
〃	解説, 試験, 勉強, 資格, 経験, 初心者, 便利, 経営, 会社, 仕事
閲覧履歴と付与したいタグの関連度用いた推薦	gem,gems,irb,rexml,yugui,sdl,gettext, プログラミング,exerb,dsl
閲覧履歴とページ内特徴語の関連度を用いた推薦	プログラム, 教科書,dhh, クラス, 自動,uri, 育成,buffer, 製作, 分散

表 5 “研究” をクエリとしてデータセットから抽出されたサイトへの従来手法と提案手法のタグ推薦結果

付与したいタグ	研究, 論文
タグ履歴と付与したいタグの関連度を用いた推薦:A	文章,trac, 猫, アルゴリズム,cygwin, 音声合成, スレッド, 画像認識, 音声認識,gps
〃	レポート, 研究, 読書感想文, 徹底, 再出発, 会, 攻略法, 国民, 冬のソナタ, コンタクトレンズ
タグ履歴と付与したいタグの関連度を用いた推薦:B	資料,module,plugin, まとめ,os, ネタ ,browser,google,network,wordpress
〃	プログラム, マネー, 日本, ブラウザ, 男女, 仕事, まとめ, デジカメ,os, ネタ
タグ履歴と付与したいタグの関連度を用いた推薦:C	物理学, 医学, 化学, 教育, 中心, 飛行機, 病気, 英語, 勉強, 学習
〃	診断, 不安, 進化, 構造, 小川, 上田, 試験, 中心, 飛行機, 病気
閲覧履歴と付与したいタグの関連度を用いた推薦	論文, 研究, 書き方, 大学院, 仮説, 発表, 大学, サイエンス, 学会, 記述
閲覧履歴とページ内特徴語の関連度を用いた推薦	締切, 場合, 会議, 参加, 投稿, ファイル, 論文, 治験, 講師, 注文

の分野に関するタグはまだ蓄積されていないが，閲覧履歴は蓄積されていると考えられる．

5.2.3 研究に関するページへのタグ推薦結果

表 5 は付与したいタグと各手法で推薦されたタグのリストである．サイトの内容としては国際学会に投稿する際に気をつけておきたいことが書かれている．結果に対する考察として，実際にページを確認した所，研究のクエリで取得した 30 件中，従来手法では，各推薦候補で平均 0.6 個マッチしており，提案手法では 3.8 個マッチしていた．従来手法だと，研究に関するコンテンツにふさわしくない “冬のソナタ” や “コンタクトレンズ” などが推薦されており，研究に関する興味がないユーザが選ばれており，うまくタグ推薦が行われていないことがわかる．

5.3 閲覧履歴が少ない場合における提案手法によるタグ推薦

前節までの実験では，“東京”，“Ruby”，“研究” に関するページを 220 件ずつ閲覧履歴として用いたが，ここでは閲覧履歴が少ない場合での提案手法によるタグ推薦を考える．これは，5.2.2 節で述べた，ユーザが新たに興味を持った際に，閲覧履歴を用いた提案手法に比べ，従来手法では興味の分野での適切なタグ候補の蓄積に時間がかかると考えられるため実験を行った．各キーワードで検索を行い，PageRank のスコアを Google Toolbar [27] を用いて確認し，高いスコアをもつページ上位 10 件，50 件を閲覧履歴とした．表 6 は閲覧履歴が少ない場合の提案手法の推薦結果である．

表 7 は閲覧履歴が 10 件，50 件，220 件の時に適切に推薦されたタグの個数の平均である．10 件や 50 件の時のように閲覧履歴が少ない，つまり興味を持ち始めた状態でも適切にタグ推薦が可能である．また，10 件から 50 件のように閲覧履歴が増えた場合ではタグ推薦の個数が微増している．実際に推薦されたタグを確認したところ 10 件の時に推薦されていたタグが 50 件の時には推薦されていなかったり，まったく変動がなかったものがあった．閲覧履歴が少ない場合でも提案手法では Web ページの内容に適したタグを推薦できていると考えられる．情報源としての閲覧履歴を選定することで，閲覧数が少ない場合でもより適切に推薦が可能になると考えられる．

5.4 考　　察

ユーザが興味を持っているものに対しては，従来手法，提案手法でも適切にタグ付けが行えたと考えられる．しかし，ある分野に対して興味を持ち始めたユーザがブックマークしたページに対してタグ推薦を行う際，従来手法では，興味のある分野に適切なタグが蓄積されるのに時間がかかるのに対して，閲覧履歴を用いる提案手法では，興味を持つ間にユーザはその分野に関するコンテンツを閲覧している．提案手法ではコンテンツから特徴語を抽出し，推薦候補群とするためタグ候補が蓄積されており，ユーザがブックマークした際に，適切にタグ推薦が行うことができると考えられる．

また，閲覧しているページの文章の少ない場合はページ内特徴語を用いたタグ推薦の単語数が少ない．本研究の場合，閲覧しているページ内の特徴語と閲覧履歴の関連度を調べ，閲覧しているページの TF-IDF の上位 5%までの特徴語を用いるため，特徴語となる単語が少なくなり，推薦する単語の数も少な

表 6　閲覧履歴が少ない場合の提案手法の結果

クエリ	手法	推薦されたタグ
東京	タグ (10 件)	東京, 場所, 銀座, 渋谷, 福岡, 埼玉, 周辺, バス,jr, 浅草
	ページ内特徴語 (10 件)	
	タグ (50 件)	東京, 場所, 銀座, 渋谷, 大阪, 福岡, 名古屋, 神奈川, 埼玉, 周辺
	ページ内特徴語 (50 件)	表示
Ruby	タグ (10 件)	gem, プログラミング,perl, インタプリタ, 実装, 開発, 言語, 分散, ライブラリ, 公式
	ページ内特徴語 (10 件)	ライブラリ,gem, 利用, 場合, 学校, 開発, 注意, 路線, 計算, 言語
	タグ (50 件)	require, ライブラリ,irb,gem, 利用, 被害, 住人, 使用,nil, 仕組み
	ページ内特徴語 (50 件)	irb,gem,gems,nokogiri,sdl, プログラミング, コンパイル,bundler,grep,perl
研究	タグ (10 件)	論文, 研究, 仮説, 公開, 教育, 分散, 段階, 検索, 時間, 問題
	ページ内特徴語 (10 件)	論文, 動作, 検討, 研究, 復元, 注意, 運営, 仮説, 編集, 掲載
	タグ (50 件)	論文, 研究, 大学院, 仮説, 発表, 学会, 記述, 理論, 科学, 修正
	ページ内特徴語 (50 件)	会議, 投稿, ファイル, 論文, 領域, 表示, 動作, 検討, 研究, 復元

表 7　閲覧履歴がそれぞれ 10 件，50 件での推薦結果

	10 件での平均	50 件での平均	220 件での平均
東京	2.21	2.43	3.4
Ruby	2	2.33	2.7
研究	2.02	2.33	3.8

くなったと考えられる．

6 ま と め

本研究では，ユーザの Web ページの閲覧履歴を用いたソーシャルブックマークにおけるタグ推薦を提案した．

まず，ユーザは日常的に興味のある Web ページを閲覧しているとし，その閲覧履歴を蓄積する．蓄積された閲覧履歴から Web ページを取得する．次に，Web ページから本文を抽出し，形態素解析を行い，単語単位に分割する．分割後，単語の品詞と用法を識別子，名詞に限定し，Web ページごとの TF-IDF 値を計算する．TF-IDF 値が高い単語をその Web ページの特徴語として抽出し，閲覧履歴の全ての特徴語を併合し，推薦候補群とする．

次に，ユーザがソーシャルブックマークサービスを利用し，ブックマークを行う．ブックマークを行う Web ページの特徴語を閲覧履歴の時と同様の手法で抽出する．また，ブックマークを行う際に，そのページに付与したいタグを入力するので，その付与したいタグを取得する． Web ページから抽出された特徴語と付与したいタグをそれぞれ閲覧履歴の推薦候補群との関連度を計算する．関連度の計算には，NGD を用いた．NGD では，対象となる二単語の出現数が必要となるため，本研究では Livedoor が提供しているデータセットを用いた．また，比較のため，従来用いられている既に他のコンテンツに付けたタグを用いたタグ推薦を行った．

結果として，提案手法では，従来手法で推薦されたタグとは異なった単語を推薦することができ，また，内容にマッチしたタグを推薦できていた．また，ユーザが新たに新しい分野に対して興味を持ち始めたと仮定して実験を行った結果，閲覧履歴が少ない場合でも適切にタグ推薦が行われていた．従来手法では，ユーザが新たに新しい分野に興味を持ち始めた場合，推薦候補となるタグの蓄積には時間がかかるが，提案手法では短時間で推薦候補となるタグを蓄積できる点で有用だと言える．

今後の課題として，本研究では，ユーザの興味の度合いなどを考慮していなかった．興味の度合いに応じて重み付けを行い，評価を行う必要がある．また，本研究では興味のある分野を著者が手動で指定し，閲覧履歴を取得したが，興味のある分野を閲覧履歴を元に機械学習などで推定することによって，推薦システムを自動化することが必要である．

本研究では，閲覧履歴を全て併合し，推薦候補群としたが，推薦候補群に階層的クラスタリングを用い，ユーザが興味を持ち閲覧を薦めていくことによる知識の蓄積に応じて，ユーザが付与しようとしているタグやページ内特徴語に対して，浅い知識を持ったユーザには一般的な単語，深い知識を持ったユーザには専門的な単語を推薦することでよりユーザのとって有益なタグ推薦が行うことが可能であると考えられる．

文　献

[1] 松尾豊, 福田隼人, 石塚満. ユーザ個人の閲覧履歴からのキーワード抽出によるブラウジング支援. 人工知能学会論文誌, Vol. 18, No. 4, pp. 203–211, 2003.

[2] Thorsten Joachims, Dayne Freitag, and Tom Mitchell. Webwatcher: A tour guide for the world wide web. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pp. 770–775. Morgan Kaufmann, 1997.

[3] Liren Chen and Katia Sycara. Webmate: a personal agent for browsing and searching. In *Proceedings of the second international conference on Autonomous agents*, AGENTS '98, pp. 132–139, New York, NY, USA, 1998. ACM.

[4] Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka. Can social bookmarking enhance search in the web? In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '07, pp. 107–116, New York, NY, USA, 2007. ACM.

[5] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing web search using social annotations. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pp. 501–510, New York, NY, USA, 2007. ACM.

[6] 百田信, 伊東栄典. ソーシャルブックマークに基づく情報発見. データ工学ワークショップ, 2008.

[7] はてなブックマーク. `http://b.hatena.ne.jp/`.

[8] Oded Nov and Chen Ye. Why do people tag?: motivations for photo tagging. *Commun. ACM*, Vol. 53, pp. 128–131, July 2010.

[9] del.icio.us. `http://www.delicious.com/`.

[10] livedoor クリップ. `http://clip.livedoor.com/`.

[11] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, ed-

itors, *The Semantic Web: Research and Applications*, Vol. 4011 of *Lecture Notes in Computer Science*, pp. 411–426, Heidelberg, June 2006. Springer.

[12] 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. 情報処理学会論文誌, Vol. 47, No. 5, pp. 1382–1392, 2006-05-15.

[13] 佐々木祥, 宮田高道, 稲積泰宏, 小林亜樹, 酒井善則. Social bookmark におけるコンテンツクラスタ間の類似度を用いた web コンテンツ推薦システム. 情報処理学会論文誌. データベース, Vol. 48, No. 20, pp. 14–27, 2007-12-15.

[14] 小野裕作, 當間愛晃, 遠藤聡志. ソーシャルブックマークサービスを利用したタグ付け自動化システム開発に関する一考察. 第 22 回人工知能学会, 2008.

[15] Flickr. `http://www.flickr.com/`.

[16] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition*, CVPR'04, pp. 1002–1009, Washington, DC, USA, 2004. IEEE Computer Society.

[17] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pp. 119–126, New York, NY, USA, 2003. ACM.

[18] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*. MIT Press, 2003.

[19] Nikhil Garg and Ingmar Weber. Personalized, interactive tag recommendation for flickr. In *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys '08, pp. 67–74, New York, NY, USA, 2008. ACM.

[20] Börkur Sigurbjörnsson and Roelof van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pp. 327–336, New York, NY, USA, 2008. ACM.

[21] Taiki Takashita, Tsuyoshi Itokawa, Teruaki Kitasuka, and Masayoshi Aritsugi. Tag Recommendation for Flickr Using Web Browsing Behavior. *Computational Science and Its Applications–ICCSA 2010*, pp. 412–421, 2010.

[22] Taiki Takashita, Yuki Abe, Tsuyoshi Itokawa, Teruaki Kitasuka, and Masayoshi Aritsugi. Design and implementation of a system for finding appropriate tags to photos in Flickr from Web browsing behaviour. *International Journal of Web and Grid Services (IJWGS)*, Vol. 7, No. 1, pp. 75–90, 2011.

[23] Yahoo!デベロッパーネットワーク. `http://developer.yahoo.co.jp/`.

[24] Extractcontent. `http://rubyforge.org/projects/extractcontent/`.

[25] Mecab. `http://mecab.sourceforge.net/`.

[26] Edge datasets. `http://labs.edge.jp/datasets/`.

[27] Googole toolbar. `http://toolbar.google.com/`.