DEIM Forum 2010 D2-5

ソーシャルブックマークにおけるトピック分析と活性度推定に基づくWebページのランキング

高橋　翼† 渡邊　桂太† 北川　博之†,††

† 筑波大学 大学院システム情報工学研究科　〒 305–8573 茨城県つくば市天王台 1–1–1
†† 筑波大学 計算科学研究センター　〒 305–8573 茨城県つくば市天王台 1–1–1
E-mail: †{tsubasa,ein_vogel}@kde.cs.tsukuba.ac.jp, ††kitagawa@cs.tsukuba.ac.jp

あらまし　ブックマーク情報をタグ付けし，管理・分類・共有するソーシャルブックマークが注目されている．被ブックマーク数はWebページの質を測る一つの指標と言えるが，その時間変化は加味されていない．本研究では，クラスタリングや隠れマルコフモデルによる時系列分析を用いて，扱うトピックの違いや，潜在的な注目度・持続度の違いを考慮し，Webページの活性度を推定する．また，活性度を基にしたWebページのランキング手法の提案も行う．
キーワード　ソーシャルブックマーク，活性度分析，トピック分析，ランキング

A Ranking Method for Web Search based on Topic-aware Activation Analysis Using Social Bookmarks

Tsubasa TAKAHASHI†, Keita WATANABE†, and Hiroyuki KITAGAWA†,††

† Graduate School of Systems and Information Engineering, University of Tsukuba　Tennodai 1–1–1, Tsukuba, Ibaraki, 305–8573 Japan
†† Center for Computational Sciences, University of Tsukuba　Tennodai 1–1–1, Tsukuba, Ibaraki, 305–8573 Japan
E-mail: †{tsubasa,ein_vogel}@kde.cs.tsukuba.ac.jp, ††kitagawa@cs.tsukuba.ac.jp

Abstract　Social bookmarking services have recently made it possible for us to register and share our own bookmarks on the web and are attracting attention. The number of bookmarks is a barometer of web page values, but there is no consideration for the changes of its attractiveness over time. If most of the bookmarks are very old, the page may be obsolete. And the indicator of freshness and obsolescence depends on the content characteristics and the topics of web page. In this paper, focusing on the timestamp sequence of social bookmarkings on web pages, we model their activation levels representing current values. In the activation analysis, we also consider the potential attractiveness of topics. Further, we improve our previous ranking method for web search by introducing the proposed activation level concept. Finally, through experiments, we show effectiveness of the proposed ranking method.
Key words　Social Bookmark, Activation Analysis, Topic Analysis, Ranking Algorithm

1. まえがき

Web検索エンジンの普及によって，情報の宝庫であるWebから様々な知識を簡単に獲得できるようになった．Web検索エンジンでは，Webページ (以降，ページ) 間のリンク構造や，ページコンテンツ内の単語の出現頻度，閲覧頻度の高さと言った情報に基づき，ページの評価が行われている．しかし，そのような評価は必ずしもページを閲覧した人々の感想や意見，嗜好を反映したものではない．

一方，Web上に個人のブックマーク情報を作成し，管理，分類，共有するサービスであるソーシャルブックマーク (SBM) に注目が集まっている．SBMでは，ユーザはブックマーク情報に独自の観点でタグを用いて注釈付けすることができる．ブックマークされたページは役に立つ，おもしろいなど，ブックマークしたソーシャルブックマークユーザ (以降，ユーザ) にとって，何らかの価値のある情報を持ったページと考えることができる．また，ユーザがブックマークするという振舞は，ページに対して評価を与える行為と考えることもできる．我々は，ペー

ジとユーザ間のブックマークの関係から，良質なページをブックマークしているユーザは良質なハブであり，良質なユーザがブックマークしているページは良質な情報であるという考えに基づくランキング手法 S-BITS [1] を提案し，有効性を示した．

SBM のブックマーク情報には，ユーザがブックマークした時刻のタイムスタンプがメタデータとして与えられている．SBM 上には，ニュース記事のように人々に注目されやすいページが非常に多く登録されている．それらのコンテンツは，時間経過と共に価値が衰退すると考えることができ，時間変化を加味しない静的なブックマークの関係だけでは，現在のコンテンツの価値を的確に捉えているとは言い難い．先行研究 [2] では，現在におけるページの活性度を，そのページのブックマークの時系列パターンと平均ブックマーク頻度から推定する手法を提案した．しかし，新規にブックマークされたページや短期間のブックマーク系列しか持たないページに対しては，基準とする情報が少なく，ロバストな評価を行うことができなかった．また，活性度の評価は対象ページの過去の系列との比較であり，他のページと比較可能な指標ではなかった．

一方，コンテンツのトピックが類似するページは，ブックマークするユーザが共起したり，与えられるタグが類似すると言ったことが SBM では起こる．そのため，トピックの類似するページは潜在的な注目度が類似する可能性が高い．先行研究で提案したページベースの活性度推定がロバストでないと言う問題を，基準となるブックマーク頻度を類似ページ間の時系列から導出すると言った，よりマクロな視点を用いることで解決を図る．類似ページの集約は，ページのブックマーク情報に付加されているタグの類似性を基に，ページのクラスタリングを行うことで実現する．トピックごとの基準を設けることで，トピック内で活性度の差異を比較することが可能となり，活性度を用いたランキングでは，検索対象トピック内で活性度が高いページをより上位にリフトアップさせる効果を得られる．このランキング手法を S-BITS*TB と呼び，被験者実験を通して，提案ランキング手法の有効性を示す．

本稿の以降のセクションの構成は以下の通りである．まず，2 章では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，本研究と関連のある過去の研究について概観する．4 章では，SBM におけるページのトピッククラスタリングについて述べる．5 章では，トピックに基づいたページの活性度とその評価手法について述べる．6 章では，ページのランキング手法について述べる．7 章では，提案手法の有効性を測るための評価実験について述べる．最後に，8 章で本稿のまとめを述べると共に，今後の課題について述べる．

2. ソーシャルブックマーク

ソーシャルブックマーク (SBM) は近年注目を集めている Web2.0 の概念を持つサービスの一つである．主要な SBM サービスに，Delicious [3] やはてなブックマーク [4] などがある．多くの SBM サービスでは，ユーザは独自の価値観で選択した任意のキーワードをタグとして利用し，ページに注釈を付けることができる．また，ブックマーク時刻はユーザがブックマーク

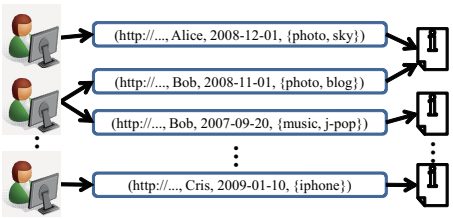

図 1　ソーシャルブックマーク

対象のページにいつ興味を持ったかを表す情報である．このような SBM サービスからはブックマーク情報を (URL，ユーザ名，ブックマーク時刻，タグ集合) というような構造化された形で取得できる (図 1)．これはユーザがページに対して，いつ興味を持ち，どのような情報として認識しているのかを表す一種の嗜好情報と考えることができる．またタグは，ユーザがページコンテンツを閲覧した結果を踏まえて，ブックマーク情報を効果的に管理・分類するために与えられる．ゆえにタグ集合は，コンテンツの内容を高度に要約したキーワード群であり，トピックを表す情報源と言える．

本研究では，ユーザ u_j のページ p_i に対するブックマーク情報 $b_{i,j}$ を以下のようにモデル化する．

$$b_{i,j} = (p_i, u_j, t_{i,j}, A_{i,j}) \tag{1}$$

$$A_{i,j} = \{a_0, a_1, ..., a_{n-1}\} \tag{2}$$

t はブックマーク時刻，$A_{i,j}$ はユーザ u_j がページ p_i に注釈として与えたタグの集合を表し，a_k は各タグを表す．

自分の興味があるトピックについて情報を得たいと思ったとき，そのトピックについて詳しい人からの意見は参考になる可能性が高い．また，多くの人々から評価を得ている情報もまた参考になる可能性が高い．何人のユーザがブックマークをしているかという情報は，ページの信頼度や品質を測る 1 つの指標と言える．Yanbe ら [5] は SBM 上でのページの被ブックマーク数を SBRank という Web 検索の際の尺度として用い，PageRank [6] との比較実験を行っている．加えて，様々なユーザによるページへのブックマークの時系列情報は，ページに対するユーザ群の興味の変化，注目度の変遷と捉えることができる．あるユーザがどんなタグを使い，どんなページをブックマークしているかは，そのユーザの嗜好を表す情報であり，利用するタグや，ブックマークページが類似するユーザ同士は，嗜好や興味も類似している可能性が高い．ページ間においても同様に，与えられているタグの集合が類似するページ同士は，扱うトピックが類似している可能性が高いと言える．Li ら [7] は，タグがページ中に出現する単語よりも高度で精度の高い要約を実現し，頻出なタグの共起集合が，ページのトピックを代表する情報源として利用可能であることを示した．

3. 関 連 研 究

SBM の普及と共に SBM を含む Folksonomy に関する研究は盛んになってきている．Xian Wu ら [8] は，アノーテーションが与えられた Web リソースに対するセマンティックな検索モ

デルを SBM を取り上げ，提案している．Hotho ら [9] は SBM のページのランク付けに PageRank を応用することで，トピックに特有のトレンドを発見する手法を提案している．Heymann ら [10] は，様々な観点から SBM を分析し，その詳細な報告を行っている．毛受 [11] らは SBM に新しく投稿されたページの注目度を予測する手法を提案している．Capocci ら [12] は，特定タグの利用間隔に着目し，タグの利用形態の共起や相関のパターンの発見を行った．本研究は，ページ固有のブックマークの生起間隔に着目し，ブックマーク頻度を基に活性度を推定している点で毛受らの研究や Capocci らの研究と異なる．

4. ページのトピッククラスタリング

まずトピックの類似するページを集約するために，ページに付与されているタグ集合群の類似性に基づいてクラスタリングを行う．形成されたクラスタをトピッククラスタと呼ぶ．トピッククラスタに属するページは，類似する特徴を持つと考えることができ，トピックごとの特徴の分析が可能になる．

4.1 タグ付けの類似性に基づくトピッククラスタリング

SBM において，あるページは様々なユーザにタグ集合を付加されブックマークされる．ユーザごとに価値観や興味が異なるため，対象とするページに付与するタグは異なり，タグを付与しないユーザも存在する．また，タグはユーザがページのコンテンツを閲覧し，ページに対する何らかの価値基準，分類基準に基づいて抽象化されたキーワードがタグとして用いられる．各ユーザのタグ付与行動の差異はあるが，複数のユーザによって付与されたタグを集約すると，ページが閲覧者の観点からどのようなキーワードで代表され，どのような話題・トピックを扱っているのかを俯瞰できる．そのため，ページのコンテンツ内の頻出単語を用いるよりも，複数ユーザによって付加されたタグの頻出集合を用いる方が，ページのコンテンツを高度に抽象化し，少ないキーワードでページの話題やトピックを代表し，高度な要約を行うことができる [7].

ページのトピックを代表する情報として，対象ページをブックマークしているユーザ群が付与したタグ集合群の *TF-IDF* を利用する．TF-IDF は，ページ中の出現単語の重要度を測る方法であり，ページ中の特徴的な単語を抽出することができる．

本研究では，TF は，ページごとにタグ a の使用回数に対応し，IDF は，クローリングしたデータセットに対し，対象タグの付与されているページ数から算出する．ページ p に対するタグの使用頻度の TF-IDF ベクトルをトピックベクトル tv_p と呼ぶ．トピックベクトルの属性となるタグには，DF が 20 以上かつ，NG タグ (表 1) に列挙されているタグ以外のものを用いる．これらのタグは，コンテンツの内容を示すものとしてふさわしくないタグと考えられるため除外した．

トピックベクトルの類似性に基づき，トピックの類似するページをクラスタリングによって集約する．ページ間のトピックの類似度は，トピックベクトル間の類似度計算により算出する．トピックベクトルの類似度計算には，コサイン類似度を利用する．また本研究は，適切にページ群をクラスタ化するために，以下のような 2 段階のクラスタリングを行う．

表 1　NG タグ

これはひどい	これはすごい	これはエロい	これはこわい
これはいい	これはえがい	これは欲しい	あとで行く
あとでよむ	あとで読む	あとで見る	あとでみる
あとで買う	あとで試す	ほしい	読んだ
欲しい	行きたい	おもしろい	なるほど
興味深い	かわいい	作ってみた	歌ってみた
踊ってみた	演奏してみた	描いてみた	何か質問ある？
質問ある？	あたまがわるい		

（1） 被ブックマーク数 20 以上のページ群を用いたベースクラスタの形成

（2） 1. で対象にならなかったページ群のクラスタへの併合

多くのユーザにブックマークされているページのトピックベクトルはある程度信頼できる情報である．しかし，ブックマーク数の少ないページは，付与されているタグの種類や使用回数が少なく，トピックを同定する際に用いる情報として信頼に足りるとは言い難い．一段階目のクラスタリングでは，信頼できると思われるページ群のみを対象とすることで，トピッククラスタのベースとなるクラスタをより適切に形成する．この一段階目のクラスタリングでは，クラスタリングツール bayon [13] による bisecting k-means のアルゴリズム [14] を用いた，非階層クラスタリングを行う．bisecting k-means によって形成されたクラスタは，クラスタ重心 (セントロイド) を持つ．各セントロイドは，各タグを基底ベクトルとするクラスタの重心座標を持つ．この座標へのベクトルをセントロイドベクトルと呼ぶ．クラスタリングを行ったページには，メンバーとなったクラスタのクラスタラベル (1〜$|C|$) を付与する．

第二段階のクラスタリングでは，第一段階で対象にならなかったページと，セントロイドのセントロイドベクトルとの類似度を測り，最も類似するセントロイドを持つクラスタへ併合する．タグが全く付与されていないページには，クラスタラベルとして 0 を付与し，クラスタリングの対象とならなかったことをマークする．

4.2 クラスタリング結果

本節では，前節で述べたクラスタリング手法を用いて SBM データをクラスタリングした結果を例示する．クラスタリングの対象としたデータは，はてなブックマークからクローリングしたデータであり，約 120 万ページ，約 2000 万ブックマークのデータセットである．これらのデータは，2009 年 12 月 12 日時点のはてなブックマークで提供されている API を用いて取得できるデータに基づいている．

まず，クラスタ数 $|C|$ の値を 300〜2000 の範囲でクラスタリングしたときの，クラスタ内のページ間類似度 (IntraSim) とクラスタ間のセントロイドの類似度 (InterSim)，それらに基づくクラスタの凝集度 (Cohesion) を図 2 に示す．それぞれ，以下のような計算式によって求めることができる．

$$IntraSim(c) = \sum_{p_i, p_j \in c} CosineSim(tv_{p_i}, tv_{p_j}) \qquad (3)$$

$$InterSim(c_i, c_j) = CosineSim(cv_{c_i}, cv_{c_j}) \qquad (4)$$

表 2　クラスタの構成要素

クラスタ ID	ページ数	第一主成分タグ	係数	第二主成分タグ	係数	第三主成分タグ	係数	第四主成分タグ	係数
50	55	mindmap	0.816	マインドマップ	0.501	software	0.221	freemind	0.136
100	193	vimperator	0.952	firefox	0.280	vim	0.101	plugin	0.0406
150	97	photoshop	0.790	素材	0.522	フリー素材	0.145	テクスチャ	0.144
200	166	rss	0.829	blog	0.522	feed	0.121	bloglines	0.0712
250	430	livedoor	0.975	ライブドア	0.200	きっこ	0.0291	fon	0.0241
300	385	雑学	0.978	資料	0.0994	漢字	0.0639	言葉	0.0570
900	553	法律	0.699	law	0.637	社会	0.196	国籍法	0.142
950	2746	ruby	0.988	programming	0.104	プログラミング	0.0395	perl	0.0360
1000	4270	ニコニコ動画	0.981	niconico	0.110	nicovideo	0.0641	ひろゆき	0.0453

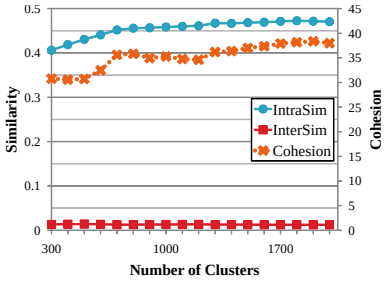

図 2　クラスタ数の違いによるクラスタ類似度の違い

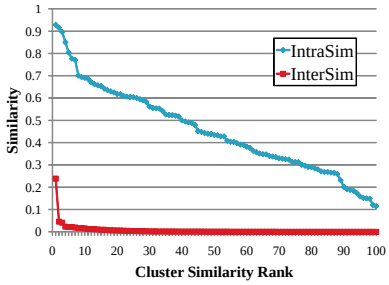

図 3　クラスタ内・間の類似度

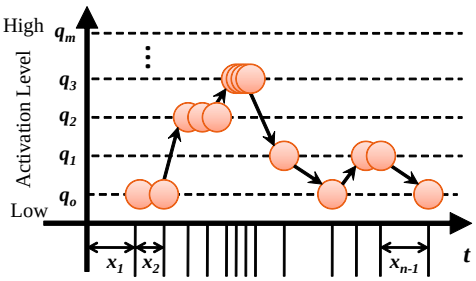

図 4　活性度の状態遷移系列

$$Cohesion(C) = \frac{average_{c \in C}(IntraSim(c))}{average_{c_i, c_j \in C}(InterSim(c_i, c_j))} \quad (5)$$

cv_c はクラスタ c のセントロイドの基底ベクトルである．クラスタ数が 700 以降は凝集度は大きく変わらない．また，クラスタ数が 1000 を超える，クラスタが扱うトピックの粒度が細かくなりすぎてしまうことを目視で確認した．そのため，本研究ではクラスタ数 $|C| = 1000$ を採用した．表 2 は，クラスタリングで得られたいくつかのクラスタの構成要素を示している．

次にクラスタ数 $|C| = 1000$ のときの各クラスタの IntraSim 及び，InterSim を示す．クラスタ数が 1000 の場合，クラスタの組み合わせは，約 100 万となる．これをグラフとして例示することは困難であるため，無作為に 100 クラスタを抽出し，さらにこの 100 クラスタ間の 100 クラスタペアを無作為に抽出した．図 3 には，この 100 クラスタの intra sim と 100 クラスタペアの InterSim を，それぞれの値の降順に並び替えた結果を示す．IntraSim は，InterSim よりも非常に高い類似度を示している．また，InterSim は非常に小さな値であり，ほとんどが 0 に近い．このことから，各クラスタは類似するトピックを扱うページによって構成され，異なるクラスタ間では，トピックがほとんど類似しないことが分かる．

5.　活性度推定

5.1　ページベースの活性度推定

時系列データの活性度に関する研究には，Kleinberg [15] の研究がある．Kleinberg は，文章ストリームにおける文章の到着頻度に着目し，特定のトピックの活性度 (バーストの強度) を分析する手法を提案した．Kleinberg による手法では，隠れマルコフモデル (HMM) を用いて，次の文章ストリームの到着確率が指数関数的に異なる複数の状態を内部状態とし，確率的に状態遷移が行われるモデルを提案している．文章ストリームのトピック分析や注目語の抽出に Kleinberg の手法は有効な手法として知られ，よく用いられている．

先行研究 [2] では，SBM 上のページの活性度を，Kleinberg の手法と同様に HMM を用いてモデル化した．あるページに対するブックマークの活性度が高い状態では頻繁にブックマークがなされ，次にブックマークされるまでの時間間隔が短くなる可能性を高くする．活性度が低い状態では稀にしかブックマークされないため，次にブックマークされるまでの時間間隔が長くなる可能性を高くする．

HMM の内部状態として，平均ブックマーク頻度 $\alpha_0 = \hat{g}^{-1}$ を持つ状態 q_0 を活性度の基準状態とする．$\hat{g}^{-1}$ は平均ブックマーク頻度を表し，ページのブックマークの時系列情報から導出する．状態 q_i は，状態番号 i の値に従って，異なる高さの活性度を持つ．また，次のブックマークの出現確率が以下のような式で表わされる指数確率密度関数 f_i に従う．

$$f_i(x_i) = \alpha_i e^{-\alpha_i x_i} \qquad \alpha_i = \hat{g}^{-1}\,\beta^i \quad (6)$$

このとき，$\beta(>1)$ は各状態のブックマークの生起頻度の分解能を表す．状態番号 i が増加する程，短い時間間隔で次のブックマークが生起しやすく，減少する程，次のブックマークが生起するまでの時間間隔が長くなる (図 4)．提案するモデルでは，q_0 を基準に，$M = 2m+1$ 個の異なるブックマークの生起頻度を持つ状態から成る内部状態系列 $\mathbf{q} = \{q_{-m}, ..., q_{-1}, q_0, q_1, ..., q_m\}$ を規定する (図 5)．

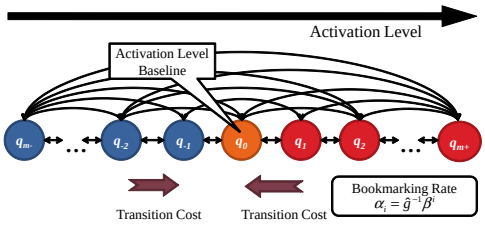

図 5　SBM の活性度モデル

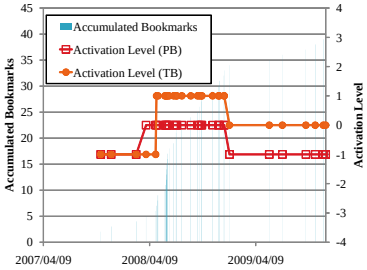

図 6　活性度推定の例

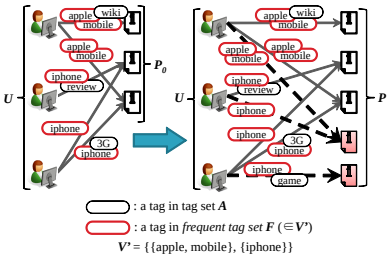

図 7　S-BITS の二部グラフ構築

一連の n ブックマークの時間間隔 $\mathbf{x} = (x_1, x_2, ..., x_{n-1})$ が与えられたとき，各時間間隔 x_i は付随する HMM の内部状態に応じて確率的に出力される記号とみなす．この HMM を用い，ブックマークの時間間隔 $\mathbf{x} = (x_1, x_2, ..., x_{n-1})$ から最適な状態遷移系列 $\mathbf{s} = (s_1, s_2, ..., s_{n-1})$ (s_j は状態番号) を状態遷移コストが最小になるように，ビタビアルゴリズムを用いて推定する．これによって，あるページの活性度の時間変化系列を分析することができる．

また，各状態 q_k から q_l へ状態遷移するとき，以下の式で表わされるような状態遷移コスト $\tau(k, l)$ が発生するものとする．

$$\tau(k, l) = |l - k|\gamma \ln n \ln SD \qquad (7)$$

このとき，$\gamma(> 0)$ は状態遷移の起こりやすさを設定するパラメータである．この状態遷移コスト $\tau(k, l)$ により，ブックマークの生起頻度が大きく異なる状態への遷移は難しくなり，たまたま頻繁なブックマークの追加が発生したり，ブックマークが発生しなかったりといった，偶然発生するようなノイズに強くすることができる．また，n はブックマーク数，SD はブックマーク時刻の標準偏差を表す．これらのパラメータにより，多量のブックマークを得ており，長期間に渡ってブックマークされているページほど，状態遷移にかかるコストが大きくなるため，頻繁な状態遷移はより起こりづらくなる．

SBM サービスからは，現在までのブックマーク系列のスナップショットを取得できる．この系列情報だけでは，現在時刻における活性度を推定不可能であるため，現在時刻に仮想的なブックマークを付加したブックマーク系列 $\mathbf{x}' = (x_1, x_2, ..., x_{n-1}, x_n)$ を対象に活性度を推定する．推定された活性度の状態遷移系列 $\mathbf{s}' = (s_1, s_2, ..., s_{n-1}, s_n)$ の最後の状態番号 s_n が表す状態 q_{s_n} が，現在における対象ページ p の活性度 act_p となる．

5.2　トピックベースの活性度推定

ページベースの活性度推定では，過去のブックマークの系列の情報量が十分でない場合，平均ブックマーク頻度 $\hat{g}_c$ の導出がロバストではなく，活性度の推定結果も安定的ではない．また，ページ独自の基準であるため，他のページと単純に活性度の比較を行うことができない．

一方，同じトピックを扱うページは，嗜好の類似するユーザにブックマークされる傾向にあり，ブックマークのパターンも類似すると考えられる．そこで，トピックに属するページ群のブックマーク系列を考え，その平均ブックマーク頻度を，そのトピックのページ群に対するブックマーク生起頻度基準とする．これにより，よりロバストで適切な活性度の推定，及び，特定トピック内で活性度が高いページの発見が可能となる．

トピック c に属するページの平均ブックマーク間隔 $\hat{g}_c$ は，式 8，9 のブックマーク時刻の系列から式 10 のように導出する．

$$T_p = \{t_i | t_i \in b_p \wedge t_{i-1} < t_i \wedge 0 \leqq i \leqq n-1\} \qquad (8)$$

$$T_c = \cup_{p \in c} T_p \qquad (9)$$

$$\hat{g}_c = \frac{t_{|T_c|-1} - t_0 + 1}{|T_c|} \qquad (10)$$

b_p はページ p をブックマークする全ユーザのブックマークの集合を表す．上記より，平均ブックマーク頻度 $\hat{g}_c^{-1}$ が導出できる．

内部状態数 $M = 9(m = 4)$，内部状態 q_{-4}〜q_4 の HMM を用い，パラメータ $\beta = 4$, $\gamma = 10$ と設定し，活性度の分析を行う．活性度を推定した例を図 6 に示す．この例では,「朝バナナダイエット(注1)」に関する Web ページの活性度変化を示している．このページ自体は注目度が低下しているが，属するクラスタ内の平均程度の注目度があるため，トピックベースの推定 (TB) では，活性度が標準であり鮮度がある程度保たれていると推定されている．

6.　ランキング

先行研究では，S-BITS にページの活性度を加味した手法 S-BITS*を提案した．過去に提案した手法は，ページベースの活性度推定に基づくものである．これを S-BITS*PB と呼ぶこととする．本研究で提案したトピックベースの活性度推定に基づく手法を S-BITS*TB と呼び，手法の有効性を検証する．

6.1　S-BITS

S-BITS(Social-Bookmarking Induced Topic Search) は，SBM におけるページとユーザ間の 2 部グラフを対象にしたページのランキング手法である．S-BITS では，SBM ユーザの Hub 度によって表わされる専門性と，ページとユーザとの相互関係から導かれるページの Authority 度によって，ページを評価する．また，HITS [16] がページ間の in-link， out-link を利用して，対象とするページ集合を拡張しているのに対し，S-BITS では，タグの共起に着目し，共通のタグ集合を利用することで，対象ページ集合の拡張を行う (図 7)．S-BITS のアルゴリズムの概要は以下の通りである．

(注1)：http://www.asabanana.net/

Algorithm 1 Scoring Algorithm of S-BITS

$p\_score^0 := \{1,1,1,...,1\}$ $u\_score^0 := \{1,1,1,...,1\}$
$k := 0$
repeat
  $k := k+1$
  for all $p_i \in P$ **do**
    $p\_score_i^k := \sum_{b_{j,i} \in B} u\_score_j^{k-1}$
  end for
  for all $u_i \in U$ **do**
    $u\_score_i^k := \sum_{b_{i,j} \in B} p\_score_j^{k-1}$
  end for
  normalize(p_score^k) and normalize(u_score^k)
until $|p\_score^k - p\_score^{k-1}|_1 < \epsilon_p$ and $|u\_score^k - u\_score^{k-1}|_1 < \epsilon_u$
return p_score^k and u_score^k

（1） 検索クエリ q を与え，q を用いて検索エンジンから上位 n 件のページを収集する (初期ページ集合 P_0)．SBM サービスを利用し，初期ページ集合 P_0 の各ページ p_i のブックマーク情報 $b_{i,j}(=(p_i,u_j,t_{i,j},A_{i,j}))$ を収集する (ブックマーク集合 B)．収集した全ユーザからユーザ集合 U，全タグ集合からルートタグ集合 V を生成する．

（2） タグ集合を利用して関連性のあるページを収集する．V 中において，頻出なタグ集合 F を抽出する (V')．頻出タグ集合 F は相関ルールマイニングの極大頻出アイテム集合抽出によって得る [17]．頻出タグ集合 $F \in V'$ を包含するタグ集合 A でブックマークされているページを収集し，P_0 とマージする (ページ集合 P)．

（3） ページ集合 P，ユーザ集合 U，ブックマーク集合 B からなるグラフ G を対象に，ページの Authority スコア (p_score)，ユーザの Hub スコア (u_score) を計算し，Authority スコアを基に，ページのランキングを行う (アルゴリズム 1)．

6.2 活性度を考慮する手法

前節で提案した活性度の評価をページのランキングに反映させ，ランキングの適合率を向上させることを考える．本研究では，ページに対して評価を行った活性度の値を，ページの重みとしてユーザ-ページ間のブックマークに与え，S-BITS のページスコア，ユーザスコアの計算の際に，この重みを反映させる．提案した活性度の評価手法では，活性度を整数で表わしている．

検索者の時間軸に関する評価の多様性に対応するために，活性度がページの評価に対して与える影響力を変更可能にすることを考える．これを実現するために、シグモイド関数 (式 11) を用い，活性度を 0〜1 に正規化する．

$$\varsigma_\lambda(z) = \frac{1}{1+\exp(-\lambda z)} \tag{11}$$

シグモイド関数では，パラメータ λ の値を 0 に近づけることで，$\varsigma_\lambda(z) = 0.5$ に漸近し，∞ に近づけることで，大きさ 1 のステップ関数に漸近する．シグモイド関数を用いて活性度を正規化することで，活性度の高低をより顕著にしたり，鈍化させたりと言ったように，0 から 1 の範囲で活性度が与える影響を調整することができる．正規化した活性度を評価値伝播の重みとして加味し，以下のような評価式でスコアを算出する．

$$p\_score_i^k = \sum_{b_{i,j} \in B} \varsigma_\lambda(act_{p_i})\ u\_score_j^{k-1} \tag{12}$$

$$u\_score_i^k = \sum_{b_{j,i} \in B} \varsigma_\lambda(act_{p_j})\ p\_score_j^{k-1} \tag{13}$$

この評価式を S-BITS のアルゴリズムに当てはめ，ページスコアベクトルとユーザスコアベクトルが平衡を迎えるまで計算を繰り返し，定常状態となったページスコアを基に，ページのリランキングを行う．これにより，現在においても，情報が価値を維持しているかどうかを考慮した上で，ユーザの検索対象ページから成るコミュニティにおける専門性を考慮したページのランキングが可能となる．S-BITS に活性度を考慮した手法を S-BITS*と呼ぶ．本研究では，$\lambda = 1$ のシグモイド関数を用いる．これは，予備実験により，人々の価値基準に近く，S-BITS に対する評価を最も向上することが可能であったためである．

7. 評価実験

この節では，S-BITS*TB の有効性について議論する．有効性を測るために，被験者実験を行った．

7.1 比較手法

比較対象のランキングは，S-BITS に活性度を加味する手法である S-BITS*TB，PB，時間経過に基づく衰退を扱う Aging S-BITS (半減期 $t_{1/2} = 30$ 日, 60 日, 120 日)，SBRank に活性度を加味する手法である SBRank*TB，PB，S-BITS，SBRank，Yahoo!のランキングである．SBRank*，Aging S-BITS については，以下に詳細を記載する．

7.1.1 SBRank*

Yanbe らは，Web ページのブックマーク数を基に Web ページのランキングをリランキングする手法を提案した [5]．彼らの研究では，ブックマーク数を SBRank 値と呼び，ページの有用性を示す指標としている．Yanbe らが提案した SBRank に活性度評価を導入することで，改善することを目指し，活性度を加味することの有効性を示す．S-BITS*と同様にシグモイド関数で活性度を正規化し，SBRank 値に加味させる．この手法を SBRank*と呼ぶ．計算式は以下の通りである．

$$SBRank*(p) = \varsigma_\lambda(act(p))\ SBRank(p) \tag{14}$$

$act(p)$ は現在のページ p の活性度，$SBRank(p)$ はページ p の SBRank 値 (被ブックマーク数) を表す．この評価式を用いて，ページを評価しランキングを行う．

S-BITS*と同様に，各トピックの平均ブックマーク頻度を活性度の基準として用いたものを SBRank*TB，ページ固有の平均ブックマーク頻度を活性度の基準として用いたものを SBRank*PB と呼ぶ．

7.1.2 Aging S-BITS

ブックマークの時系列情報から鮮度を推定する簡単な方法として，ブックマークを行った時刻からの経過時間に基づき，鮮度を減衰させていく方法がある．これは，ブックマークが行われた時刻における鮮度を，鮮度の最大値 F_0 とし，以下のよう

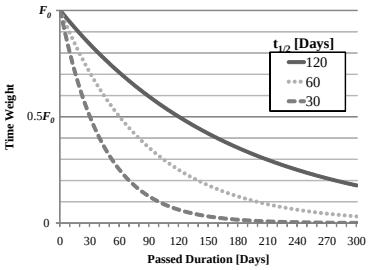

図 8　TimeWeight

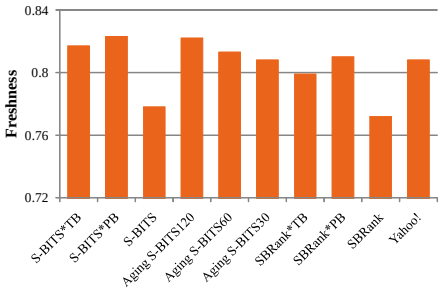

図 9　鮮度評価の平均

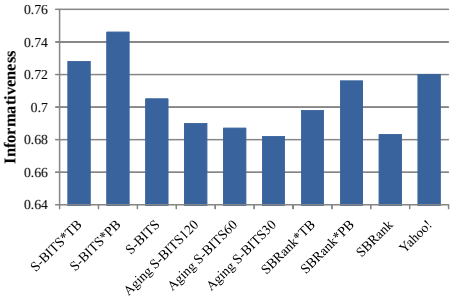

図 10　有用性評価の平均

な指数関数によるエイジングを施す方法である (図 8).

$$TimeWeight(t) = F_0\ e^{-\frac{log2}{t_{1/2}}(t'-t)} \qquad (15)$$

t はブックマークを行った時刻であり，t' はこの評価を行う時刻を表す．$t_{1/2}$ は，エイジングにより出力される値が，$0.5F_0$ になるのに要する経過時間である半減期を表している．ゆえに，$TimeWeight(t)$ は，経過時間に基づいたブックマークの老化減少を振舞う関数の一つである．

ブックマークのエイジングの概念を取り入れると，ブックマークのタイムスタンプの古さ，年齢に応じて，その鮮度が減衰する．これを S-BITS の各エッジに重みとして与え，ページ，ユーザ双方に伝播させることで，ブックマークからの時間の経過を加味させることができる．ページ，ユーザのスコアの計算式は，以下のようになる．

$$p\_score_i^k = \sum_{b_{i,j} \in B} TimeWeight(t_{i,j})\ u\_score_j^{k-1} \qquad (16)$$

$$u\_score_i^k = \sum_{b_{j,i} \in B} TimeWeight(t_{j,i})\ p\_score_j^{k-1} \qquad (17)$$

ブックマークのエイジングを加味する手法を *Aging S-BITS* と呼ぶ．Aging S-BITS は，各ブックマークがされてからの経過時間を加味している．しかし，単純に経過時間のみを加味した場合，ページのコンテンツの違いによる鮮度の持続時間の違いを加味することができない．また，SBM は再利用を目的としたメディアである従来のブックマークという性質も持ち合わせているため，時間の経過と共に価値が失われていくというよりも，その注目度の変化に着目することによる鮮度の衰退や持続の度合いを測る手法の方がより適切であると考えられる．評価実験では，Aging S-BITS を比較対象の一つとして利用する．

7.2　評価方法

本実験で用いた SBM のデータセットは，はてなブックマークから 2008 年 7 月〜2009 年 12 月の期間にクローリングを行った，約 120 万ページ，約 2000 万ブックマークを持つデータセットである．各手法の初期ページ集合 P_0 は，Yahoo! Web 検索 API [18] を利用して，検索語を入力して得られる上位 200 件を取得する．Yahoo!のランキングも同様に，Yahoo! Web 検索 API から取得する．また，クラスタ分析と活性度推定はあらかじめ行っておく．検索実行時には，データベースから参照することで活性度の値を取得する．

20 人の被験者を募り，以下のような検索語で評価を依頼した．

- 検索語: iphone, java, ruby, php, web design,
  論文 書き方, 英語 学習, レシピ,
  ダイエット, プロジェクトマネージメント

被験者にはそれぞれ 5 つの検索語について，各ランキングの上位 20 件のページの評価を依頼した．どの手法であるのかと言った心理的な影響を回避するために，各検索語ごとに全手法の評価対象ページ群を一つのリストに集約し，重複をなくし，ランダムに順序を並び替えて被験者に提示する．評価項目は以下の 3 点であり，それぞれの有無の評価を依頼した．

- ページコンテンツが検索語と関連があるか (関連度)
- ページコンテンツが鮮度を保っているか (鮮度)
- ページコンテンツが有益な情報であるか (有用性)

図 9 に，関連度 AND 鮮度の評価結果を示す．図 10 に，関連度 AND 有用性の評価結果を示す．以降，有用性と言った場合には，関連度 AND 有用性を指し，鮮度についても同様である．比較手法が多いため，上位 10 件における結果のみを示す．

7.3　実験結果・考察

鮮度に関する評価では，S-BITS*PB と S-BITS*TB が高い評価を得ている．S-BITS*PB，S-BITS*TB 共に，活性度推定に基づくランキング手法である．また，S-BITS*PB にやや劣っているが，クラスタごとのブックマーク頻度基準に基づく活性度推定を利用する S-BITS*TB も良い評価を得ている．活性度推定に基づくランキング手法高い評価を得ていることが分かる．活性度推定の方法の違いが差として現れていると考えられ，ページベースの手法の方が高い評価を得ている．Aging S-BITS は S-BITS を上回る結果を示しているが，S-BITS*には劣る．S-BITS を上回っていることから，鮮度においては Aging S-BITS にある程度の有効性があると言える．

有用性に関する評価でも同様に，S-BITS*PB と S-BITS*TB が高い評価を得ている．また，Yahoo!のランキングも高い評価を得ている．S-BITS のようなリランキング手法は，Yahoo!のランキングを基に，SBM 情報を利用してリランキングをする．特に S-BITS や S-BITS*は，SBM に登録されていないページは検索対象としていないため，Yahoo!のオリジナルのランキングが非常に高い評価を得ている場合，精度を下げてしまっている．しかし，Yahoo!があまり良い結果を示していない場合 (有用性評価が 0.8 未満) のときは，提案手法で評価を改善することができた (図 11)．SBRank*TB は S-BITS*TB と同様に，SBRank*PB にやや劣るが，SBRank より良い評価を得てい

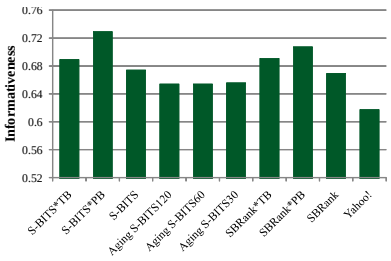

図 11　Yahoo!の評価 0.8 未満の
検索語における有用性評価の平均

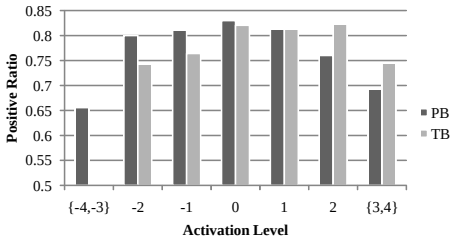

図 12　活性度推定結果と
被験者の鮮度評価との相関

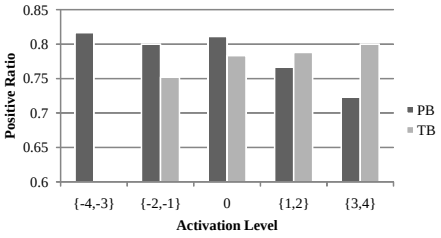

図 13　活性度推定結果と
被験者の鮮度評価との相関
(ブックマーク数 20 以下)

る．Aging S-BITS は S-BITS を上回るに至っていない．Aging S-BITS によるランキングは，単純な時間の経過と，最近のブックマークの量に依存する．ページのブックマーク頻度や，その時間変化を扱うことはできるが，ページの潜在的な注目度の異種性を考慮することができない．S-BITS*は，活性度を加味することでブックマーク頻度や，ブックマークの持続期間と言った情報を扱っている．活性度推定では，単なる時間の経過だけでなく，その潜在的な注目度に基づいた時間軸の評価を行っている．これが人々の活性度や鮮度の概念をより的確に捉えたため，S-BITS*PB，TB が良い結果を示したと推測できる．

最後に，提案した活性度評価が妥当であるかについて考察を行う．活性度の値と被験者の鮮度評価の関係を図 12 に示す．活性度が低くなるに連れて，被験者に鮮度があると評価される割合が減少している．活性度の推定値と被験者の鮮度評価との間には一定の関連があると言える．また，ページベースの活性度推定では，ブックマークの時系列情報が十分な大きさでないとき，適切に活性度を推定できないという問題点があった．トピックベースの活性度推定は，被験者の鮮度評価と関連のある結果を示している (図 13)．以上より，活性度推定にある程度の妥当性があり，特にトピックベースの推定法がページベースの推定法の欠点を改善し，有効であることが示唆できる．

8. ま と め

SBM のページに対して，トピックを同定し，属するトピックの平均的なブックマーク頻度を基に，活性度を推定する手法の提案を行った．また，トピックベースの活性度を加味した S-BITS の改良手法の提案も行った．被験者実験では，先行研究で提案した手法や，他の手法との比較を行い，今回の提案手法にある一定の有効性があることを示した．加えて，トピックを同定に必要なクラスタリングの妥当性について議論を行い，活性度推定の妥当性についても被験者実験で示した．今後は，活性度のモデルの洗練や，検索システムの開発を考えている．

謝辞　本研究の一部は科学研究費補助金特定領域研究 (＃21013004）による．

文　献

[1] T. Takahashi and H. Kitagawa, “S-BITS: Social-Bookmarking Induced Topic Search,” Proceedings of the Ninth International Conference on Web-Age Information Management (WAIM 2008), pp.25–30, 2008.

[2] T. Takahashi, H. Kitagawa and K. Watanabe, “Social Bookmarking Induced Active Page Ranking, ” IEICE Transaction of Information Systems. (to appear)

[3] “Delicous.” `http://delicious.com`.

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

[5] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka, “Towards Improving Web Search by Utilizing Social Bookmarks,” Proceedings of the 7th International Conference on Web Engineering (ICWE 2007), pp.343–357, 2007.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web,” Technical report, Stanford Digital Library Technologies Project, 1998.

[7] X. Li, L. Guo, and Y. Zhao, “Tag-based social interest discovery,” Proceedings of the 17th International World Wide Web Conference, pp.645–654, 2008.

[8] X. Wu, L. Zhang, and Y. Yu, “Exploring social annotations for the semantic web,” Proceedings of the 15th International Conference on World Wide Web (WWW 2006), pp.417–426, 2006.

[9] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, “Trend detection in folksonomies,” Proceedings of the First International Conference on Semantics And Digital Media Technology (SAMT), pp.56–70, 2006.

[10] P. Heymann, G. Koutrika, and H. Garcia-Molina, “Can social bookmarking improve web search?,” Proceedings of the international Conference on Web Search and Web Data Mining (WSDM 2008), pp.195–206, 2008.

[11] T. Menjo and M. Yoshikawa, “Trend Prediction in Social Bookmark Service Using Time Series of Bookmarks,” Proceedings of WWW2008 Workshop on Social Web Search and Mining (SWSM 2008), 2008.

[12] A. Capocci, A. Baldassarri, V. Servedio, and V. Loreto, “Statistical properties of inter-arrival times distribution in social tagging systems,” Proceedings of the 20th ACM conference on Hypertext and hypermedia, pp.239–244, 2009.

[13] “bayon, a simple and fast hard-clustering tool..” `http://code.google.com/p/bayon/`.

[14] M. Steinbach, G. Karypis, and V. Kumar, “A comparison of document clustering techniques,” KDD workshop on text mining, pp.525–526, Citeseer, 2000.

[15] J. Kleinberg, “Bursty and hierarchical structure in streams,” Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.91–101, 2002.

[16] J. Kleinberg, “Authoritative sources in a hyperlinked environment,” Journal of the ACM, vol.46, no.5, pp.604–632, 1999.

[17] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” Proceedings of the 1993 ACM SIGMOD international conference on Management of data, pp.207–216, 1993.

[18] “Yahoo! Web Search API.” `http://developer.yahoo.co.jp/search/web/V1/webSearch.html`.