ソーシャルブックマークと Wikipedia を用いた Web ページ推薦と重要文抽出手法

吉田 拓実† 井上 潮‡

†‡東京電機大学大学院 工学研究科 〒120-8551 東京都足立区千住旭町５番

E-mail: †12kmc36@ms.dendai.ac.jp, ‡inoue@c.dendai.ac.jp

あらまし ユーザが Web ページを検索する際，検索クエリの試行錯誤が必要である．また，興味・関心のある Web ページを検索結果から探す必要がある．ユーザの興味のある Web ページを提示する方法として，様々な推薦システムが研究，開発されてきた．しかし，これらの推薦システムでは，検索結果と同様にページのタイトルと URL のリストで提示されることが多く，ユーザは興味のあるページを探す必要がある．

本研究では，ソーシャルブックマークと Wikipedia のカテゴリ情報を用いて，タグに関連した情報を含んだ Web ページを推薦対象とするとともに，推薦対象ページから HTML タグを手掛かりに重要文を抽出して提示することで，ユーザの煩わしさを軽減するシステムを開発する．

キーワード 情報推薦，ソーシャルブックマーク，Wikipedia，情報抽出

1. はじめに

近年，ブログや SNS，Web 上の情報をまとめるキュレーションサービスなど，ユーザが自由に情報を発信することができるサービスの増加につれ，Web 上の情報量は増加傾向にあり「情報過多」が問題になっている．情報過多の影響により検索エンジンで Web ページを検索する際，目的とするページを見つけ出すためには検索クエリの試行錯誤が必要である．

さらに，目的とする情報が存在するかは，検索結果で提示された Web ページを閲覧していく必要がある．検索結果で提示された Web ページは膨大であるため，目的の情報を得るまで手間や時間がかかるという問題がある．

検索クエリの試行錯誤に関しては，閲覧履歴から関連情報の検索・自動提示を行う研究[1]や、ソーシャルブックマークを用いて情報を提示する研究[2]などがある。しかしながら，Web ページのタイトルと URL を提示するだけで，情報の取捨選択はユーザに一任されていることが多い．

目的とする情報を見つけやすくする研究としては，Web ページからキーワードや重要文を抽出することで要約を提示することにより，ユーザの煩わしさを解消することが行われている．

本研究では，検索時の煩わしさの解消として，ブックマークを不特定多数のユーザと共有するサービスであるソーシャルブックマークのタグ情報と Web 上の百科事典である Wikipedia のカテゴリ情報を用いて，ユーザのブックマーク傾向と同じような Web ページを推薦対象とする．その後，推薦対象ページから目的の情報を探し出す手間の軽減のために，HTML 制作者が重要な語句を強調するために用いる HTML タグを手掛かりとして，重要文を抽出することでユーザの負担を軽減する手法を提案する．

2. 関連研究

2.1. 情報推薦

石橋ら[3]は，ソーシャルブックマークからタグのグループを作成し，ユーザの嗜好推定を行い，類似ユーザを発見し Web ページを推薦する手法を提案している．ユーザが新規にブックマークを登録する際のタグと URL の関係の推定をベイズ推定によって行う．しかし，推薦される URL は類似度が近いユーザ 1 名のブックマークしている URL のリストであるため，ユーザの嗜好と異なるページが推薦される場合がある．

2.2. 情報抽出

柴田ら[4]は，Web ページ内からキーワードと重要文を抽出する手法により，要約提示を行うモデルを提案している．キーワードの重要度は，TF-IDF 法による重みづけの手法を用いている．また，重要文抽出は，重要キーワードを多く含むほど文章重要度が高くなる重みづけ手法を用いている．その結果，人間の負担は軽減できるが，処理時間がかかるため HTML タグに着目し，単語の重要度の精度を高めるべきと述べている．

3. 本研究のアプローチ

本研究では，検索時の試行錯誤の煩わしさの解消と，目的の情報を探し出す際の煩わしさの解消のために情報推薦と情報抽出を組み合わせたシステムを目指す．本システムは，Web ページ推薦部と情報抽出部から構成する．Web ページ推薦部では，ソーシャルブックマークのタグ情報に着目し，タグの関連語として，Wikipedia のカテゴリ情報を用いた推薦を行う．既存の

推薦システムの研究の場合，推薦される Web ページが単調になってしまう場合がある．本研究では，タグがつけられている Web ページだけではなく，タグに関連する情報を含む Web ページをユーザに推薦することで，単調さをなくし新たな発見をユーザに与えることが可能である．

情報抽出部では，推薦された各 Web ページの HTML タグに着目し，各ページの重要なテキストを抽出することで，ユーザに重要な情報を提示する．

4. Web ページ推薦手法

本研究の Web ページ推薦手法は 2 つの段階で行う．1 つめは，類似ユーザの抽出である．2 つめは，推薦 Web ページの抽出である．

4.1. 前提条件と仮説

Web ページ推薦に関して以下の前提条件を設ける．

- ソーシャルブックマークユーザは興味や関心がある Web ページに対して，タグ付けを行う．
- ユーザはタグ付けにより Web ページを整理する．

以上の前提条件より以下の仮説を立てる．

(1) ユーザが Web ページにつけたタグは，ユーザの興味関心を示す．

(2) ユーザが特定のタグを多く使用するほどそのタグに対して興味関心が強い．

(3) あるユーザにとって，類似したタグ付傾向を示す他のユーザのブックマークは役に立つ．

なお，(1), (2), (3)の仮説については参考文献[5][6][7]などで述べられており、本研究の提案手法もこれらの仮説をもとに行う。

4.2. Web ページ推薦アルゴリズム

上記の仮説をもとに Web ページ推薦アルゴリズムを提案する．Web ページ推薦アルゴリズムのフローチャートを図 1 に示す．

はじめに，類似ユーザの抽出として，ユーザがタグ付けしている回数が多いタグ上位 n 件を取得する（以下タグセット）．その後類似ユーザとしてタグセットに含まれるタグをすべて付けている他のユーザを取得する．類似ユーザが x 人以下の場合，推薦できる Web ページ数が少ないことが考えられるため，タグセットの中で上位 m 件のタグを Wikipedia のカテゴリ情報を用いて解析し解析結果をタグとしてタグ付けしているユーザを追加取得する．このことにより，タグの関連語をつけているユーザも類似ユーザとして取得できる．n=5 の時の類似ユーザ抽出例を図 2 に示す．

次に推薦 Web ページの抽出として

タグセットに含まれるタグ全てを Wikipedia のカテゴリ情報を用いて，タグを拡張した結果を拡張タグセットとする．類似ユーザがブックマークした Web ページの中から，タグセットに含まれるタグ，および拡張タグセットに含まれるタグのいずれかが付けられている Web ページを抽出する．推薦する Web ページの有用性を考慮して複数人がブックマークしている Web ページを推薦する．

Wikipedia カテゴリ情報を用いた，タグセットの拡張例を図 3 に示す．

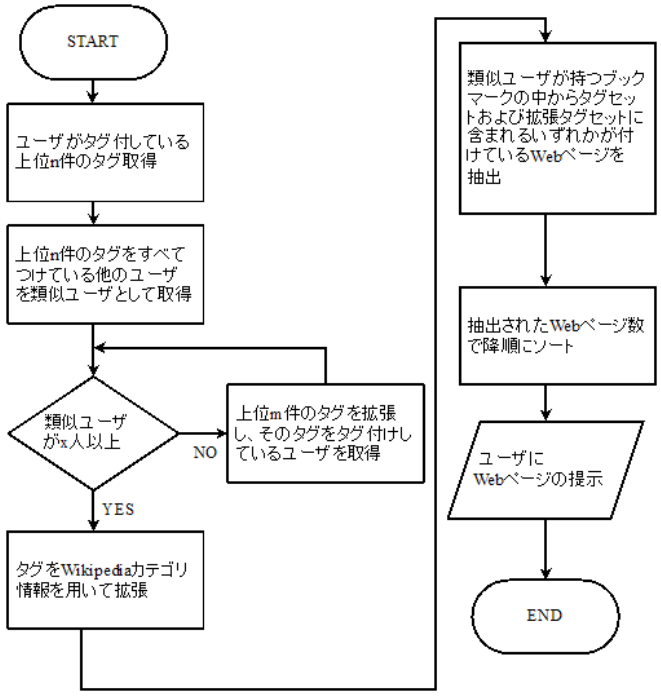

図 1 Web ページ推薦アルゴリズム

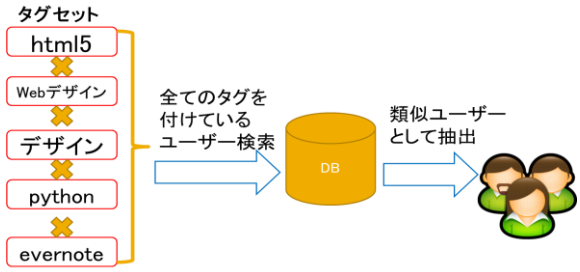

図 2 類似ユーザ抽出例

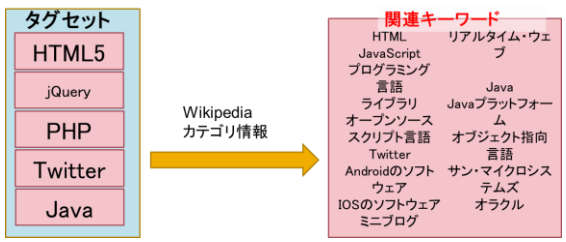

図 3 タグセットの拡張例

5. 情報抽出手法

本研究では，Web ページの制作者は重要な語句を強調するために HTML タグを使うと仮定し，Web ページ内の重要な文章を抽出する．

5.1. 前提条件と仮説

情報抽出に関して以下の前提条件を設ける

- ユーザが知りたいことは，重要な文章である．
- 重要な文章は，重要な単語を多く含む．
- 重要な文章は，Web ページ上でスクロールを行わなくとも閲覧できる範囲内にある．

以上の前提条件より以下の仮説を立てる．

(1) Web ページ制作者は，重要な単語に HTML タグで強調を行う．

(2) 近年は，サービス固有のエディタを使用することにより，HTML タグを意識せずに単語を強調する可能性がある．そのため，強調 HTML タグの使用方法はドメインごとに異なる可能性がある．

(3) 従ってドメインごとの特性を考慮した強調 HTML タグから重要な文章を抽出できる．

この仮説の裏付けとして以下の 2 つの調査を行い確認した．

(1) 強調 HTML タグの調査

(2) ドメインごとの強調 HTML タグの調査

5.1.1. 強調 HTML タグの調査

仮説 1 の確認のために，本研究で収集した URL40 万件から，無作為に抽出した 500 件について，どの強調 HTML タグが多く使われているかを調査した．調査対象とした HTML タグと個数，用途を表 1 に示す．この結果，Font・B・Strong などの，単語を直接強調する HTML タグの割合が多いことが分かった．そのため，Web 制作者は HTML タグで強調を行うことが確認できた．

表 1　調査対象 HTML タグと用途及び個数

タグ	用途	個数
H1	見出し	589
H2		1835
H3		2555
H4		1439
H5		329
H6		40
Strong	強調	3764
em	強い強調	844
Font	フォントの色，サイズ	6218
B	ボールド（太字）	4400
I	斜体	610
U	下線	163

5.1.2. ドメインごとの強調タグ調査

次に，仮説 2 の確認のために，情報推薦システムでは，どのドメインが多いかを調査した．上位 10 ドメインの結果を表 2 に示す．

その後，表 2 に含まれる 1 ドメインにつき 500 件を無作為に抽出し，調査した．なお，調査対象 HTML タグは前節と同様である．結果の一例としてブログサイトである d.hatena.ne.jp(以下 hatena)，ニュースサイトである gigazine.net（以下 gigazine）を図 4 に示す．

表 2 収集した上位 10 ドメイン

ドメイン URL	サイト種別	ページ数(件)
blog.livedoor.jp	ブログ	21452
d.hatena.ne.jp	ブログ	20952
matome.naver.jp	まとめ	18508
blog.fc2.com	ブログ	13513
lifehacker.jp	ニュース/ブログ	8848
cookpad.com	レシピ	7935
togetter.com	まとめサービス	7837
gigazine.net	ニュース/ブログ	5532
yomiuri.co.jp	新聞社	3379
nikkei.com	新聞社	3284

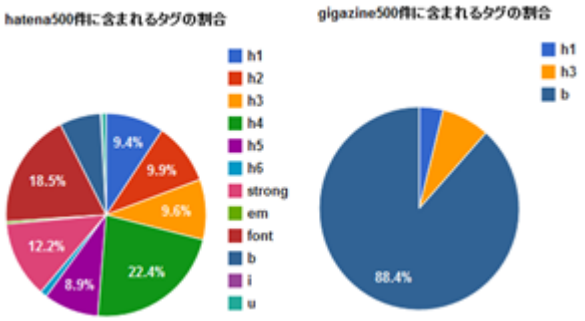

図 4 ドメインごとの強調タグ

図 4 から，gigazine のドメインでは B タグが多く使われていることがわかる．また，hatena のドメインでは様々なタグが使われていることがわかる．以上のことからドメインごとに強調 HTML タグの使用傾向が異なることが確認できた．

5.2.重要文抽出アルゴリズム

以上の仮説をもとに重要抽出アルゴリズムを提案する．重要文抽出アルゴリズムのフローチャートを図 5 に示す．本研究では，情報推薦が行われた Web ページのリストに対して重要文を抽出する．また，ドメインごとの特徴を考慮し，強調 HTML タグが含まれる文を重要文として提示する．そのため，初めに推薦された Web ページのリストから最初の 1 件の URL を対象 URL として取得するその後，対象 URL の HTML タグ

から表 1 の見出し以外の HTML タグの個数を取得する．次に，対象 URL のドメインを取得し，ドメインがデータベースに登録されている場合，これまでの総タグ数を取得し，対象 URL の HTML タグ数に加算する．加算後の総 HTML タグ数をデータベース内で上書きする．ドメインが登録されていない場合は，ドメインと総 HTML タグ数を新規登録する．

次に，HTML タグ内から
タグと対象ドメインで一番多く使われているタグ以外を除去する．

最後に対象ドメインで一番多く使われている HTML タグが含まれる文章を取得し提示する．ページ前半に重要な文章があるとした仮説を踏まえ，x 件の文章を重要文として提示する．これを推薦された Web ページに対してすべて行う．

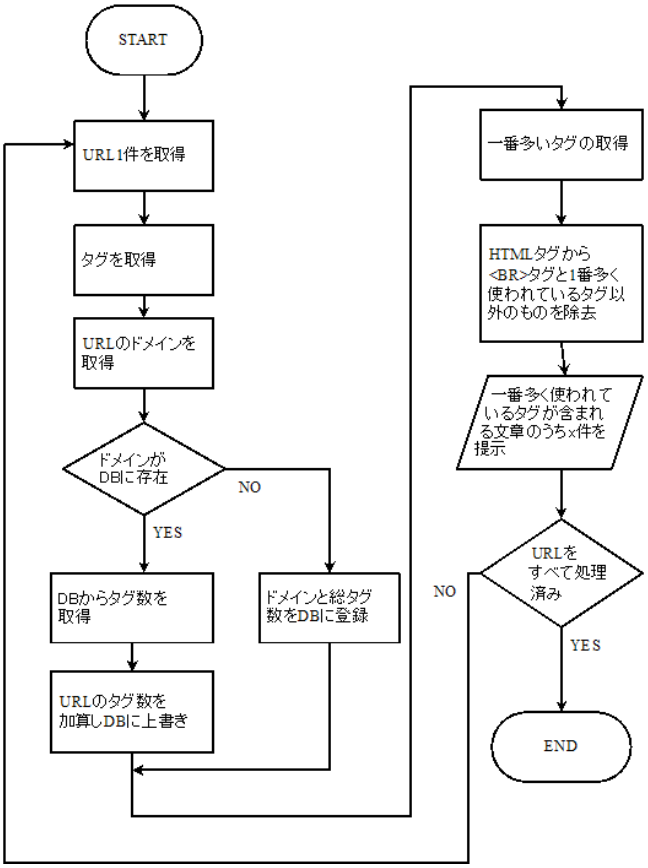

図 5 重要文アルゴリズム

6. 評価

本手法の有用性を検証するために 2 つの評価を行った．1 つめは，ユーザにとって有用な推薦 Web ページが増加したかどうかである．

2 つめは，推薦 Web ページから重要文が抽出できたかどうかである．

6.1. 実験システム

評価を行うため，実験システムを構築した．実験システムの構成を図 6 に示す．本システムは，ソーシャルブックマークのデータを収集するクローラと Web ページの推薦・表示を行う推薦部，重要文を抽出する重要文抽出部からなる．今回は推薦，重要文抽出それぞれの有用性を確認するため評価は別々に行った．

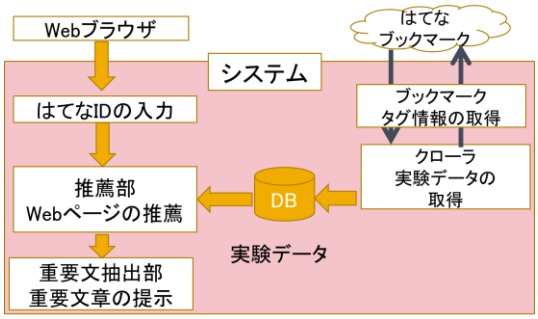

図 6 実験システム構成図

本実験では，国内で良く用いられているはてなブックマーク[9]のデータを用いた．使用したデータは，クローラを用いて収集しデータベースに格納したものを利用した．また，Wikipedia のカテゴリ情報は提供されている 2012 年 5 月 4 日のダンプファイルを用いる．

データの収集に関しては，ソーシャルブックマークのデータを収集するクローラを構築する．クローラは，はてなブックマークエントリー情報取得 API と，はてなブックマークページのユーザ個人の RSS の解析によってブックマークしている Web ページタイトル，URL，タグを取得する．

推薦部においては，4 章で述べた方法により類似ユーザの抽出と Web ページの抽出を行う．なお，評価に当たり処理時間の関係上，類似ユーザが 50 人以上の場合，50 人になるように無作為に抽出した．また，Wikipedia のカテゴリ情報を使用しない場合と使用する場合の推薦を行えるようにした．

重要文抽出部においては，5 章で述べた方法により重要文抽出を行った．

6.1.2.評価パラメータ

今回の評価にあたり，パラメータとして上位 n=5 件のタグを取得し，類似ユーザ x=10 人以下の場合上位 m=1 件を Wikipedia のカテゴリ情報を用いて解析し解析結果をタグとしてタグ付けしているユーザを追加取得した．

サンプルユーザとしては，2012 年 6 月 6 日時点で，はてなブックマークで多く用いられているタグ上位 n=5 件をタグ付けしているユーザを 6 人ずつ計 30 人を無作為に抽出し，評価に使用するサンプルユーザとした．

実験データとして 2012 年 5 月 19 日から 6 月 6 日までクローラで取得したデータを使用した．表 3 に 6 月 6 日時点のデータ数を示す．

表 3　実験データ

ユーザ数(人)	27,643
総ブックマーク数（重複ふくむ）	110,886
総タグ数（重複ふくむ）	114,253

6.2. Web ページ推薦評価

本手法の有用性を検証するために Wikipedia のカテゴリ情報を使用しない場合と使用した場合の適合サイト数，不適合サイト数，適合率を評価した．

推薦された Web ページがユーザにとって有益かどうかは主観で判断した．主観で判断するにあたり，以下の基準を設けた．Web ページのタイトルにタグセットに含まれるタグが含まれる場合，適合していると見なす．タイトルに入っていない場合，はてなブックマークおすすめタグの中にタグセットに含まれるタグの有無で適合・不適合を判断する．

サンプルユーザ 30 人中 27 人に対して Web ページの推薦が行えた．しかし，3 人に対しては Web ページの推薦ができなかった．Web ページを推薦できた 27 人に対しての評価結果を表 4 に示す．また，Web ページを推薦できなかったユーザのタグセット例を表 5 に示す．

表 4 推薦システム評価結果

		Wikipedia カテゴリ情報使用	
		なし	あり
Web ページ数平均	適合	17.56	34.07
	不適合	0.19	1.0
適合率(%)		98.96	96.51

表 5　推薦に失敗したタグセット例

ユーザ 1	ユーザ 2	ユーザ 3
24contest	**	mentalhealth
開発	生活	Politics
あとでもう一度読む	2ch	Joke
プログラミング	ネタ	ux
ruby	開発	Haskell

Wikipedia のカテゴリ情報を用いた場合，用いない場合と比較して平均 34 件の適合している Web ページが推薦できるたことがわかる．また，適合率も 96.5%と良い結果を得ることができた．

しかし，30 人中 3 人のユーザでは Web ページを推薦することができなかった．ユーザ 1 の使用していたタグのデータベース上での使用回数と使用人数を表 6 に示す．

表 6 推薦に失敗したタグの使用回数

タグ名	データベース上の使用回数	使用人数
24contest	10	3
開発	338	188
あとでもう一度読む	2	1
プログラミング	629	313
ruby	337	185

タグセットに含まれるタグを使用している人数があまりにも少なく，結果としてタグセット全てをタグ付けしている類似のユーザが存在しないため，推薦に失敗している．

6.3. 情報抽出システムの評価

6.3.1. 重要文抽出評価

情報抽出システムの評価として，クローラで取得した URL のうちドメインが gigazine のものを 1 件無作為に抽出し対象とした．また，提示される重要文は x=2 とした．評価は主観で行った．正しく重要文が抽出されたかの判断基準は，抽出された文章に推薦に使用されたソーシャルブックマークのタグ情報が含まれるかどうかで判断した．タグ情報が含まれる文章の場合は重要文とし，それ以外は非重要文とする．生成した重要文と Web ページ全体において，抽出に使用したタグ(B タグ)が含まれる文章のうち重要文章数，非重要文章数を比較した．結果を表 7 に示す．また，図 7 に重要文抽出の一例を示す．

表 7　特定のサイトに対する重要文数の比較

ページ全体		抽出した重要文	
重要文章数	非重要文章数	重要文章数	非重要文章数
3	12	2	0

TwitterからのRSS取得が2013年3月5日で打ち切りへ – GIGAZINE

重要文

Twitterで8月にアナウンスされていた**APIリクエスト回数を基本毎時60回以内にする**など利用ルールを改定したTwitter API ver1.1の運用が現地時間8月5日(水)から始まりました。この規約変更を受けて**「Twit」のように開発終了するクライアントも登場**するなどしていますが、今後、大きな影響を及ぼしそうなのは**XML、RSS、Atomがサポート外となる**ことです。

図 7　重要文抽出の一例

ページ全体における B タグが含まれる文章を抽出した場合，非重要文章が多く抽出されている．一方，抽出した重要文では非重要な文章が含まれていない．これは，5.1.3.節で述べたようにページの前半に重要文が存在することが多いためこのような結果になったと考えられる．さらに，ページの後半は他ページへのリンクなど重要文とは関係ないものに B タグが使用されてしまうため，ページ全体では重要文抽出がうまくいか

ないとも考えられる.

また，この Web ページには，タグとして「Twitter」，「これはひどい」，「RSS」，「api」，「Web」が付けられていた．重要文として抽出された文章にもそのような単語が存在した.

6.3.2. コンテンツの主観分析調査

5.2.節で提案したアルゴリズムが 6.3.1.節で行った Web サイトでも適用できるかを調べるためにコンテンツの主観分析を行った。

5.1.2.節で抽出した URL の中から 1 つのドメインにつき 5 件を無作為に抽出した．また，それ以外のドメインの中から 50 件を無作為に抽出し合計 100 件を調査対象 URL とした．主観分析では，主観で重要な単語を見つけた後，使用されている HTML タグを HTML ソースから取得する．重要/非重要な単語において表 1 に含まれる単語を直接強調する HTML タグの使用個数を取得した．また，ページ内前半として，Web ページをスクロールせず閲覧できる範囲内に含まれる，文字を強調するタグについて取得した．なお，閲覧領域は 1280*800dpi である．表 7 に重要/非重要な文字列に使用されていた HTML タグの個数を示す.

表 7 の hatena と gigazine に着目してみると，hatena ではページ前半の strong タグに，gigazine ではページ前半の B タグに着目すれば良いことがわかる．その他のドメインでも，同様にページ前半の強調タグに着目すれば良いことがわかった。

以上のことから他の Web ページにおいても本研究の提案システムが適用可能であると言える。

表 7　主観評価結果(数値はタグの個数)

タグ	hatena				gigazine			
	ページ全体		ページ前半		ページ全体		ページ前半	
	重要	非重要	重要	非重要	重要	非重要	重要	非重要
Strong	30	6	13	0	0	0	0	0
font	17	6	4	0	0	0	0	0
em	0	0	0	0	0	0	0	0
b	0	1	1	0	26	55	17	0
i	0	3	0	0	0	0	0	0
u	0	0	0	0	0	0	0	0

7. まとめと今後の課題

本稿では，ソーシャルブックマークと Wikipedia のカテゴリ情報を用いてユーザの嗜好に合う Web ページの推薦を行う手法を提案し，その有効性を検証した．その後，ソーシャルブックマークでよくブックマークされている Web ページのドメインと使われているタグについて調査をし，重要文抽出アルゴリズムを提案した．また，実験システムを構成し，情報推薦，情報抽出の各手法の有効性を確認した.

情報推薦手法に関しては，タグ付方法が特殊なユーザの場合類似ユーザが存在せず情報推薦を行うことができないことがあった．これに対応するため，特殊なタグの場合にはストップワードに追加して対応する方法を行う予定である

情報抽出手法に関しては，アルゴリズムを提案し，プロトタイプとして実装した．しかしながら，ドメインごとのタグ付傾向によりうまく重要文抽出ができない場合があるため，TF-IDF 法などを組み合わせる必要がある.

今回はシステムの各部の評価を個別にしか行っていないため，今後はシステム全体の評価を行っていく.

参　考　文　献

[1] 吉田大我，中村聡史，田中克己，“ブラウジングと検索の融合：閲覧履歴からの関連情報の検索・自動提示にもとづくウェブ閲覧”，日本データベース学会論文誌，vol.7,No.1,pp.133-138,2008.

[2] 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”，DEWS2008.

[3] 石橋智幸，顧優輝，脇屋達，真部雄介，菅原研次，”ソーシャルブックマークを用いた Web 推薦システムの開発”，電子情報通信学会技術研究報告．AI，人工知能と知識処理，vol. 109(211), pp. 7-12, 2009.

[4] 柴田裕子，山内知子，石川千里，高田雅美，城和貴”複数 Web ページの重要文抽出および直感的理解を支援するための GUI の開発”，情報処理学会研究報告．BIO.バイオ情報学，vol.2007(128),pp.81-84,2007.

[5] Scott A,Golder, Bernardo A.Humberman, ”Usage Patterns Of Collaborative Tagging Systems”, Journal of Information Science, Volume 32 Issue 2,pp.198-208,2006.

[6] Ziyu Guan,Can Wang,Jiajun Bu,Chu Chen, Kun Yang,Deng Cai,Xiaofei He, ” Document Recommendation in Social Tagging Services”,WWW ’10,pp.391-400,2010.

[7] 山家雄介，中村聡史，アダムヤトフト，田中克己，“ソーシャルブックマークの特性分析とそれに基づく Web 検索結果の再ランキング手法”，情報処理学会論文誌データベース，Vol.1,No.1,pp88-100,2008

[8] “Delicious”，http://delicious.com/

[9] “はてなブックマーク”，http://b.hatena.ne.jp/

[10] 中山浩太郎，原隆浩，西尾章治朗，” Wikipedia マイニングによるシソーラス辞書の構築手法”，情報通信学会論文誌， vol. 47, pp. 2917-2918, 2006.