DEIM Forum 2011 F7-4

閲覧履歴からの閲覧モードの推定に基づくブラウジング支援

木村　清尭† 　湯本　高行† 　新居　学† 　高橋　豊† 　角谷　和俊††

† 兵庫県立大学大学院工学研究科　〒 671–2280 兵庫県姫路市書写 2167
†† 兵庫県立大学環境人間学部　〒 670-0092 兵庫県姫路市新在家本町 1-1-12
E-mail: †er09m018@steng.u-hyogo.ac.jp, ††{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp,
†††sumiya@shse.u-hyogo.ac.jp

あらまし　ユーザが Web ブラウジングを行っている際の状態を閲覧モードと定義し，情報を検索するという強い目的に基づいて探索と散策に分類する．閲覧モードの推定は Web ブラウジングの履歴を用いて，閲覧している内容の集中度とサイトの集中度から推定する．散策での支援では，閲覧した Web ページを基準にソーシャルブックマークを用いて Web ページを推薦する．推薦する Web ページは入力の Web ページに付与されたタグ間の上位下位関係を活用して，複数の観点に基づき Web ページを推薦する．推薦の対象とする Web ページは入力ページに対して，類似関係にある Web ページ，汎化関係にある Web ページ，主要な一部分が異なった Web ページであり，これらをソーシャルブックマークから取得する．

キーワード　情報推薦，閲覧モード，ソーシャルブックマーク

Web Browsing Assistance based on Estimation of Browsing Mode from Browsing History

Kiyotaka KIMURA†, Takayuki YUMOTO†, Manabu NII†, Yutaka TAKAHASHI†, and Kazutoshi SUMIYA††

† Graduate School of Engineering, University of Hyogo.2167 Shosha, Himeji, Hyogo,671–2280 Japan
†† School of Human Science and Environment, University of Hyogo.1-1-12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
E-mail: †er09m018@steng.u-hyogo.ac.jp, ††{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp,
†††sumiya@shse.u-hyogo.ac.jp

1. はじめに

近年，Web ブラウジングが一般的になり，日常生活を過ごす上で様々な用途に使用されている．Web ブラウジングを行うユーザは必要に応じて検索エンジンを用いて所望する知識を得たり，ニュースサイトにてニュースを購読したりすることで，興味のある分野の情報をより容易に把握できるようになった．しかし Web 上での情報の爆発的な増加により，ユーザは求める情報の発見に至るまでに多種多様な情報に目を通す必要があり，これはユーザの負担となっている．このような背景から，情報発見までの負担を軽減するための研究や，ブラウジングをより快適に行うための研究などが活発に行われている．これには閲覧している Web ページに対して関連する内容の Web ページを推薦するものや，ブラウジングにより発生する閲覧履歴からユーザの好みを判定し，それに応じた Web ページを推薦するものなどが存在する．

しかし，これらの Web ページの推薦によってユーザを支援する研究では，ユーザの目的意識の強さを考慮しておらず，ユーザの状態を一定としている．このために，これらの手法ではユーザが求めるものを推薦できない場合が考えられる．また推薦方法は推薦される内容に関わらず統一されているため，推薦される内容に適さない場合も考えられる．

実際のブラウジングを想定すると，ユーザが自身の興味を持つ分野の Web ページを閲覧する際，ソーシャルブックマークサービス (以下，SBM サービス) を活用することができる．Web ページの管理形式の 1 つである SBM では，それぞれの Web ページが属している分野名を “タグ” とよばれるテキストで設定することができる．一般的に，タグとして用いられる単

語はブックマークの対象である Web ページに基本的に出現するとは限らない．例えば，C 言語の入門に関する Web ページに “プログラミング”や “コンピュータ”という単語が必ずしも含まれているとは限らない．このために，ユーザが自身の興味を持つ分野の Web ページのために検索エンジンを用いることは適切とは言い難い．また，SBM サービスではサービス利用者の各々のブックマークは Web 上で管理されており，そのデータは公開されている．このような特徴から，ユーザが自身の興味を持つ分野の Web ページを閲覧する際には，SBM サービスに興味のある分野名で問い合わせを行うことで，求める分野の情報を発見することができる．また，検索を行う際，ユーザは求める Web ページの本文中に含まれている単語を推測し，それを検索クエリとして検索を行う．このようなユーザが情報を求める際のアプローチは，状況に応じたものを選択することが適切であり，重要といえる．

そこで本論文では，ブラウジングを行っている際のユーザの状態を閲覧モードと呼び，これは探索と散策から成るとし，それぞれに適した形式支援を行うこととする．本論文ではまず閲覧モードの定義，これの推定を閲覧履歴から行う手法について述べる．そして，興味に基づいたブラウジングを行っているモードの支援として，ユーザが興味を示す Web ページを SBM サービスから取得し推薦する手法について述べる．

以下，2. で関連研究，3. で閲覧モードの定義と推定手法，4. で閲覧モードの 1 つである散策における支援内容とその手法を述べる．そして 5. で実験，6. で結論を述べる．

2. 関 連 研 究

検索行動を分析する研究として，Broder [1] が検索クエリについて調査している．Broder は検索クエリを Navigational(特定の Web サイトを閲覧するためのクエリ)，Informational(求める情報が記載された 1 つのもしくは複数の Web ページを閲覧するためのクエリ)，Transactional(Web を介した行動をするためのクエリ) に分類できると述べている．これらのクエリによって発生するブラウジングを次章に記載している図 1 での分類に当てはめると，Navigational は分類 C に，Informational, Transactional は閲覧するページ数に依存し，それぞれ分類 C, もしくは D に相当する．

また，ブラウジングの 1 つであるユーザの散策的閲覧行動に対する研究では，是津らはユーザの興味を持つページを閲覧する際のブラウジングスタイルが都市での散策行動に類似しているとして，“Web での散策”というコンセプトを提案している [2]．更にこれに基づいて，閲覧中の Web ページの周辺空間を呈示するナビゲーション手法を提案している．この手法では，実際の都市での周辺空間に合わせて呈示することが特徴となっている．本論文で述べる散策に対する支援で推薦する Web ページ集合は，是津らの研究で閲覧ページに対する内容的な周辺空間に属するものと考えられる．

最後に本論文でも研究対象としているブラウジングにおける情報推薦についての研究を紹介する．佐々木ら [3] は，ユーザが用いている SBM のあるタグをベクトル化し，ベクトルを補完するという観点から，他のユーザのブックマークデータから，そのタグに関係を持つ Web ページを推薦する手法を提案している．佐々木らの研究とは異なりタグを階層的には用いていないが，同様に幅広い情報を推薦することが本論文での研究では可能と考えられる．丹羽ら [4] は SBM のデータを用いた Web ページの推薦手法を提案している．この手法では，タグ間の類似度に基づくクラスタの作成することでタグ情報の抽象化を行い，加えてタグとユーザ間の関連度を求めている．これによって各タグクラスタにおけるそれぞれのユーザに適した Web ページを推薦することができる．本論文での散策支援と同様に SBM のタグの階層構造を用いているが，本論文ではクラスタを作成せずにタグの上位下位関係のみを用いている．しかし，これらの手法では入力，もしくは基準とする情報の分野と同じ分野の Web ページが主に取得されると考えられ，本論文で重要としている他分野の推薦対象とする Web ページの取得には適さない．

3. 閲覧モード

閲覧モードとは，ユーザがブラウジングを行っている状態と定義する．本論文ではさらに情報を探すという目的の強さに基づいて探索と散策に分類する．なお，閲覧モードを目的の強さによって詳細に分類することも可能であるが，最も単純な分類として，以下で詳細を述べている探索と散策の分類を本論文では採用する．

3.1 探　　索

探索とは，ユーザが強い目的を持って情報を探すブラウジングをしている状態である．具体例としては以下が挙げられる．

- 専門用語について，意味や用法を調べている
- 以前閲覧した Web ページを，記憶を頼りに探している
- C 言語によるプログラミングの方法について調べている
- 希望する機能を完全に有するソフトウェアを探している

また，1，2 ページの閲覧で達成されるようなブラウジング (テレビの番組表を確認しようと検索を行うブラウジングなど) は探索ではなく，次に述べる散策である．ここでの強い目的はブラウジング開始時にユーザが持ち合わせている目的意識に加えて，情報を発見するまでに閲覧したページ数にも関係している．

このように探索では目的が明確な閲覧を行っており，支援として推薦すべき内容も，より詳細化されている情報が適しているといえる．よって，この場合の支援としては検索クエリに追加するためのキーワードの推薦などが考えられる．これは，ユーザが検索によって情報を求めている状況において，より目的となる情報を容易に発見できるようにするためである．ユーザはクエリに追加するキーワードを選択し，検索クエリとして追加することで，より的を絞った情報を容易に取得できると考える．

このような探索における支援，クエリ拡張については既に数多くの研究が行われている [5] [6] [7]．また，クエリに追加するキーワードの推薦としては，すでに Google がクエリ入力補助

サービス (GoogleSuggest)[注1]を行っている．これらの状況から探索における支援については本論文の対象としていない．

3.2 散　　策

散策とは，ユーザが自身の興味に基づいたブラウジングを行っている状態である．具体例としては以下が挙げられる．

- 日常的に閲覧しているニュースサイトを閲覧している
- タイトル，概要から興味を持ったブログを閲覧している
- 今日の運勢を確認している
- 動画共有サイトにて動画を鑑賞している

このように散策では，ユーザの閲覧する際の目的はあいまいであり，特定の Web ページを求めている状況でないため，ユーザが興味を示すと思われる他分野の Web ページや，興味を示す分野内で汎化された内容の Web ページなどを推薦する支援を行えばよい．ユーザが興味を抱くだろうこれらのような Web ページであれば推薦対象としては問題なく，様々な情報が含まれたものを推薦することが重要である．

探索と散策の観点からの Web ブラウジングの特徴を表 1 に示す．このように，閲覧モードによって傾向が明らかに異なっており，それぞれに応じた支援が必要といえる．

表 1　閲覧モード別のブラウジングの比較

	探索	散策
閲覧の目的	明確	あいまい
支援概要	的を絞った情報の取得	多様な内容の推薦
推薦内容	検索クエリの推薦	Web ページ集合の推薦

3.3 閲覧モードの推定手法

ここで提案する推定手法は，Web ブラウジング時にある Web ページを閲覧しているときの閲覧モードを推定することを目的としている．手法の概要としては，次で述べる探索におけるブラジングの特徴を尺度として用いて，探索であるか (探索でなければ散策として) 判定する．

探索の特徴として以下の 2 つの集中度が挙げられる．

- 連続して同様の内容を閲覧している (内容の集中度)
- 連続で特定のサイトの Web ページを閲覧している (サイトの集中度)

そこでこれらの特徴に基づいて，閲覧モードが探索であるかを判定する．これらからを用いて図 1 のように分類できる．すなわち，内容の集中度，サイトの集中度が共に高い場合が探索である．便宜上，それぞれを分類と呼ぶ．次節以降でこれらの特徴を定式化する．

まず，入力を閲覧履歴のリストとする．入力のリストを，モードの推定対象とする Web ページ p_i と，p_i を含めて直近で閲覧した N 件の Web ページをリスト P とし，以下のように表す．

$$P = \langle p_{i-N+1}, \cdots, p_{i-1}, p_i \rangle \tag{1}$$

ここでは $N = 5$ とした．これを用いて内容の集中度，サイトへの集中度を算出する．

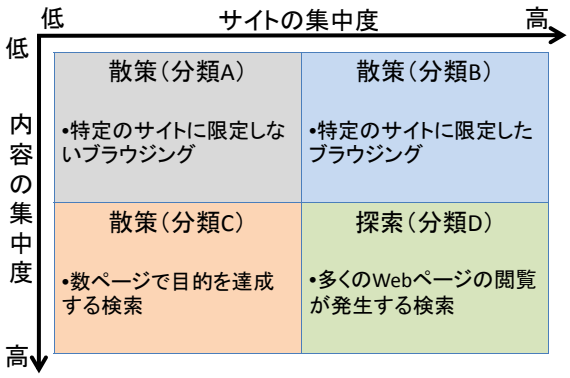

図 1　ブラウジングの分類

3.3.1 内容の集中度による評価値

リスト P において，Web ページ p_i を閲覧した際の文章の類似度による評価値を内容集中度 $\mathrm{ConCon}(p_i, P)$ として以下のように定める．

$$\mathrm{ConCon}(p_i, P) = \underset{p_\alpha, p_\beta \in P}{\mathrm{Average}}(\mathrm{Cos}(p_\alpha, p_\beta)) \tag{2}$$

これはリスト P から順序なしの組合せで選出した Web ページ p_α，p_β のコサイン類似度の平均を表している．探索では同様の内容を連続して閲覧する傾向があるためこのような定式化を行った．なお，コサイン類似度に用いるそれぞれ Web ページの特徴ベクトルは (代名詞を除く) 名詞の TF ベクトルである．

なお，Web ページの文章の本文のみを用いて，広告，ヘッダ，フッダなどは取り除く．これにより同じ Web サイトの Web ページ間の類似度が著しく高くなることを防ぐ．本文取得には ExtractContent[注2]を用いた．これは HTML ソースから ，<td> タグに囲まれたテキストをそれぞれブロックとして取得する．そして各ブロックに，句読点の個数，Web ページでのテキストの位置，ハイパーリンクの有無などからスコアを付与する．高スコアのブロックが連続して存在する場合にはこれらをクラスタとしてまとめ，最終的なスコアが最大のクラスタを本文であると判定する．

3.3.2 サイトの集中度よる評価値

リスト P において，Web ページ p_i を閲覧した際の特定の Web サイトへの集中度よる評価値をサイト集中度 $\mathrm{SiteCon}(p_i, P)$ として以下のように定める．

$$\mathrm{SiteCon}(p_i, P) = 1 - \frac{|UniqueSite(P)|}{|P|} \tag{3}$$

$UniqueSite(P)$ はリスト P 中の Web ページが存在する DNS ホストの集合である．特定のサイトを閲覧している場合，サイト集中度の値は高くなる．

3.3.3 モード推定の最終的な評価式

以上で定めた文章の類似度による内容集中度と，Web サイトへの集中度によるサイト集中度を用いて，リスト P における Web ページ p_i を閲覧した際の探索の度合いを探索度 $\mathrm{SearchDegree}(p_i, P)$ として以下のように定める．

(注1)：http://www.google.com/webhp?complete=1&hl=en

(注2)：http://search.cpan.org/dist/HTML-ExtractContent/

$$\mathrm{SearchDegree}(p_i, P) = \mathrm{ConCon}(p_i, P) \cdot \mathrm{SiteCon}(p_i, P) \tag{4}$$

これにより探索である場合のみ高い値となる．この値がしきい値以上である場合，Web ページ p_i を閲覧している際の閲覧モードは探索と推定する．しきい値は予備実験から 0.08 とした．

4. 散策での支援手法

前述したように散策ではユーザの閲覧する際の目的はあいまいである．このような状況では，ユーザが興味を示すであろう他分野の Web ページや，汎化された内容の Web ページなどを推薦すればよい．これは，ユーザが興味を抱くであろう Web ページであれば推薦対象としては問題なく，様々な情報が含まれたものを推薦することが重要であるためである．推薦対象としては良質な Web ページ群を用いることと，取得する分野を容易に指定できることが重要である．これを解決するために本論文では SBM に登録されている Web ページを用いる．

SBM には頻繁に閲覧する Web ページや，有益な情報が記載された Web ページが登録されるため，Web 中の良質な Web ページが厳選されている．またタグには “programming” や “料理” といった分野を表すものが数多く存在している．分野名は Web ページ中に出現するとは限らず，分野を指定するにあたっては貴重な情報である．このように SBM のデータセットを用いることでユーザが閲覧している内容に対して様々なアプローチで情報を推薦することが可能である．本論文では入力である閲覧した Web ページに対して，以下に示す観点に基づいた Web ページを推薦する．

- **類似ページ**：閲覧している内容と同分野の Web ページ
- **関連ページ**：同じ系統ではあるが，述べられている主要な一部分が別の内容となっている Web ページ
- **汎化ページ**：上位の分野について述べている Web ページ

例として，Perl によるプログラミングについての Web ページを閲覧している場合を挙げる．この例では汎化ページとしてはプログラミングそのものについての Web ページ，類似ページとしては同様に Perl でのプログラミングにおける Web ページ，関連ページとしては Perl ではなく Ruby でのプログラミングについての Web ページを SBM から取得し推薦する．

これらの Web ページを取得するために本論文では SBM のタグを活用する．入力とする Web ページに付与されたタグ集合に対し，各タグの上下関係の分析や一部のタグの置換などを行い，SBM に問い合わせを行うことで，上で述べた各観点に基づく Web ページを取得する．

4.1 Web ページからのタグペアの生成手法

入力とする Web ページは推薦する各 Web ページの基点となるものである．そこで入力ページを基点ページとし，基点ページ p_b のブックマークに付与されているタグの集合を $Tag(p_b)$ として以下のように表す．

$$Tag(p_b) = \{t_1, t_2, \cdots, t_n\} \tag{5}$$

なお基点ページは SBM に登録されていることを前提とする．このタグ集合から上位下位の関係にある 2 つのタグをタグペアとして作成する．タグペアは 2 つのタグから成り，上位階層にあたるタグ t_u，下位にあたるタグ t_l として，これらを一対としたものである．これを (t_l, t_u) と表記する．基点ページ p_b から取得したタグペアの集合を $TagPair(p_b)$ として以下に示す．なお，$t_l \prec t_u$ は上位タグ t_u と下位タグ t_l に上位下位関係があることを示す．

$$TagPair(p_b) = \{(t_l, t_u) | t_l, t_u \in Tag(p_b), t_l \prec t_u\} \tag{6}$$

後の処理では基点ページから生成した複数のタグペアからタグのツリーを生成し，これを用いて各 Web ページを取得する．

タグペアの 2 つのタグが満たす条件として以下を定める．

- 基点ページのブックマークにおいて，多用されている
- 同じ分野のタグである (例. (料理という分野で) “カレー” と “ラーメン”)
- SBM 内で概念的な上位下位の関係が見受けられる (例. “料理” と “カレー”)

それぞれの条件に付いて具体的に述べていく．まず，多用されている条件として以下の条件を定める．

$$\frac{\mathrm{DF}(t, BM_{Page}(p_b))}{|BM_{Page}(p_b)|} \geqq \theta_1 \tag{7}$$

上記の条件を満たすタグ t を用いてタグペアを構成するものとする．$BM_{Page}(p_b)$ は基点ページ p_b に対するブックマーク集合，$\mathrm{DF}(t, BM_{Page}(p_b))$ は $BM_{Page}(p_b)$ でのタグ t の使用数である．この条件によりごく一部のユーザが付与したタグをタグペアに用いることを防ぐ．なお，θ_1 は経験的に 0.075 とした．

そして同じ分野のタグは，SBM で高頻度で共起していると考え，判定に Jaccard 係数を用いて同分野の判定の条件を設けた．Jaccard 係数の算出は以下のように行った．

$$\mathrm{Jaccard}(t_l, t_u) = \frac{|BM_{Tag}(t_l) \cap BM_{Tag}(t_u)|}{|BM_{Tag}(t_l) \cup BM_{Tag}(t_u)|} \tag{8}$$

$BM_{Tag}(t)$ はタグ t が付与されているブックマークの集合である．この値がしきい値以上のとき同じ分野のタグであると判定する．ここではしきい値を経験的に 0.01 とした．

最後に，上下関係の判定として，各タグの共起タグの種類数を用いて以下のように条件を定めた．

$$|CoTag(t_u)| > |CoTag(t_l)| \tag{9}$$

$CoTag(t)$ はタグ t が付与されているブックマークで共に付与されているタグの集合である．この条件は上位のタグは下位のタグより多用され，様々な下位の分野のタグと共起しているという仮定に基づいている．

4.2 タグペアの統合によるタグツリーの生成手法

基点ページから生成されたタグペアには，下位タグとなっているタグが別のタグペアでは上位タグとして用いられることがある．そこでタグの多階層関係を表現すべく，複数のタグペアを統合したものを木として扱い，これをタグツリーとする．図 2 にタグペア集合からのタグツリーの生成例を示す．図 2 はタグツリーを 2 つ生成している．このように生成したタグツリーを活用して SBM から Web ページを取得する．

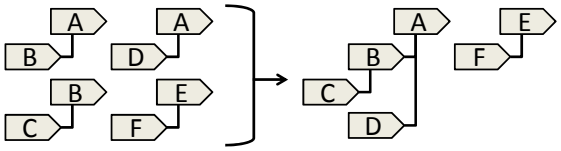

図 2　タグツリーの生成例

4.3　類似ページの取得手法

提案手法では類似ページに限らず，関連，汎化ページとして推薦する Web ページは，以下の式から得る Web ページ集合である．

$$MatchPage(p_b, f) = \bigcup_{tt \in TagTree(p_b)} \left(\bigcup_{t_{lf} \in Leaf(tt)} f(t_{lf}, tt) \right) \tag{10}$$

$TagTree(p_b)$ は基点ページ p_b から生成されるタグツリー集合，$Leaf(tt)$ はタグツリー tt における最下位タグの集合である．そして $f(t_{lf}, tt)$ は最下位タグ t_{lf} とタグツリー tt を用いて Web ページ集合を返す関数である．この節で示す類似ページや，次節以降で示す関連，汎化ページを取得するには式 10 で用いる関数 $f(t_{lf}, tt)$ をそれぞれに適した関数とすることで，目的の Web ページを得ることができる．

まず，類似ページの取得手法について述べる．類似ページは，最下位タグから根ノードにあたるタグ (最上位タグ) に至るまでに経由する各タグが付与されている Web ページ集合とする．これを得る関数を以下のように表す．

$$f_s(t_{lf}, tt) = \bigcap_{t \in ancestor^*(t_{lf}, tt)} Page(t) \tag{11}$$

$ancestor^*(t_{lf}, tt)$ は，タグツリー tt において最下位タグ t_{lf} から最上位タグに至る経路上のタグの集合である．これは最下位タグ t_{lf} 自身を含む．また，$Page(t)$ はタグ t が付与されている Web ページ集合である．よって，基点ページと類似関係にある Web ページ集合 $SimPage(p_b)$ を以下のように取得する．

$$SimPage(p_b) = MatchPage(p_b, f_s) \tag{12}$$

例として，図 2 でのタグツリー集合を用いた場合の類似ページ集合は，

$$\begin{aligned} &\big(Page(A) \cap Page(B) \cap Page(C)\big) \\ &\cup \big(Page(A) \cap Page(D)\big) \cup \big(Page(E) \cap Page(F)\big) \end{aligned} \tag{13}$$

となる．この例での処理の流れを図 3 に示す．

4.4　関連ページの取得手法

関連ページを取得では，入力とするタグツリー集合を類似ページの取得手法と同様に分割する．そして，それぞれの最下位タグを別のタグに置換したものを生成する．これを類似ページ取得と同様の手法を用いることで関連ページを取得する．

タグツリー tt とそれに含まれる最下位タグ t_{lf} から得られる関連ページ集合を得る関数を f_r として以下に示す．

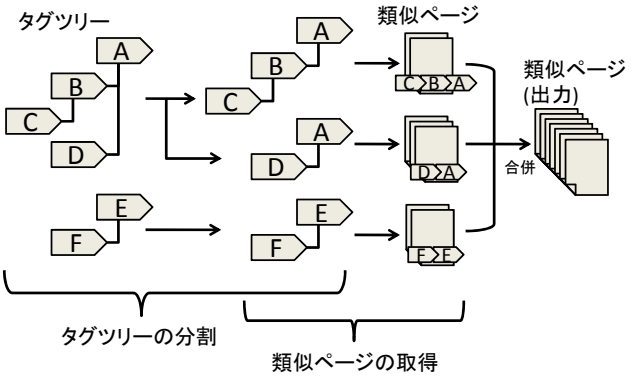

図 3　類似ページの取得イメージ

$$\begin{aligned} &f_r(t_{lf}, tt) \\ &= \bigcup_{t_s \in SibTag(t_{lf}, tt)} \left(\bigcap_{t \in ancestor(t_{lf}, tt)} Page(t) \cap Page(t_s) \right) \end{aligned} \tag{14}$$

なお，$ancestor(t_{lf}, tt)$ はタグツリー tt において最下位タグ t_{lf} から最上位タグに至る経路上のタグの集合である．ただし最下位タグ t_{lf} 自身は含まないものとする．そして，前述した最下位タグと置き換えるタグをここでは兄弟タグとし，これの集合を $SibTag(t_{lf}, tt)$ としている．

タグツリー tt の最下位タグ t_{lf} における兄弟タグ集合 $SibTag(t_{lf}, tt)$ は以下のように取得する．

$$SibTag(t_{lf}, tt) = \underset{t \in CoTag'(t_{lf})}{\text{Arg Top}} (\text{Conf}(t \to t_u), num) \tag{15}$$

$$\begin{aligned} &CoTag'(t_{lf}) \\ &= \{t | t \in CoTag(t_{lf}), \text{Supp}(t \to t_u) \geqq \theta_2, |User(t)| \geqq \theta_3\} \end{aligned} \tag{16}$$

$\underset{x \in X}{\text{Arg Top}}(f(x), num)$ は関数 f を適用する集合 X の要素 x のうち，f の値が上位 num 件となる x の集合である．$num = 3$ とした．また，タグ t_u はタグツリー tt におけるタグ t_{lf} の上位タグである．タグ t_α とタグ t_β の相関ルール $t_\alpha \to t_\beta$ の確信度と支持度は式 17，18 にあるように算出する．

$$\text{Conf}(t_\alpha \to t_\beta) = \frac{|BM_{Tag}(t_\alpha) \cap BM_{Tag}(t_\beta)|}{|BM_{Tag}(t_\alpha)|} \tag{17}$$

$$\text{Supp}(t_\alpha \to t_\beta) = \frac{|BM_{Tag}(t_\alpha) \cap BM_{Tag}(t_\beta)|}{N'} \tag{18}$$

N' は SBM サービスで管理されている全ブックマークの個数である．最後に $User(t)$ はタグ t を SBM で使用しているユーザ集合である．各しきい値を経験的に $\theta_2 = 0.00001$，$\theta_3 = 50$ とした．

以上から，関数 f_r と前節で定めた式 10 を用いて，基点ページ p_b の関連ページを以下のように求める．

$$RelPage(p_b) = MatchPage(p_b, f_r) \tag{19}$$

この例での処理の流れを図 4 に示す．図中では分割したタグツリーの一部を用いて処理の流れを表している．

4.5　汎化ページの取得手法

汎化ページを取得では，まずタグツリーをこれまでと同様に分割し，分割したタグツリーの最下位タグ以外のタグを用いて

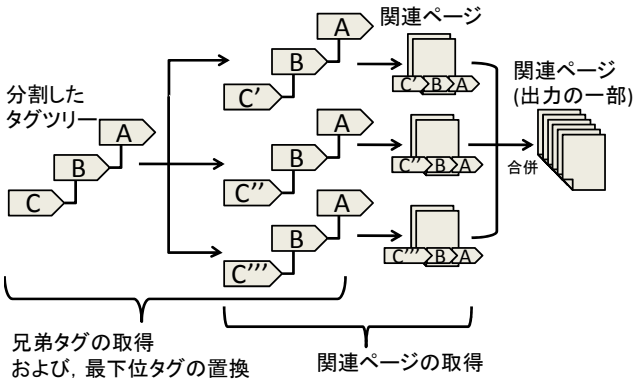

図 4　関連ページの取得イメージ

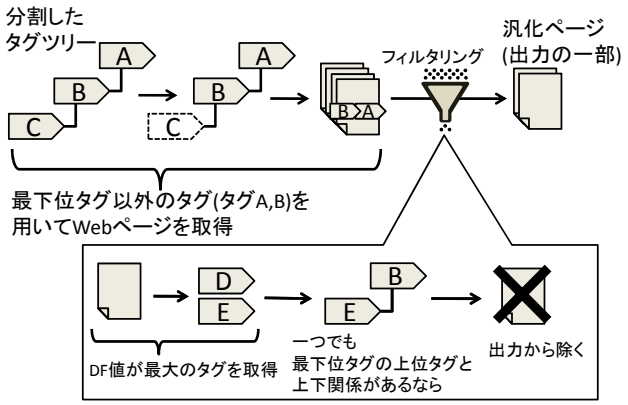

図 5　汎化ページの取得イメージ

Web ページを取得する．そして，取得した Web ページのうち，それぞれの Web ページのブックマーク中で出現頻度が最大のタグが，分割したタグツリーの最下位タグの代わりに最下位タグとなるものは取り除く．この条件を満たした Web ページ集合を汎化ページとする．

基点ページから生成したタグツリー tt と，これを成している最下位タグ t_{lf} を用いて，汎化ページとなる Web ページ集合を得る関数を f_g として以下に示す．

$$f_g(t_{lf}, tt) = \bigcap_{t \in ancestor(t_{lf}, tt)} Page(t) \tag{20}$$

ただし，この関数から得られる Web ページのうちは式 21 から式 23 を満たす Web ページ p に限定して推薦する．関数 f_g で得られる Web ページのうち，この条件も満たす Web ページを得る関数を f'_g とする．

$$t_f \not\prec t_u \tag{21}$$

$$t_f \in \{t | \operatorname*{Arg\,Max}_{t \in Tag(p)} \mathrm{DF}(t, BM_{Page}(p)), t \neq t_u\} \tag{22}$$

$$t_u \in \bigcup_{tt \in TagTree(p_b)} \left(\bigcup_{t_{lf} \in Leaf(tt)} \left(Upper(t_{lf}, tt) \right) \right) \tag{23}$$

t_f は Web ページ p のブックマークに用いられているタグのうち，DF 値が最大のタグ，つまり内容を最も表しているタグをである．t_u はタグツリー tt に含まれる最下位タグの上位タグを表している．$Upper(t_{lf}, tt)$ はタグツリー tt の最下位タグ t_{lf} の上位タグ集合である．そして，$t_f \not\prec t_u$ はタグ t_f が上位タグ t_u の下位タグとならないことを表している．これにより，下位タグに重点を置かずに上位タグに重点を置いた Web ページを得ることができる．

以上から，関数 f'_g と前節で定めた式 10 を用いて，基点ページ p_b の汎化ページを以下のように求める．

$$GenPage(p_b) = MatchPage(p_b, f'_g) \tag{24}$$

この例での処理の流れを図 5 に示す．図中では分割したタグツリーの一部を用いて処理の流れを表している．

5. 評価実験

5.1 閲覧モード推定実験

被験者は，Web ブラウジングに Firefox を標準利用している同研究室の学生 5 名である．被験者には実験実施日から遡り 3 日分の履歴に，閲覧モードの定義を説明したうえで履歴のラベリングを指示した．たたし，対象とする履歴は 10 ページ以上の Web ページを閲覧していることを条件としている．また，ログインが必要な Web ページや，各ユーザによってカスタマイズが可能な Web ページは推定の対象外としている．

そして，評価基準を “ラベリング結果と推定結果がどれほど一致しているか” とし，一致率を以下のように定めた．

$$\text{一致率} = \frac{\text{被験者のラベリング結果と推定結果の一致数}}{\text{対象の全閲覧ページ数}} \tag{25}$$

実験結果を表 2 に示す．表中の平均以外の一致率は，それぞれ一日分の履歴を評価して得た一致率である．

表 2　閲覧モード推定結果 (一致率)

被験者	一致率	被験者	一致率	被験者	一致率
	0.22		0.61		0.93
A	0.79	B	0.35	C	0.88
	0.56		0.69		0.74
	0.43		0.58		
D	0.64	E	0.42	平均	0.60
	0.58		0.61		

提供された閲覧履歴と推定結果を調査したところ，一致率が低下する要因としては実験結果から以下の要因が考えられた．

- 探索に属するブラウジングの開始に対して，探索度の増加が数ページ後に生じた誤判定 (探索を散策と誤判定)
- 散策に属するブラウジングにおいて連続して同じサイトを閲覧した際に生じた誤判定 (散策を探索と誤判定)

推定の遅延については，現時点では減少させることは困難である．そこで提供された閲覧履歴を調査したところ，被験者によるが Web ブラウジングの多く (およそ 6 割から 9 割) は散策に属するブラウジングであった．この傾向から，モード推定は探索を如何に検出するかが重要だと考えられ，特定の Web ページ (検索サイトなど) に対して重みを与えることで，遅延の抑制が期待できると考える．

散策での同サイトの連続閲覧時には，内容集中度が高くなる

ことで誤判定が発生していた．しかし，探索度のしきい値を上昇させることは探索の検出を見逃す恐れがある．よって，独自の本文抽出に用いた ExtractContent ではなく，独自の本文抽出を行うことで幾らかの改善が見込めると考えている．

また，画像を求めるブラウジングが複数の履歴に含まれていたが，これらの Web ページは文章量が少ないために内容集中度が上昇せずに散策と誤判定をしていた．これに対してはタイトルなどのテキストを用いるなどの本文への依存を減らす必要性が考えられる．

最後に，文章量の少ない Web ページでは本文抽出を適用すると，ごく少量の文章を用いた文章の類似度判定を行うことになる．このため，一定量に満たない文章が記述された Web ページに対しては本文抽出を適用しないなどの対策が必要である．

5.2 散策支援の評価実験

基点ページから提案した手法によって類似ページなどの Web ページを正しく取得できたかを評価の基準とした実験を行った．5.1 節と同じ被験者 (5 名) に表 3 に示す基点ページの一覧と，各基点ページから提案手法によって取得した Web ページを提示し，類似などの定義を説明したうえで分類を指示した．実験には livedoor クリップの 2009 年 12 月時点の SBM のデータを用いて，タグの大文字小文字は区別せずに扱った(注3)．提示する推薦ページは，提案手法によって取得したものから抜粋したものであり，それぞれがどの Web ページかを伏せて提示している．また，類似ページなどのいずれにも属さないページ (他ページと表記) を含めた．他ページは基点ページのタグツリーに含まれているタグを用いて SBM から取得した Web ページであり，広義で同じ分野の Web ページである．

実験結果の評価として，分類正答率を以下のように定義し，結果の分析を行った．

$$\text{分類正答率} = \frac{\text{正しく分類できたページ数}}{\text{提示した全ページ数}} \tag{26}$$

被験者 A から E までの分類正答率を表 4 に示す．

表 4　散策における推薦ページの分類結果 (分類正答率)

基点ページ	A	B	C	D	E	平均
No.1	0.18	0.27	0.36	0.27	0.64	0.34
No.2	0.09	0.36	0.36	0.09	0.36	0.25
No.3	0.13	0.38	0.50	0.13	0.38	0.30
No.4	0.38	0.50	0.38	0.25	0.25	0.35
No.5	0.30	0.10	0.40	0.10	0.10	0.20
No.6	0.33	0.33	0.11	0.00	0.33	0.22
No.7	0.10	0.10	0.20	0.20	0.20	0.16
No.8	0.20	0.40	0.20	0.40	0.50	0.34
No.9	0.00	0.11	0.22	0.00	0.11	0.09
No.10	0.00	0.38	0.38	0.00	0.38	0.23

表 4 に示しているように，全体的に各 Web ページの正しい取得が行えていないと判断できる．この原因を調査するために，各ページをどのように誤判定しているか (類似ページを関連ページや汎化ページと誤判定しているか，など) を調査した．そこで，被験者による分類結果の偽陽性率を算出し，評価を行った．偽陽性率は，陰性であるものを陽性と誤判定した割合である．本手法では，類似，関連，汎化，その他，の 4 値であるため，それぞれの偽陽性率を以下のように定めた．

$$\text{偽陽性率 (類似)} = \frac{\text{類似ページと誤判定したページ数}}{\text{類似ページ以外のページ数}} \tag{27}$$

$$\text{偽陽性率 (関連)} = \frac{\text{関連ページと誤判定したページ数}}{\text{関連ページ以外のページ数}} \tag{28}$$

$$\text{偽陽性率 (汎化)} = \frac{\text{汎化ページと誤判定したページ数}}{\text{汎化ページ以外のページ数}} \tag{29}$$

$$\text{偽陽性率 (他)} = \frac{\text{他ページと誤判定したページ数}}{\text{他ページ以外のページ数}} \tag{30}$$

それぞれ偽陽性率を表 5 に示す．表 5 から，関連ページの偽陽性率が他の項目と比較して高い値となっている．これは取得した Web ページがそれぞれの取得を目指した Web ページではないが，基点ページに対して一定の関連性を持っていると被験者が判断したためだと考えられる．

表 5　散策における推薦ページの偽陽性率 (平均)

基点ページ	類似	関連	汎化	他
No.1	0.10	0.45	0.28	0.07
No.2	0.04	0.49	0.11	0.36
No.3	0.33	0.20	0.16	0.31
No.4	0.04	0.49	0.23	0.03
No.5	0.29	0.45	0.20	0.13
No.6	0.02	0.57	0.27	0.25
No.7	0.00	0.22	0.04	0.78
No.8	0.00	0.42	0.20	0.20
No.9	0.09	0.54	0.20	0.37
No.10	0.20	0.50	0.20	0.14
平均	0.11	0.43	0.19	0.26

また，他ページをどれだけ取り除くことができたかを調査した．評価指標として推薦正答率を以下のように定めた．

$$\text{推薦正答率} = \frac{\text{他ページ以外を他ページ以外と判定，または 他ページを他ページと判定したページ数}}{\text{提示したページ数}} \tag{31}$$

表 6 にその結果を示す．表 6 にあるように，各基点ページで平

表 6　推薦正答率

基点ページ	A	B	C	D	E	平均
No.1	0.82	0.73	0.82	0.82	0.91	0.82
No.2	0.36	0.82	0.73	0.27	1.00	0.64
No.3	0.50	0.50	0.88	0.50	0.88	0.65
No.4	0.75	0.75	0.75	0.75	0.75	0.75
No.5	0.70	0.70	0.80	0.60	0.70	0.70
No.6	0.78	0.67	0.67	0.44	0.78	0.67
No.7	0.20	0.70	0.30	0.20	0.30	0.34
No.8	0.80	0.50	0.50	0.60	0.80	0.64
No.9	0.33	0.33	0.78	0.33	0.67	0.49
No.10	0.50	0.88	0.75	0.75	0.88	0.75
平均	0.57	0.66	0.70	0.53	0.77	0.64

(注3)：http://clip.livedoor.com/

表 3　散策支援に用いた基点ページ一覧

タイトル (通し番号を付与) および URL
No.1 入門から実践まで Java で学べる「ログ」の常識　(1/4) - @ IT http://www.atmarkit.co.jp/fjava/rensai4/programer10/programer10_1.html
No.2 珈琲の淹れ方・器具 総合スレ:アルファルファモザイクだった http://alfalfa.livedoor.biz/archives/51322316.html
No.3 ぬこ可愛すぎハムスター速報　2ろぐ http://2log.blog9.fc2.com/blog-entry-1696.html
No.4 ASCII.jp：jQuery で作る“ Amazon 風 ”インターフェイス ｜29 分でできる！ あのサイトの“ 技 ”を盗め http://ascii.jp/elem/000/000/173/173575/
No.5 無意識の力で早起きをするテクニック — Lifehacking.jp http://lifehacking.jp/2007/07/using-the-subconscious-to-wake-up-early/
No.6 脳腐ってるんじゃないかと思わせるアメリカのイカレた州法 - 適宜覚書はてな異本 http://d.hatena.ne.jp/dacs/20080705/1215230996
No.7 元日限定！１３０円で関東一周…一筆書きの旅大公開 社会：ZAKZAK http://www.zakzak.co.jp/top/200812/t2008122709_all.html
No.8 [映画] 時をかける少女（細田守監督）：極東ブログ http://finalvent.cocolog-nifty.com/fareastblog/2008/07/post_6c62.html
No.9 携帯電話の「簡単ログイン」は個体識別番号を使ってこんなふうに作れます ｜WEB クリエイターの木 http://ameblo.jp/yosswi/entry-10036647527.html
No.10 ソフトバンク孫氏、「ケータイで楽しめる世界を作る」 ケータイ Watch http://k-tai.impress.co.jp/cda/article/news_toppage/43817.html

均して 6 割を超える他ページの検出率を得た．これから，同じ分野の Web ページでも他ページのような関連性が弱い Web ページを提案手法では取り除くことが可能であるといえる．

以上から，提案手法では類似ページなどの提案手法で定めた Web ページを正しく取得しているとは言い難い結果だが，入力である基点ページに関連のある Web ページを取得できているといえる．手法の改良としては，タグの表記ゆれの解消によるタグツリーの生成手法の改善が考えられる．

6. おわりに

本論文では，Web ブラウジング時に適切な支援を行うために，ユーザのブラウジングにおける状態の推定手法と，状態に応じた適切な支援手法の提案を行った．

まずユーザのブラウジング時の状態を閲覧モードとし，これを目的意識の強さによって探索と散策に分類した．閲覧モードの推定には，内容の集中度とサイトの集中度の 2 点が満たされていれば探索，そうでなければ散策とした．推定実験の結果，約 6 割の推定の一致を確認できた．誤判定例として，画像の取得を目的としたブラウジングでの探索の検出漏れや，散策時の特定サイトの連続閲覧を探索と判定するケースが目立った．対策として，探索時に用いられやすい Web ページへの重みの付加や，独自の本文抽出手法の適用が考えられる．

各閲覧モードでの支援として，探索では検索クエリの推薦が，散策時には多様な内容の Web ページの推薦が適切な支援であると述べた．散策における Web ページ推薦では基点とする Web ページに対して取得する Web ページを，類似，関連，汎化ページと定める，これらを得るために SBM から上位下位関係に基づくタグツリーを生成し，ツリーを活用して各 Web ページを取得する手法を提案した．実験として，取得した Web ページを類似ページなどに分類する実験を行い，基点ページと一定の関連を持つ Web ページを取得できることが確認できた．しかし，本手法で取得を目指した Web ページを目論見通り取得できているとは言い難い結果となった．この対策としては，タグの表記ゆれの解消によるタグツリーの生成手法の改善が考えられる．

謝　辞

本論文の一部は，平成 22 年度科研費基盤研究 (B)(2)「ユーザの潜在的意図を用いたレス・コンシャス情報検索基盤の構築」(課題番号：20300039) によるものです．ここに記して謝意を表すものとします．

文　献

[1] A. Broder: “A taxonomy of web search”, SIGIR FORUM, **36**, 2, pp. 3–10 (2002).

[2] 是津 耕司, 田中 浩也, 池田 新平, 金 星ヨン, 田中克己：“Web 上での散策行動を支援する周辺情報提示機構”, 情報処理学会研究報告. データベース・システム研究会報告, 71, pp. 343–349 (2003).

[3] 佐々木 祥, 宮田 高道, 稲積 泰宏, 小林 亜樹, 酒井善則：“Folksonomy におけるコンテンツ推薦のためのメタデータ成長モデルの提案 (情報抽出)”, 電子情報通信学会技術研究報告. DE, データ工学, **106**, 150, pp. 67–72 (2006).

[4] 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 (エージェント応用システム, ＜特集＞マルチエージェントの理論と応用)”, 情報処理学会論文誌, **47**, 5, pp. 1382–1392 (2006).

[5] B. Billerbeck and J. Zobel: “Techniques for efficient query expansion”, Proc. String Processing and Information Retrieval Symp, Springer-Verlag, pp. 30–42 (2004).

[6] J. Xu and W. B. Croft: “Query expansion using local and global document analysis”, In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 4–11 (1996).

[7] A. M. A. Singhal: “Improving automatic query expansion”, pp. 206–214 (1998).