ソーシャルブックマークを利用したユーザ嗜好に基づくページの評価

高橋翼† 北川博之††

† 筑波大学第三学群情報学類 〒 305-8573 茨城県つくば市天王台 1–1–1
†† 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@kde.cs.tsukuba.ac.jp, ††kitagawa@cs.tsukuba.ac.jp

あらまし 近年の情報量の爆発に伴い，情報の信頼度の重要性が強く認識されるようになってきた．一方，Web 上でユーザがブックマーク情報を作成し，興味や関心を持ったページを管理，分類，共有するサービスであるソーシャルブックマークが注目され，普及し始めている．ソーシャルブックマークでは，ユーザがブックマークするという行為やそれに付与されたタグをユーザの Web ページに対する興味や嗜好を表す指標とみなすことができる．本稿では，ソーシャルブックマーク上での Web ページを Authority，ユーザを Hub とみなし，HITS の概念を利用して Web ページの評価値を算出することにより，ユーザの嗜好，信頼度に基づいたページの評価手法を提案し，評価実験によりその有効性を検証する．

キーワード ソーシャルブックマーク，HITS，ユーザ指向，情報検索，ランキングアルゴリズム

Evaluating Web Pages Based on User Interests Using Social Bookmarks

Tsubasa TAKAHASHI† and Hiroyuki KITAGAWA††

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

Abstract With the recent information flood, trust of information is gaining a lot of attention in information utilization. Social bookmarking is a new type of information sharing services and allows individuals to bookmark and annotate web pages which he/she is interested in or impressed by. It is attracting more attention and has become more popular. In the social bookmark services, users' bookmarking and annotations given by tags are informative indicators of user interests to the web pages. In this paper, we propose a method to evaluate trust and significance of web pages based on the social bookmarks. Extending the HITS approach, we regard web pages as Authority and users as Hubs and evaluate trust and significance values of web pages. We show usefulness of the proposed approach by experimental evaluation.

Key words social bookmark, HITS, user oriented method, information retrieval, ranking algorithm

1. はじめに

近年の情報量の爆発に伴い，情報の信頼度の重要性が強く認識されるようになってきた．溢れかえる情報に埋もれてゆく情報も存在する．また，ユーザは膨大な情報から自分にとって必要な情報を取捨選択することをより一層強いられている．

一方，Yahoo!や Google などの Web 検索エンジンは私たちに非常に有益な情報をもたらし続けている．検索エンジンの登場，発展により，膨大な情報量を持つ Web から必要な情報を簡単に検索できる．これにより，知識の獲得，調査の効率が格

段に向上したことは言うまでもなく，Web 検索エンジンの功績は非常に大きい．

Google や PageRank [2] に代表される，ページ間のリンク構造を基にしたランキングアルゴリズムが現在の Web 検索エンジンの主要な Web ページの評価手法である．Google の Web 検索エンジンとしての成功により，PageRank は有効な Web ページの評価手法として広く認知された．また，PageRank と同様に広く認知されている Web ページの評価手法に，HITS [1] がある．HITS は特定のトピックによって代表される特定の Web ページ集合を対象とした手法である．PageRank，HITS ともに，あるページ A から他のページ B へのリンクをページ A からページ B への投票と見なしている．これは，リンクにはページ作成者の何らかの意図があるという前提によるものである．

近年，Blog や Wiki，SNS のような様々な形態の Web サイト，Web サービスの誕生により，ページ間には様々な形でリンク関係が生成されるようになってきた．リンク関係の中には，ページ作成者の意図とは無関係にアプリケーションによって自動生成されるものが存在し，前述のリンク構造を基にしたアルゴリズムの前提とは異なる環境となりつつある．また，それらの新たな Web サイトの増加に伴い，スパムトラックバックやコメントスパムなどによる悪質なスパム行為が増加し，問題となっている．このような環境においても，リンク構造によるページの評価手法が有効であることに変わりはないが，よりユーザの意図を反映した手法による評価や補完が必要であると考えられる．

インターネットの普及が始まったころから，Web ブラウザにはブックマークという機能が搭載されていた．ブックマークでは，ユーザがお気に入りの Web ページを登録し，独自に管理することができる．ブックマークされたページは，お気に入りのページであり，ユーザからのある評価を与えられた有益なページと言える．しかし，Web ブラウザのブックマーク機能はユーザ以外が知ることは難しく，オープンなものではなかった．

一方，近年ソーシャルブックマークに対し，注目が高まっている．ソーシャルブックマークは Web 上でユーザがブックマーク情報を作成し，興味や関心を持ったページを管理，分類，共有するサービスである．ユーザは各自のブックマーク情報に独自の注釈をタグによって与え，管理することができる．Web ブラウザのブックマーク機能と違い，他のユーザのブックマーク情報を閲覧することができたり，多くのユーザの評価を得ているページを知ることができる等，よりユーザ間の情報共有を意識したサービスとなっている．ソーシャルブックマークでは，ある Web ページに対して，何人のユーザがブックマークをしているかという情報を知ることができ，これを一種のページの評価の指標として用いることができる．また，興味の対象が異なる様々なユーザが存在し，様々なページがブックマークされている．

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ことである．しかし，ユーザによってその嗜好も異なり，ユーザによってはある特定の分野には詳しくないかもしれない．その分野におけるユーザの見識は，ユーザのその分野における信頼度とみなせる．そして，信頼のおけるユーザから評価されているページは信頼できる情報源である可能性が高く，また，信頼できるページを評価しているユーザも信頼できる可能性が高い．これは，HITS における Authority と Hub の関係と類似する．

そこで本稿では，ソーシャルブックマークにおける Web ページを Authority，ユーザを Hub とみなし，HITS の概念を Web ページとユーザの関係に拡張し，Web ページを評価する手法を提案する．本稿ではこの手法を S-BITS(Social-Bookmarking Induced Topic Search) と呼ぶこととする．S-BITS では，Web ページとユーザ間の相互のブックマーク・被ブックマークの関係を強化することで，ユーザの信頼度を評価し，またそれを通して，Web ページの評価値を算出する．これにより，膨大な情報に埋もれ，検索結果上位に表れないページや発見することが困難なページを抽出する．また，対象とするページ集合をユーザの評価を得ているページに拡張するために，ブックマークの際の頻出なタグの集合に着目することで，対象ページ集合の拡張を行う．

本稿の以降のセクションの構成は以下のとおりである．まず，2 章ではソーシャルブックマークおよび HITS について記述する．3 章では，我々が提案する Web ページの評価手法である S-BITS について，詳細を述べる．4 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．5 章では，過去の関連する研究について概観する．最後に，6 章で本稿の結論を述べると共に，今後の課題について述べる．

2. 前　　提

2.1 ソーシャルブックマーキング

ソーシャルブックマーク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．ソーシャルブックマークは Web 上でユーザがブックマーク情報を作成し，興味や関心を持ったページを管理，分類，共有するサービスである．ユーザは各自のブックマーク情報に独自の注釈をタグによって与え，管理することができる．“万人による注釈”という意味を持つ Folksonomy の概念を実現しており，人手によって注釈を与えられた情報源である．また，様々な価値観を持ったユーザによって評価を与えられ，フィルタリングされた一定の信頼性を有する情報源であると言える．

2003 年に del.icio.us [13] がサービスを開始し，現在では 100 万人を超えるユーザを獲得している．2004 年ごろから普及し始め，現在では，はてなブックマーク [14] などの様々なソーシャルブックマークサービスが存在する．ソーシャルブックマークが誕生した当初は，Web に関連したトピックに高い興味を示すユーザによって形成されたネットワークであったが，現在では，より広範囲なユーザに浸透し，政治や音楽，コミック，アートなど，より一般的な話題に対して興味を示すユーザやそれらに関連するユーザも加わり，多様な情報を持ち，かつ一定の信頼性を有する情報源へと成長し続けている．

ソーシャルブックマークは，様々なユーザによって作られる一種のソーシャルネットワークであるため，ユーザによって嗜

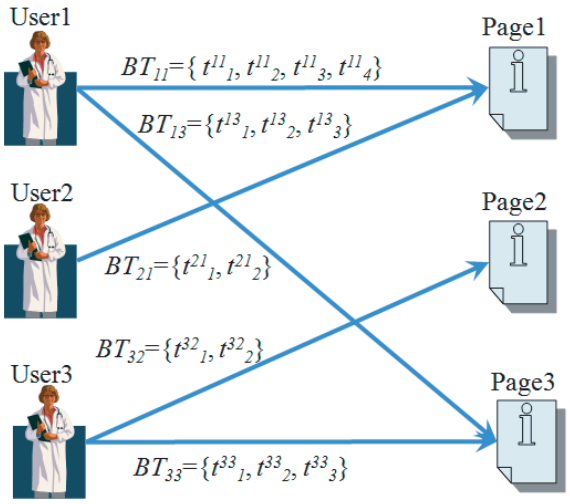

図 1　ソーシャルブックマークの構造

好も異なり，特定の情報に対する見識にはばらつきがある．なかには特定のトピックに対して非常に詳しいユーザがいたり，少しかじった程度のユーザもいたりする．一般的に，見識のある人がよい評価を与えているものは信頼度が高く，逆に見識の浅い人の評価は信頼度が低い．

ある Web ページには複数のユーザがブックマークをしており，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 Web ページの信頼度や品質を測る 1 つの指標と言える．Yanbe ら [3] はこの指標を SBRank と名付け，SBRank が Web ページの品質を測る上で有益な指標であることを示している．

また，ユーザは独自の価値観で任意のタグを用い，ページに注釈を付けることができる．多くのユーザは対象ページを表すキーワードやカテゴリーに関する単語やフレーズをタグとして用いている．興味深い点として，ユーザの感想やブラウジングの振る舞いに関するタグ，“usuful”，“おもしろい”，“あとで読む”などのフレーズもタグとして用いられることがある．多くのソーシャルブックマークサービスでは，複数のタグを 1 つの Web ページに与えることができ，タグ集合で意味を持つことも少なくない．たとえば，書籍の通信販売のサイトには “book”，“shopping”のような複数のタグが与えられる．ユーザによるタグを利用した注釈付けも，対象 Web ページの信頼度や品質を測る上で一定の有益性を有する．

ソーシャルブックマークにおける，ページ，ユーザ，タグの関係について考える (図 1)．ユーザ u_i がページ p_j をブックマークをする際，各エッジには 0 個以上のタグによって形成されるタグの集合が与えられる．本稿では，ユーザ u_i がページ p_j へのブックマークに与えたタグ集合を，ブックマークタグ集合 $BT_{ij} = \{t_1^{ij}, ..., t_n^{ij}\}$ とする．また，ユーザ u_i がブックマークタグ集合 BT_{ij} を用いて p_j をブックマークしていることを $u_i \overset{BT_{ij}}{\rightsquigarrow} p_j$ と記す．任意のタグ t_k^{ij} は，単語，フレーズによって形成される．

2.2　HITS(Hyper-link Induced Topic Search)

HITS [1] は PageRank と並び，リンク構造解析の手法として広く認知されている手法である．J. Kleinberg は Web ページに対して Hubness と Authority という 2 つの重要な概念を導いた．Hub とは多くの out-link(参照) を持つページのことである．Hub の基本的概念は ”特定のトピックの情報のまとめ役であり，多くの有用性のあるページを参照するページ ”である．一方，Authority とは多くの in-link(非参照) を持つページのことである．Authority の概念は ”特定のトピックにおいてよきコンテンツを持っていて人々に信用され，参照されるページ ”である．

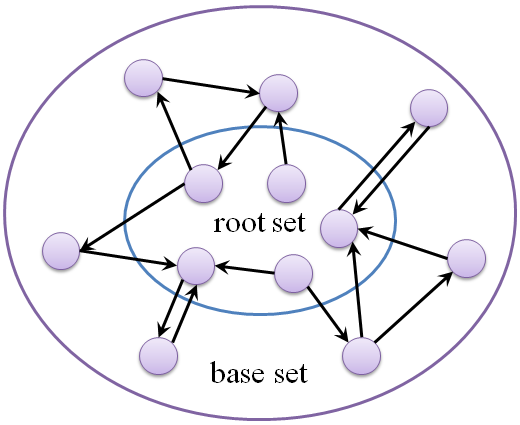

図 2　HITS におけるグラフ構造

また，良き Hub は多くの良き Authority を参照し，良き Authority は良き Hub から参照されるという関係がある．これらの関係は bipartite sub-graph によって表現される (図 2)．

各ページに対する評価は Authority スコア，Hub スコアによって評価される．ページ i の Authority スコア a_i，Hub スコア h_i は以下の式で表される．

$$a_i = \sum_{(j,i)\in E} h_j \qquad h_i = \sum_{(i,j)\in E} a_j \qquad (1)$$

ただし，(i, j) は i から j へのリンクを表し，E は (i, j) などを要素として持つ，全リンク関係の集合である．HITS では，上記の計算を両スコアが平衡の状態になるまで繰り返す．

HITS では一般的に検索エンジンにクエリを発行し，検索結果として得られる上位 n 件のページ集合を対象とする．このページ集合をルート集合と呼ぶ．その後，ルート集合に含まれる各ページが持つ out-link と in-link の参照関係にあるページを収集し，これをベース集合と呼ぶ (図 2)．ベース集合を対象とし，Authority スコア，Hub スコアを計算することにより，検索クエリによって表されるトピックにフォーカスしたグラフを対象としたリンク構造解析，およびページのランキング付けを可能とする．

3.　提 案 手 法

本章では，ソーシャルブックマークから得られる情報を利用し，ユーザ視点での有益なページの抽出手法について提案する．

本研究では，ソーシャルブックマークから得られる情報について以下の仮定をする．

- ユーザのブックマークをするという振舞はページに対して正の評価を与えることである．
- 多くの良きユーザからブックマークされているページは良きページである．
- 多くの良きページをブックマークしているユーザは良きユーザである．

上記の仮定は HITS の概念に類似するものであり，HITS の概念をソーシャルブックマークのページとユーザとの関係に適用したものである．本稿で取り上げるこれらのページの評価手法を S-BITS(Social-Bookmarking Induced Topic Search) と呼ぶ．

3.1 概　　要

HITS では，ページとページとの間の参照，非参照 (hyper-link) の関係によって作られるグラフを対象としているが，本手法では，ユーザからページへのブックマーク，ページからユーザへの被ブックマークから作られるグラフを対象としている．これにより，ユーザからのブックマークという形での評価による多数決的なページの評価が可能になる．加えて，高いスコアを与えられているより信頼できるユーザからの評価を重くすることで，ユーザの信頼度を基にした，ページの評価を行う．

また HITS では，対象とする Web ページを Web ページの持つ in-link と out-link により参照されるページの集合へ拡張する手法を取っているが，S-BITS では，頻出なタグ集合に着目し，頻出なタグ集合を包含するブックマークタグ集合でユーザがタグ付けしているページの集合に拡張する．

本手法のアルゴリズムは以下の通りである．

（1） ユーザからキーワード集合 K が与えられる．このとき，K 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 (ルート集合 R)．また，ソーシャルブックマークサービスを利用して，ルート集合 R の各ページをブックマークしているユーザを集める (ユーザ集合 U)．同時に各ページへ付与されているブックマークタグ集合 BT_{ij} を集め，それらを要素とするルートタグ集合 T を生成する．

（2） T 中において，頻出なタグ集合を抽出する (T')．U 中のユーザが頻出なタグ集合 $FT \in T'$ を包含する BT_{ij} をブックマークタグ集合として用いてブックマークしているページを収集し，R とマージする (ベース集合 B)．

（3） ベース集合 B とユーザ集合 U からなるグラフを対象に，ページの Authority スコア，ユーザの Hub スコアを計算し，Authority スコアを基に，ページのランキングを行う．

以下では，これらの各ステップについて詳しく述べる．

3.2 データ収集

まず，評価の対象となる Web ページとユーザの収集を行う．キーワード集合 K と関連のあるページを収集するために，検索エンジンから上位 n 件の Web ページを収集する．この収集したページ集合をルート集合 R と呼ぶ．ルート集合の Web ページは，従来のリンク構造や tf-idf などの評価手法によってキーワード集合 K との関連が高いページとして抽出されたものであり，ある程度の信頼度を持っていると考えられる．ルート集合 R の各ページが何人のユーザにブックマークされているか，またユーザが対象ページにどんなタグ集合を与えているかの情報をソーシャルブックマークから情報を取得することで収集する．

収集されたユーザの集合をユーザ集合 U とする．また，ユーザ $u_i \in U$ からページ p_j へのブックマークタグ集合 BT_{ij} を集め，ルートタグ集合 $T(=\{BT_{ij} | u_i \overset{BT_{ij}}{\rightsquigarrow} p_j\})$ を生成する．

3.3 対象 Web ページの拡張

HITS では，リンク関係によるグラフ作成のためにルート集合からベース集合への拡張を行った．本手法では，以下の目的のためにページの拡張を行う．

- 検索エンジンからは得られないが，ユーザの評価を得ているページを獲得する
- キーワード集合 K に関係のあるページを増やし，ユーザの信頼度をより大きなグラフで測る

より詳しく説明すると，1つ目は，多くの情報に埋もれ，検索エンジンからは発見することが困難だが，ユーザによって評価されているページを抽出するためである．2つ目は，ユーザの信頼度をより正確に計算することを通して，各ページの評価に信頼できるユーザの評価を高め，信頼できない，または，対象としているページ集合に対してあまり詳しくないユーザからの影響を小さくするためである．

対象ページの拡張は，キーワード集合 K と類似した内容を持つページに対してのみ行う．ソーシャルブックマークでは，ユーザが各ページに対してブックマークタグ集合を与え，管理している．ブックマークタグ集合は，そのページのコンテンツの内容を表すタグよって形成されていることが多数である．そこで，ルート集合に対して与えられたルートタグ集合 T 中のブックマークタグ集合に頻繁に出現するタグの組合せを抽出し，そのタグの組合せを含むブックマークタグ集合でブックマークされているページを含めることで対象ページの拡張を行うことを考える．

上記の頻繁に出現するタグの組合せを抽出するため，相関ルールマイニングの手法 [11] を応用する．各タグをアイテムとみなした場合，T 中の各ブックマークタグ集合 $BT_{ij} \in T$ はトランザクションと見なすことができる．そこで，T 中の頻出アイテム集合を抽出し，頻繁に出現するタグの組合せとして用いる．具体的には，頻出アイテム集合を最も特徴付ける極大頻出アイテム集合 [12] の集合 T' として抽出する．極大頻出アイテム集合は，最小サポート値以上のアイテム集合の中で，その超集合を持たないアイテム集合であり，最大公約数的に頻出アイテム集合を生成することができる．T' 中のタグ集合を用いることで，巨視的なトピックを表すタグ集合をできるだけ排除し，キーワード集合 K との関連が高いタグ集合を抽出することができる．

ルート集合 R を以下のように拡張して，ベース集合 B を得る．

$$B = R \cup \{\, p_j \mid u_i \overset{BT_{ij}}{\rightsquigarrow} p_j \ \wedge\ u_i \in U \ \wedge\ FT \subseteq BT_{ij} \ \wedge FT \in T' \} \tag{2}$$

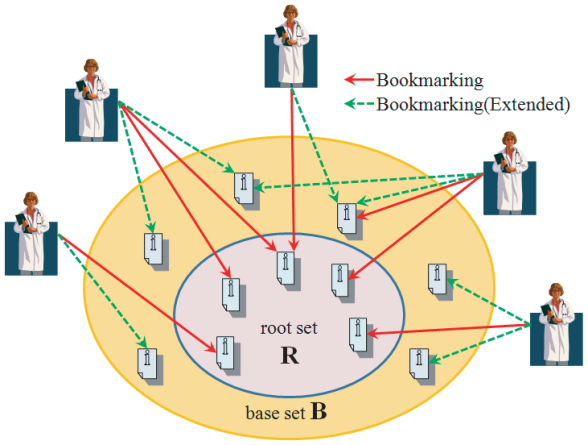

図 3　S-BITS のグラフ構造

ただし，実際には各 FT に対応して上記の条件を満たす全てのページ p_j を収集すると，収集コストが掛かる上に重複も多い．そこでランダムに選択した m ページずつ収集することとする．後述する実験では，$m = 5$ 程度で十分なページ拡張ができている．

ベース集合 B，ユーザ集合 U をノードとし，その間のブックマーク関係 E をエッジとすることで，図 3 のようなグラフが生成される．

3.4　評価値の算出

先に述べたように，ソーシャルブックマークにおけるページを Authority，ユーザを Hub とみなし，HITS の概念を拡張することで，ページ-ユーザ間の関係を相互に強化しながら，ページのスコア，ユーザのスコアを計算する．両スコアの詳細な算出方法を図 4 に記述する．ページのスコアを算出する対象は，ベース集合に含まれる各ページであり，ユーザのスコアを算出する対象は，ユーザ集合 U に含まれる各ユーザである．スコアの算出は，HITS と同様であり，ページスコア，ユーザスコアが平衡をむかえ，収束するまで計算を繰り返す．収束条件の閾値 ϵ_p，ϵ_u は十分に小さい値に設定する．α 回目の繰り返しにおける一方のスコアは $\alpha - 1$ 回の繰り返しを行った，他方のスコアの影響を受ける．繰り返し計算を行うことで，ユーザの信頼度が重みづけされてゆき，それによってページの評価値が決定される．これにより，ユーザの嗜好や信頼度を基にしたページの評価が行われる．

S-BITS
  $p^0 = \{1, 1, 1, ..., 1\}$;
  $u^0 = \{1, 1, 1, ..., 1\}$;
  $\alpha = 1$;
  Repeat
    foreach $p_i \in$ ベース集合 B
      $p_i^\alpha = \sum_{(j,i)\in E} u_j^{\alpha-1}$
    foreach $u_i \in$ ユーザ集合 U
      $u_i^\alpha = \sum_{(i,j)\in E} p_j^{\alpha-1}$

    // normalization
    $p^\alpha = p^\alpha \ / \ \|p^\alpha\|_1$
    $u^\alpha = u^\alpha \ / \ \|u^\alpha\|_1$
  until $\|p^\alpha - p^{\alpha-1}\|_1 < \epsilon_p$ *and* $\|u^\alpha - u^{\alpha-1}\|_1 < \epsilon_u$
  return p^α and u^α
end

図 4　S-BITS の評価値算出アルゴリズム

4.　評 価 実 験

提案した手法の有用性を測るために評価実験を行った．Yahoo!のオリジナルランキング，SBRank 値の優劣によるランキング，ルート集合 R のみを対象とした S-BITS のランキング，ベース集合 B を対象とした S-BITS のランキングの 4 つのランキングの妥当性を比較・検討する．以下のような環境において実装し，実験を行った．

- 検索エンジン API：Yahoo Search API [15]
- 対象ソーシャルブックマーク：はてなブックマーク
- 実装言語：Ruby 1.8.6
- キーワード集合 K：
  “open social”, “web design”, “写真 ブログ”, “年賀状”, “leopard”, “playstation3”, “action script”

SBRank，S-BITS は Yahoo! Search API から上位 200 件を取得し，そこからソーシャルブックマークサービスに登録されているページを抽出した．抽出した情報を基に，それぞれの評価手法でスコアを算出し，ランキングを作成した．この実験は，手作業による判定には作業量的限界があるため，上位 30 件のみを対象とする．

まず，キーワード集合 K に対するページの関連性を手作業で判別した．どの評価手法であるかという心理的影響を避けるために，各評価手法に関わる情報がまったく記載されていない同一フォーマットで記述されたファイルにより，検索結果の各ページがキーワード集合に適合しているかどうかを評価した．各ランキングの上位 k 件における適合率を図 5〜11 に示す．

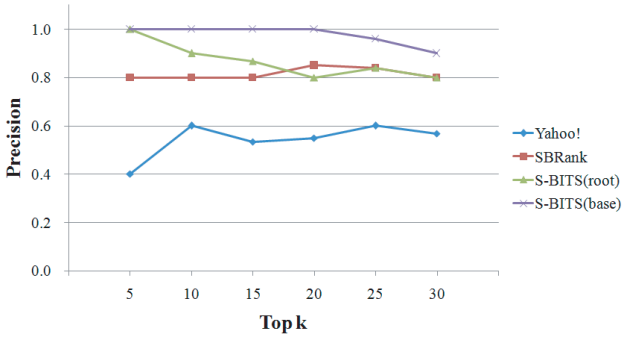

図 5　キーワード集合 “open social” におけるランク k の適合率

表 1　キーワード集合 “open social”から得られた頻出タグ集合

Tag Set	Support
api, google, sns	0.094
opensocial, google, sns	0.088
api, opensocial, google	0.072

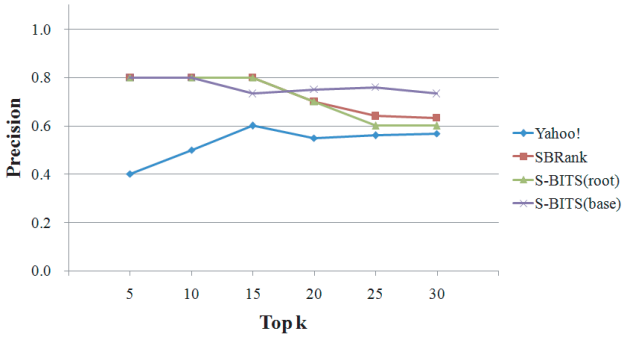

図 6　キーワード集合 “web design” におけるランク k の適合率

表 2　キーワード集合 “web design” から得られた頻出タグ集合

Tag Set	Support
web デザイン	0.242
webdesign	0.157
web, design	0.152
css, design	0.080

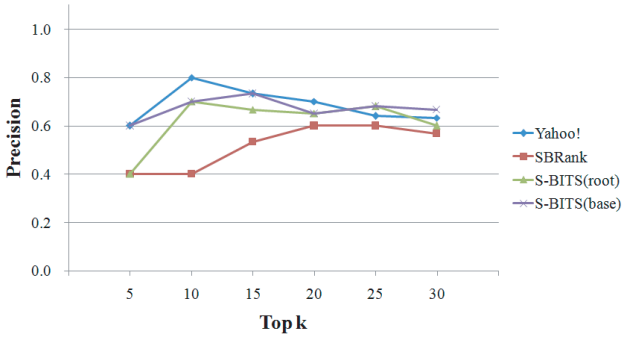

図 7　キーワード集合 “写真 ブログ” におけるランク k の適合率

表 3　キーワード集合 “写真 ブログ” から得られた頻出タグ集合

Tag Set	Support
ブログ	0.114
写真	0.110
photo, blog	0.054
ブログパーツ, blog	0.050

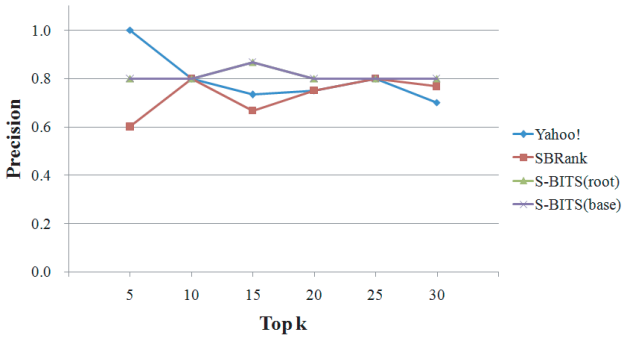

図 8　キーワード集合 “年賀状” におけるランク k の適合率

表 4　キーワード集合 “年賀状” から得られた頻出タグ集合

Tag Set	Support
印刷, 年賀状	0.069
素材, 年賀状	0.069
郵便, 年賀状	0.057

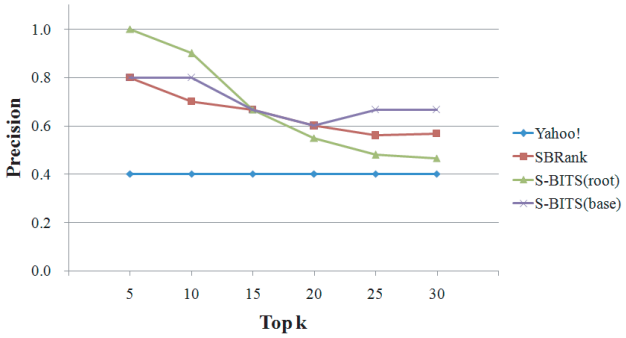

図 9　キーワード集合 “leopard” におけるランク k の適合率

表 5　キーワード集合 “leopard” から得られた頻出タグ集合

Tag Set	Support
leopard, mac	0.210
apple, mac	0.140
apple, leopard	0.074
mac os x, leopard	0.061

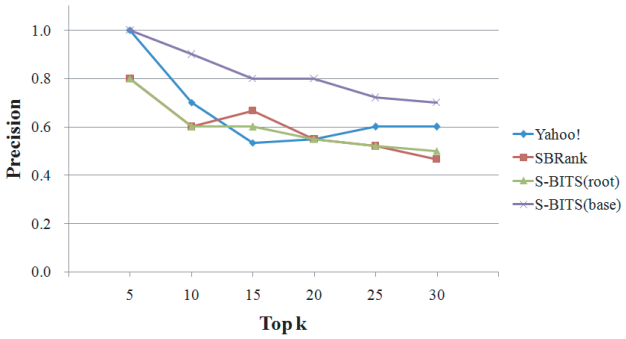

図 10　キーワード集合 “playstation3”におけるランク k の適合率

表 6　キーワード集合 “playstation3”から得られた頻出タグ集合

Tag Set	Support
ゲーム, ps3	0.073
sony, game, ps3	0.066

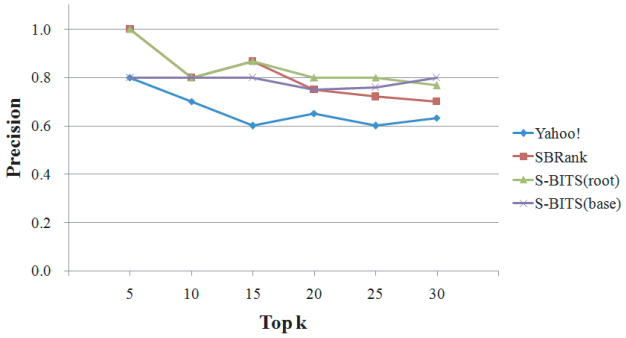

図 11　キーワード集合 “action script”におけるランク k の適合率

表 7　キーワード集合 “action script”から得られた頻出タグ集合

Tag Set	Support
actionscript, flash	0.311

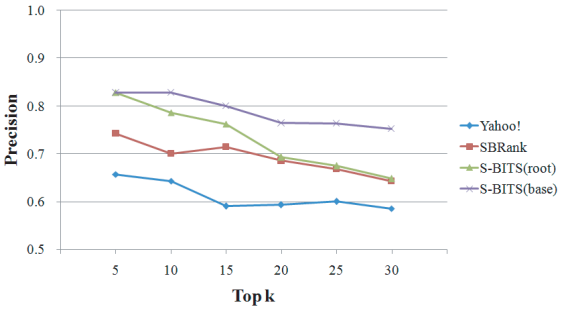

図 12　各ランク k での全クエリの平均適合率

提案手法 S-BITS は，キーワード集合 “open social”において，非常に高い適合率を示している (図 5)．対象ページを拡張したベース集合 B とした場合，キーワード集合 K と関連のある Web ページを高い精度で集めることができている．また，抽出した頻出タグ集合 $FT \in T'$(表 1) もキーワード集合 “open social”と関連のあるタグ集合を集めることができている．また，ルート集合 R を対象にした場合も，被ブックマーク数だけを評価指標とした SBRank よりも上位のランキングにおける精度が高い．純粋な Yahoo!の検索結果や SBRank 値だけの手法は “social”というフレーズに対して，多方面の Web ページがヒットしてしまったことがランキング上位で良い結果を得られないという結果になったのではないかと推測される．キーワード集合 “open social”では，頻出タグ集合 $FT \in T$ がどれも大きさ 3 という大きなアイテム集合である．例えば，表 1 の頻出タグ集合 FT である $\{api, google, sns\}$ がタグ “sns”を欠いてしまったとすると Google の様々な API を含むページ集合を表すラベルとなる．図 5 の適合率の結果からもわかるとおり，ベース集合はキーワード集合 K と関連のあるページを拡張する前と比べて適合率の高いページを多く包含しており，キーワード集合 “open social”においては対象ページの拡張手法が適切に頻出タグ集合を選択できたと考えられる．

キーワード集合 “web design”では，ソーシャルブックマークを利用した 3 手法はほぼ同様な結果だが，Yahoo!のオリジナルのランキングと比べると精度が良い (図 6)．ベース集合 B を対象とした場合には，ランキング下位のページまで関連の高いページを集めることができている．取得した頻出タグ集合 T' も，キーワード集合 K と関連の高いものを抽出できている．しかし，検索対象ページを拡張したにも関わらず，拡張していない手法や SBRank 値による手法との大きな違いを示すことができていない．また，ルート集合を対象とした場合には，SBRank 値によるものとほぼ同様の結果となっている．これは，拡張せずともユーザによる評価の高いページ集合を Yahoo!のオリジナルのランキングからルート集合 R として取得できたためではないかと推測される．加えて，SBRank 値とルート集合を対象とした S-BITS がほぼ同様な結果を示していることから，ユーザの信頼度による重みづけをせずとも，SBRank 値だけで良い評価を与えられていたということが推測される．

キーワード集合 “写真 ブログ”では，ソーシャルブックマークを利用した 3 手法ともに，よい結果を示せていない (図 7)．これは，関連の薄いトピックを扱うページがルート集合 R に含まれており，その SBRank 値が高かったため，このような結果になったと推測できる．

キーワード集合 “年賀状”では，トップ 5 の時点では差が見られるが，4 手法ともにほぼ同様な結果が得られており，精度も高い (図 8)．抽出した頻出タグ集合も，キーワード集合 K と関連が高いものと言える．

キーワード集合 “leopard”では，ソーシャルブックマークを利用した 3 手法ともに，上位のランキングにおいて高い精度を示している (図 9)．特に，ルート集合 R のみを対象とした S-BITS の精度が良い．

キーワード集合 “playstation3”では，ベース集合 B を対象とした S-BITS が非常に高い適合率を示している (図 10)．Yahoo!のオリジナルのランキングから生成されるルート集合 R を対象とした S-BITS では，ルート集合にあまり有益でない情報を持つページが含まれていたが，対象ページを拡張することにより，ベース集合 B は有益な情報を持つページを獲得することができたと推測できる．よって，キーワード集合 “playstation3”では，効果的に有益なページを収集することに成功したと言える．

キーワード集合 “action script”では，ソーシャルブックマークを利用した 3 手法が良い精度を示している (図 11)．特に，ルート集合 R を対象とした S-BITS と SBRank によるランキングが精度が高い．

上記の 7 つのキーワード集合に対する適合率の平均をとったものを図 12 に示す．我々が提案した S-BITS が良い精度を示しており，また，SBRank も Yahoo!のオリジナルのランキングに比べて良い精度を示している．Yahoo!とソーシャルブックマークを利用する 3 手法を比べると，ソーシャルブックマークを利用する手法の方が有効性が高いことが言える．ベース集合 B を対象とした S-BITS について見ると，平均して高い適合率を示すことができている．これは，検索エンジンからは得ることが難しいページにも，キーワード集合と高い関連を持つページがあることを示している．ルート集合 R を対象とした S-BITS と SBRank 値の優劣によるランキングを比較すると，我々が提案した S-BITS が，上位のランキングにキーワード集合 K と関連性の高いページを集めることができ，有効性があることを示している．

我々が提案する S-BITS は，ユーザの信頼度により重みづけすることで，SBRank と同等かそれ以上の適合率をランキング上位で示すことができた．ページの拡張手法は，キーワード集合 K と関連の高いページを収集できることを示せた．“open social”のように，キーワード集合 K が具体的かつ抽出した頻出タグ集合 T' がキーワード集合 K と高い関連性を示すことのできるページ集合に対しては，良い結果を示すことができる．逆に，“写真 ブログ”のように，頻出タグ集合自体にノイズがある場合には精度が下がる場合もある．

以上についてまとめると，提案手法 S-BITS の一定の有効性を示すことができたと言える．また，極大頻出アイテム集合を抽出した，キーワード集合 K と関連のある頻出タグ集合の抽出手法もある程度の有効性を示せていると言える．

5. 関連研究

ソーシャルブックマークサービスの普及と共にソーシャルブックマークを含む Folksonomy に関する研究 [3] [7] [8] [9] [10] はより盛んになってきている．Golder ら [10] は，ソーシャルブックマーキングの構造を詳細に分析し，ユーザの行動やタグの使用頻度などの規則性について報告している．Hammond ら [9] はソーシャルブックマークサービスに対するレビューを行っている．Xian Wu ら [7] 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，ソーシャルブックマークを取り上げ，提案している．中でも本研究と密接な関係にある研究として，山家らによる研究があげられる．山家ら [3] はソーシャルブックマーク上でページをブックマークしたユーザの数を SBRank という Web 検索の際の尺度として用い，PageRank との比較実験を行っている．また SBRank と PageRank を統合したランキング手法について提案し，SBRank 値の Web 検索の尺度としての有用性を示している．本研究は，ブックマーク件数を利用した尺度という点は同じだが，ユーザの信頼度を測ることを通してページの評価をしているという点と，検索エンジンからは得られないページにまで対象ページを拡大しているという点で山家らの研究とは異なる．

また，S-BITS の概念の基となった HITS も様々な研究 [4] [5] [6] がなされ，評価手法の改善がなされている．ただし，我々はページ-ユーザ間のブックマークの関係を利用しているため，リンク構造を対象としている彼らの研究とは異なる．

6. まとめ

ソーシャルブックマークにおけるページとユーザの関係に HITS の Authority，Hub の概念を取り入れ，ページを評価する手法，S-BITS を提案した．本稿では，ユーザの信頼度を基にページを評価することの有効性を示した．本手法は，ルート集合を対象とする場合，ルート集合 R で扱われているトピックがキーワード集合 K と関連のあるものを多く含み，複数のページがブックマークされているようなトピックに対して有効である．ベース集合 B を対象とする場合は，キーワード集合 K と密な関係にある頻出タグ集合を抽出できるときに有効性を示すことができる．しかし，ソーシャルブックマークに登録されているページの数はまだ Web 全体の数パーセントにも満たない．本手法が Web 検索全般で有効性を示すには，よりソーシャルブックマークが普及し，ユーザが増え，様々なページがブックマークされてゆく必要がある．従来のリンク構造を考慮した手法とのハイブリッドな手法や，単純な HITS の拡張による評価手法だけでなく，ブックマークされた時間やユーザのブックマークの振舞など様々な要素を取り入れた評価手法についても検討する必要がある．頻出タグの抽出手法についても，tf-idf やエントロピーのような指標を利用し，よりキーワード集合 K と密接な関係にあるものを抽出する必要がある．

謝辞

本研究の一部は科学研究費補助金特定領域研究（# 19024006）による．

文献

[1] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In Proc. of the 9th ACM SIAM Symposium on Discrete Algorithms (SODA'98), pp.668-677, 1998.

[2] 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.

[3] Yanbe, Y., Jatowt, A., Nakamura, S. and Tanaka, K. Can social bookmarking enhance search in the web?. ACM IEEE Joint Conference on Digital Libraries, 2007.

[4] Lan N., Brian D. Davidson and X. Qi. Topical Link Analysis for Web Search. In Proc. of the 29th annual international ACM SIGIR conference, pp.91-98, 2006.

[5] Ziming Z. iHITS: Extending HITS for Personal Interests Profiling. In Proc. of the 19th International Conference AINA'05, pp.747-751, 2005.

[6] K. Bharat and M.R. Henziger. Improved Algorithm for Topic Distillation in a Hyperlinked Environment. In Proc. of the 21st annual international ACM SIGIR conference, pp.104-111, 1998.

[7] X. Wu, L. Zhang and Y. Yu. Exploring Social Annotations for the Semantic Web. In Proc. of the 15th World Wide Web Conference, pp.417-426, 2006.

[8] H. Wu, M. Zubair and K. Maly. Harvesting Social Knowledge from Folksonomies. In Proc. of ACM HyperText 2006 Conference, Odense, Denmark, pp.111-114, 2006.

[9] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools(i) - a general review. D-Lib Magazine, Volume 11 Number 4, 2005.

[10] S. A. Golder and B. A. Huberman. Ther structure of collaborative tagging systems. http://www.hpl.hp.com/research/idl/papers/tags/, 2005.

[11] R. Agrawal and R. Srikant. Fast Algorithms for mining Association Rules. In Proc. of the 20th International Conference on Very Large Data Bases, pp.487-499, 1994.

[12] D. Burdick, M. Calimlim and J. Gehrke. MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases. In Proc. of the 17th International Conference on Data Engineering (ICDE '01), p.443-452, 2001.

[13] del.icio.us . http://del.icio.us/

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

[15] Yahoo! Search Web Services . Yahoo! DEVELOPER NETWORK . http://developer.yahoo.co.jp/search/web/V1/webSearch.html