情報処理学会論文誌　データベース　Vol. 1　No. 1　88–100 (June 2008)

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

山家雄介[†1]　中村聡史[†1]
アダム　ヤトフト[†1]　田中克己[†1]

ブログなどの普及により情報発信の裾野が広がるにつれて，Web検索結果から有用なページを発見するのは困難になる一方である．最近ではユーザのブックマーク行動を集約することによって価値のあるページを抽出する，ソーシャルブックマークのような取り組みがさかんになりつつある．本稿では，ソーシャルブックマークにおけるページのブックマーク数などの情報を用いて，検索結果のページの内より有用なものを上位に提示する再ランキング手法を提案する．次に，提案手法を多数のクエリに対して適用し，検索結果に含まれるページの順位変動率や，ページの種類などを調査・分類し，どのような検索目的に本アプローチが有効なのかを明らかにした．

Re-ranking Method for Web Search Based on Social Bookmark Analysis

Yusuke Yanbe,[†1] Satoshi Nakamura,[†1]
Adam Jatowt[†1] and Katsumi Tanaka[†1]

With the rise of blogs and other web applications, it is getting easier and easier to publish information. At the same time, it is getting more difficult to discern informative pages from web search results. Recently social bookmarking systems, which discover valuable pages by aggregating the bookmarking activities of many users, are getting popular. In this paper, we introduce a re-ranking method for web search results that makes use of the number of bookmarks registered with a social bookmarking service. Then, we apply our method to a number of search queries, analyze and classify characteristics of the search results, and make it clear what kind of search can be performed effectively by the method.

†1 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

1. まえがき

近年，Web検索は情報を調べるための手段として欠かせないものとなっている．初期は全文検索型の検索エンジンが一般的であり，検索エンジンが返す結果はしばしばユーザの意図に沿わないものであった．1990年代の後半に登場した，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であるPageRank[14]はWebの膨大なリンク構造の特性を利用してページの品質を推定するものであり，検索精度向上において多大な貢献を示した．PageRankはページAからページBへのリンクを，ページAからページBへの評価と見なしている．これは，リンクに人間の何らかの意図が含まれていることを前提としている．この前提は1990年代のWebの環境にうまく適合したため，PageRankはWebを組織化し人気度を測る尺度として初期のGoogleの検索結果のランキング手法に採用され，成功した．

近年増加してきたblogやwikiといった形態をとるWebサイトでは，その品質にかかわらず，アプリケーションによって自動的に生成された多数のリンクでサイトどうしが密に結び付いている．このようなリンクの多くは人の意思を反映しているとはいいがたく，PageRankがうまく働く前提に合致しない．さらにspam trackbackや，splogというスパム行為目的で自動作成されたblogサイトなどの登場により，Webのリンク構造によってページの重要性を決定するアルゴリズムの有用性は脅かされている．その結果，不適合なページが検索結果の上位にランキングされているケースが増加しつつある．リンク構造分析のアプローチは有用性をまだ失ったわけではないが，我々はWebのリンク構造を基にしたアルゴリズムに対して，何か他の補完的な尺度が必要であると考えている．

一方，近年ソーシャルブックマークに対する注目が高まっている．ソーシャルブックマークは，各ユーザのブックマーク行為を共有することでページの分類，発見などを支援するWebサービスであり，ある種の社会的アノテーションと見なすことができる．del.icio.us[⋆1]は2003年にサービス提供を開始した最初のソーシャルブックマークサービスであり，そのユーザ数は現在最も多い．del.icio.usがCGM（Consumer Generated Media）として成功を収めたのを受けて，今日では多様なソーシャルブックマークサービスがサービスを開始し利用できるようになっている．そのユーザ層も，初期はWeb関連の技術に関心があるユーザが中心だったのに比べ，音楽やスポーツ，政治といった方面へ関心を持つユーザへの広がりもみせており，近い将来，より一般的なユーザから広く利用されると期待される．

⋆1 http://del.icio.us

基本的に個人で使用することを前提とした Web ブラウザのブックマークと違い，ソーシャルブックマークはユーザによるブックマークという行為が，他のユーザに伝播し，社会的な働きをするよう設計されている．その具体例として，多数のユーザによって最近ブックマークされたページを知らせる機能や，ほかのユーザのブックマークの更新情報を購読する機能などがある．これらの機能によって，有用なページは多くのユーザの注意を惹くこととなり，最終的に広く認知されるようになっている．

ソーシャルブックマークにおいて，ページの人気度はしばしば，そのページをブックマークしたユーザの数をもって測られる．これ以降，本稿ではこの数を SBRank 値と表記する．ここで SBRank と PageRank は，一種のページの有用性を測るための尺度といえるが，評価のされ方に大きな違いがある．直感的には，我々が Web のユーザを大まかにコンテンツ作成者とコンテンツ消費者に分けたとき，PageRank は「コンテンツ作成者によるコンテンツへの評価」といえる．一方で，SBRank は「コンテンツ消費者によるコンテンツへの評価」と見なすことができるだろう．コンテンツ作成者に要求されるスキルは，コンテンツ消費者に要求されるスキルに比べ高いことは明らかであり，結果として評価者の数もコンテンツ消費者の方が多くなる．

そこで本稿では，まず SBRank と PageRank の間の比較分析を行う．この調査の目的は 2 つの評価尺度を使用した複合型 Web 検索の可能性を模索することである．また，Google が検索結果のランキングに採用されているの尺度のうち，一般に公開されている PageRank と，それを補完する尺度としての SBRank を利用したランキングにより，検索結果のランキング結果を向上させる方法を提案し，さらに実験によってその評価を行う．さらに，どのようなクエリが SBRank を用いたサーチに有効なのかを調べるため，ホットキーワードとそれに基づく膨大な検索結果を分析する．さらに，実験によりその評価を行う．最後に，この研究分野の今後の課題について概観する．

2. 関 連 研 究

ソーシャルブックマークにまつわる研究の起源は，Keller らによる Web ブラウザのブックマークの機能を協調的なアプローチで向上させる取り組み[8] まで遡る．また最近では，Bry らが関連する研究[3] を行っている．

すでにいくつかの議論や分析が行われている[6),12),15),17)] ものの，ソーシャルブックマークはそれ自体が比較的新しい取り組みであることもあり，まだ十分に研究されているとはいえない．今までに行われた研究は主に，folksonomy という，情報を整理するための新しい分類方法に着目している[15),17)]．Lei らは HACM（Hierarchy-clustering model）を用いた folksonomy の階層的概念モデル[17] を提案し，その論文の中で，「ソーシャルブックマークで用いられるタグの間にある種の階層関係が認められた」と報告している．我々の研究と密接に関係するのは，Golder らによる，ソーシャルブックマークにおけるユーザ行動，タグの使用頻度などの規則性に関する報告[6] である．彼らはユーザが使用するタグの種類に着目し，タグを機能に応じて 7 つのカテゴリに分類した．この分析は del.icio.us のシステムに対して行われた．また，最近は Marlow らが，従来の階層型分類に対するタグによる分類の利点についての議論[12] をしている．Wu らは，ソーシャルブックマークを例に，アノテーションが付加された Web 上のリソースに対する意味的な検索モデル[16] を提案した．しかしこれまでの研究では，リンク構造とソーシャルブックマークの尺度の比較分析も，それらを組み合わせるという検討もなされていない．

メタサーチエンジン[5),10),11),13)] は，我々の取り組みと密接に関係している．Web 上では今までに多くのメタサーチエンジンが実用化されてきた．メタサーチエンジンは，複数の情報ソースを利用することによって Web 全体に対する再現率を向上させるとともに，より新しい情報を提供可能とするものである．しかし，リンク構造とソーシャルブックマークからなる統合ランキング尺度に関してはまだ取り組まれていない．これは Web のリンク解析によるランキングとソーシャルブックマークの情報が異なる特性を持つことと，それらの比較的な分析がなされていないことが理由だと思われる．

本稿は，SBRank を Google が検索結果のランキングに使用する尺度の一部である PageRank と統合することの可能性を綿密に調査し，この隔たりを埋め，Web 検索に応用することを検討するものである．

3. 比 較 分 析

SBRank と PageRank を統合するランキング手法の検討にあたり，各手法の実態について分析を行う．まず SBRank と PageRank の分布を中心に，それら単独での基本特性について分析する．また両尺度の相関関係について調べることで，補完の可能性について議論する．そして時間的な特性についても分析し，特に即時性の面で SBRank が PageRank を補完する可能性について議論する．各分析についてその結果を示したうえで，分析結果のまとめを行う．

3.1 データセットの特性

ソーシャルブックマークでブックマークされているページの特性を分析するために，我々

は del.icio.us とはてなブックマーク[⋆1]からそれぞれデータセット A および B の収集を行った．ここで，ソーシャルブックマークでは，多数のユーザがブックマークする際に付加するタグ（tag）という，分類用の短いキーワードによって，ページの分類が行われる．ユーザはタグをクエリとし，そのタグで分類されたページを取得することができる．使用頻度の高いタグはある種の分類として利用することができるため，データセットの取得にあたり高頻出のタグを使用する．ユーザはタグをクエリとし，そのタグで分類されたページを取得することができる．ここで，データセットの取得の方法は大まかに以下のとおりである．

(1)　ソーシャルブックマークで頻出するタグリストを取得.
(2)　タグリスト中のタグを用いてブックマークされている URL のリストを取得し，重複を省く.
(3)　各 URL について SBRank や PageRank などのページ情報を取得.

del.icio.us やはてなブックマークには，ユーザが最近使用する頻度が高いタグの集合を取得する機能がある[⋆2,⋆3]．この機能を使用して，まず我々は 2006 年 12 月 6 日にデータセット A として del.icio.us から 135 種，2007 年 2 月 16 日にデータセット B としてはてなブックマークから 742 種類のタグの集合を取得した．一方，あるタグ A について，その時点で人気のあるページの URL もそれぞれ取得可能である[⋆4,⋆5]．そこで，高頻度で利用されているこれらのタグそれぞれについて，人気のあるページを過去 2 年分，重複を排除したうえで取得し，データセット A で 1,079 件，データセット B で 8,029 件の URL を得た．収集した URL についてそれぞれページ情報は [Tag，URL，firstDate，SBRank 値，PageRank 値] を取得した．ただし，firstDate はその URL がソーシャルブックマークで最初にブックマークされた日時を示す．SBRank 値は与えられた URL をブックマークしたユーザの数であり，PageRank 値は Google ツールバーにより提供されるものを収集した．

3.2　PageRank 値と SBRank 値の分布の特徴

図 1 および **図 2** は PageRank 値の分布である．ここで，データセット A では 67%のページが，またデータセット B では 81%のページが PageRank 値が 0 であることが判明した．Google が検索結果のランキングを決定する際の主な要因がページの PageRank 値である以上，これらのページは検索結果の下位に表示され，見つけ出すのが困難なことが多い．しかし興味深いことに，多くのソーシャルブックマークのユーザはこれらのページの品質が高いと判断し，ブックマークしている．このことは，通常の Web サーチエンジン以外の情報源，おそらくは個々のソーシャルブックマークのシステム内の相互作用によって発見されたと推測できる．

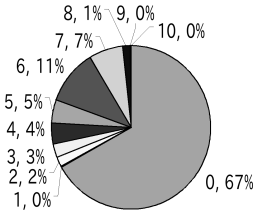

図 1　PageRank 値の分布（データセット A）
Fig. 1　Distribution of PageRank (Dataset A).

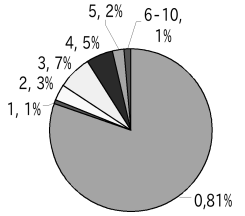

図 2　PageRank 値の分布（データセット B）
Fig. 2　Distribution of PageRank (Dataset B).

これらのページに現れる SBRank 値と PageRank 値とのギャップついては次のような原因があると考えられる．まず，PageRank 値は Web 全体のページ間のリンク関係が対象になるため，計算量が膨大となり，新たに作成されたページが PageRank によって評価されるまでに最大で 3 カ月かかる[18] といわれている．一方で SBRank 値は平均で 10 日でブックマーク数がピークに達するという報告[6] があり，ページが作成されてからそのページに対する評価が定まるまでにかかる時間が PageRank より短いという特長がある．

また Web サイト管理者がリンクを作成する行為とブックマークするという行為を比べた場合，ユーザへの負荷が大きく異なる．そのため，「存在は認められているが特に他のページからリンクされていない．しかしブックマークはされている」といったページは PageRank では正しく評価されない可能性がある．

⋆1 http://b.hatena.ne.jp
⋆2 http://b.hatena.ne.jp/t
⋆3 http://del.icio.us/tag/
⋆4 http://del.icio.us/A/popular
⋆5 http://b.hatena.ne.jp/t/A

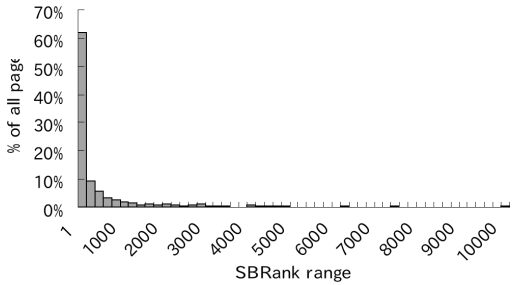

図 3　SBRank 値の分布（データセット A）
Fig. 3　Distribution of SBRank (Dataset A).

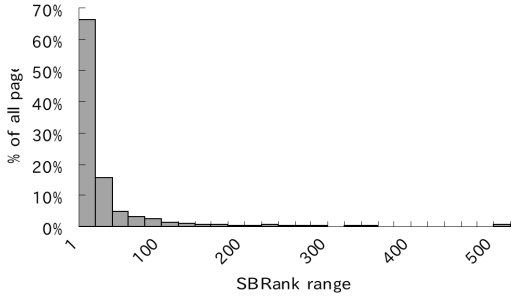

図 4　SBRank 値の分布（データセット B）
Fig. 4　Distribution of SBRank (Dataset B).

図 3 および **図 4** は SBRank 値の分布である．横軸は SBRank 値の区間，縦軸はデータセット中の割合を示している．

これらの図から，データセット A では 1 から 200 の範囲，データセット B では 1 から 20 の範囲に，両データセット中の過半数のページが収まっており，比較的大きな SBRank 値を持つのは少数のページのみであることが分かる．また，両データセットを比較すると，SBRank 値の絶対値は異なるものの，分布としては同様の傾向を持っていることが分かる．

なお，データセット A では SBRank 値の中央値が 144 であるのに対して平均は 1,115 であった．また，データセット B では SBRank 値の中央値が 80 であるのに対して平均は 736 であった．

データセット A の中央値が図 3 の見た目に対して意外に低い理由としては，図 3 をよく見ると 10 件以上 100 件未満のブックマーク数を持つページがデータセットに 50%近く含まれていることによる．また，平均値が意外に高いことについては，SBRank が 1,000 から 10,000 以上の少数のページの存在により，平均が大幅に押し上げられているのがその理由である．「検索エンジンにおいて上位にランキングされることが多いページは人の目に触れられることが多く，その結果としてさらに参照される機会が増加するため，その順位を保持し続ける傾向がある」という報告[4] が Cho らによってなされているが，本研究の調査より，ブックマーク行為にも同様の傾向があることが分かる．

図 5　firstDate の分布（データセット A）
Fig. 5　Histogram of firstDate (Dataset A).

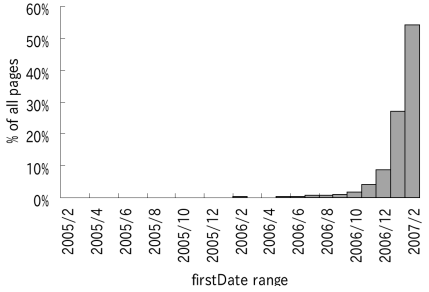

図 6　firstDate の分布（データセット B）
Fig. 6　Histogram of firstDate (Dataset B).

3.3　時間軸による分析

図 5 および **図 6** は各データセットについて各ページが初めてブックマークされた日付（firstDate）を古い順に並べたものである．これらの図から，約半分のページは最初に del.icio.us でブックマークされてから 3 カ月以内であり，約 9 割のページは最初にはてなブックマー

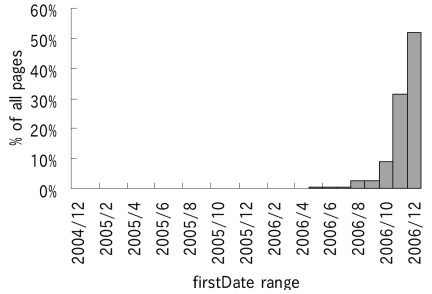

図 7　firstDate の分布（データセット A，PageRank = 0 のページのみ）
Fig. 7　Histogram of firstDate (Dataset A, PageRank = 0 pages only).

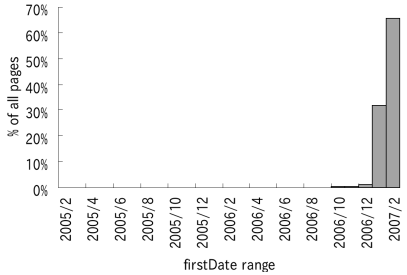

図 8　firstDate の分布（データセット B，PageRank = 0 のページのみ）
Fig. 8　Histogram of firstDate (Dataset B, PageRank = 0 pages only).

クでブックマークされてから 3 カ月以内，ということが分かる．このことから，ソーシャルブックマークのユーザは新鮮なページを好む傾向があると推測できる．この結果は，ソーシャルブックマークユーザの典型的な振舞いとして興味深い．

図 7 および **図 8** はデータセットのうち PageRank 値が 0 であるページに限ってプロットしたものである．図 7 は図 5 に比べ，プロットされている点が実験日から見て最近の日付に偏っている．図 8 も図 6 と比較すると同様の傾向がみてとれる．つまり，PageRank 値が 0 と評価されたページの多くは，最近初めてブックマークされたページでもあるといえる．Golder と Huberman による「del.icio.us で過去にブックマークされたページにおいて，その半数を超えるページは，最初にブックマークされた日から最初の 10 日以内に人気度がピークに達する」という報告[6] にもあるとおり，ソーシャルブックマークにおけるペー

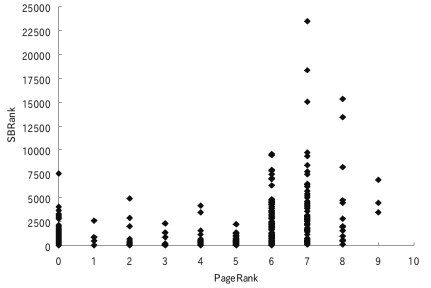

図 9　PageRank 値と SBRank 値の散布図（データセット A）
Fig. 9　Scatter plot between PageRank and SBRank (Dataset A).

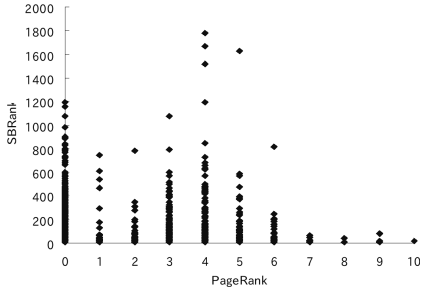

図 10　PageRank 値と SBRank 値の散布図（データセット B）
Fig. 10　Scatter plot between PageRank and SBRank (Dataset B).

ジの人気度は安定するまでの時間が短い．この事実から，SBRank は新しいページに対する PageRank の評価のバイアスを解消できる可能性が高い．そのため，リンク構造による重要度とブックマーク人気度の組み合わせるアプローチは検討に値すると考える．

3.4　PageRank 値と SBRank 値の相関

図 9 および **図 10** は，各ページの PageRank 値と SBRank 値をプロットしたものである．横軸が PageRank，縦軸が SBRank を示している．この図は，PageRank 値と SBRank 値の関係を提示している．図 9 からは，それぞれの PageRank 値で SBRank 値の値域が重なる部分が大きいことと，PageRank 値が大きくなるにつれて SBRank 値も大きくなる傾向が見てとれる．これと関連してデータセット中の SBRank 値と PageRank 値の相関係数を求めたところ，ある程度の正の相関（$r = 0.49$）を示した．

図 10（データセット B）には図 9 と違い「PageRank 値が大きくなるにつれて SBRank

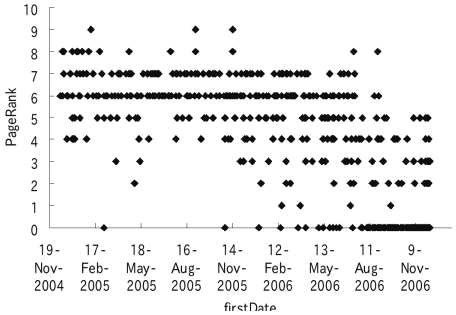

図 11　firstDate と PageRank 値の散布図（データセット A）
Fig. 11　Scatter plot between firstDate and PageRank (Datset A).

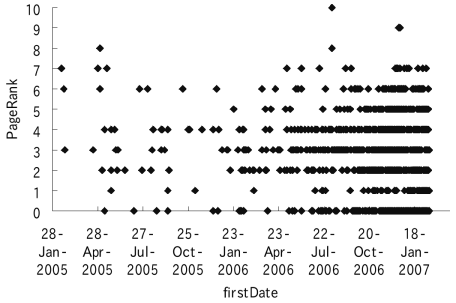

図 12　firstDate と PageRank 値の散布図（データセット B）
Fig. 12　Scatter plot between firstDate and PageRank (Dataset B).

値も大きくなる」という傾向は見られない．相関係数も $r = 0.10$ とそれほど相関があるわけではない．このような違いが見られた原因は，図 8 に示したとおり，データセット B はデータセット A に比べ比較的新しいページが占める割合が高く，結果として多くのページが PageRank による評価が安定する途中であるからだと考えられる．図 9 および図 10 を見ると PageRank が 0 のページを除けば，データセット A では分布のピークが PageRank が 7 の部分にあるのに対して，データセット B では PageRank が 4 の部分にある．時間が経過するとともに，データセット B のピークが右に移動し，データセット A のような分布になることが考えられる．

図 11 および **図 12** は各データセットにおいて，横軸に firstDate，縦軸に PageRank 値をとった散布図である．これらの図は，ページが最初にブックマークされた時期と，そのページの PageRank 値との関係を提示している．

図 11 からは，最初にブックマークされてから時間が経過しているページは PageRank 値が高く，初めてブックマークされてから日が浅いページは PageRank 値が低いという傾向が見てとれる．しかし，図 12 からは図 11 のような明確な傾向は見てとれない．この原因としては，これは図 6 が示すとおり，データセット B として収集されたページは PageRank が低いものがほとんどであり，さらに図 6 で確認できるとおり古いページはほとんど含まれていないことによる．ここで firstDate と PageRank 値の間の相関係数を調べたところ，どちらも負の相関（データセット A：$r = -0.85$，データセット B：$r = -0.51$）を観測した．これは Baeza-Yates らの「Web サイトが公開されてから経過した時間の長さと PageRank 値との間には高い相関がある」という報告[1] に合致するものである．本節の分析結果から，SBRank は PageRank に比べて，ページに対する一定の評価が定まるのに必要な時間が短いという点で優位であることが分かった．また，PageRank を SBRank で補完することは，PageRank の精度だけでなくその即時性も強化できる可能性があるといえる．

3.5　比較分析のまとめ

ここでは，上記の比較分析によって得られた結果と，それに基づく我々の推測について列挙する．

- ソーシャルブックマークのユーザは検索エンジン以外の情報源からブックマークしたページを発見している可能性が認められたことから，ソーシャルブックマークには，検索エンジンでは発見するのが困難であるものの，重要なページが含まれている可能性がある．
- SBRank 値と PageRank 値の間にある程度の正の相関が認められたことから，SBRank が PageRank を補完することで Web 検索におけるランキング精度を向上させることができる可能性を秘めていると考えられる．
- firstDate の日付が古いほど PageRank 値が高いという関係が認められたことから，SBRank は PageRank の即時性を強化できる可能性があると考えられる．

4.　SBRank と PageRank の統合検索

本章では，両ランキング尺度を組み合わせた統合型 Web 検索手法の有効性を調査する．まず，統合したランキング尺度を求めるにあたって次の式を利用する．

$$NewRank = \alpha \cdot SBRank + (1-\alpha) \cdot PageRank \quad (0 \leq \alpha \leq 1) \tag{1}$$

ここで，α は $SBRank$ と $PageRank$ の最適な重みを決定するためのパラメータである．

4.1　実際の検索結果における提案手法の振舞い

提案手法を適用するには，既存の検索エンジンにおける検索結果において，ある程度のページがソーシャルブックマークにおいてブックマークされている必要がある．また再ランキングの際にはブックマーク数と検索結果の順位に乖離がある必要がある．そこで提案手法を評価するに先だち，これらの特性について，異なる特徴を持つ 2 つのクエリ集合を用いて比較調査を行った．通常のクエリ集合として，Web 検索エンジンの goo[⋆1]が公開している，“goo において 1 カ月単位に使用頻度の急上昇したキーワードの上位 50 件[⋆2]” を 2006 年 1 月分から 2007 年 10 月分まで，重複を取り除いたうえで計 806 件取得した．**図 13** は両クエリ集合から頭文字 1 つにつき 1 クエリずつサンプリングしたものである．またもう 1 つのクエリ集合として，“はてなブックマークで最近頻繁に利用されているタグ[⋆3]” を 531 件取得した．goo キーワードは goo を利用している一般的なユーザの興味を表しているといえ，人の名前やイベント名，番組の名前などの固有名詞が目立つ．一方で，はてなブックマークで使用されているタグは，はてなブックマークを使用しているユーザの中心的な興味，多くは技術的な話題を表しているといえ，どちらかというと一般名詞が目立つ．

次に，2 つのクエリ集合において Yahoo!の検索結果を検索クエリごとに 500 件ずつ取得し，さらに検索結果の各ページについて，はてなブックマークにおけるブックマーク数および PageRank 値を取得した．以降はこの 2 つのデータセットの特性と，それについての考察を述べる．

図 14 は各クエリの検索結果上位 500 件それぞれにおける PageRank 値の平均である．どちらも上位ほど高い PageRank 値を持つ傾向を示しており，Google の PageRank によるページ評価尺度は Yahoo!の検索結果とも相関があることが見てとれる．

図 15 は Yahoo!で Web 検索を行った際に，検索結果の上位 500 件のうち，ブックマークされているページの数の分布を示している．横軸が検索結果の上位 500 件のうちブックマークを持つページ数の区間を 10 ページきざみにとったもの，縦軸が各区間に該当する検索クエリの，クエリ集合全体における割合を示したものである．図 15 によると，goo キーワードでは多くのクエリにおいてブックマークされているページは，500 件中 20 件から 50 件，つまり 5%から 10%程度と比較的狭い範囲に収まっているのが分かる．それに対して，はてなタグではブックマークされているページが 40 ページから 140 ページ，つまり 10%から 30%に比較的幅広く分布していることが分かる．検索結果に対して被ブックマークページが多い状態，つまり図 15 において分布が右側に偏るほど，提案手法によって検索結果下位のページが上位に再ランクされる可能性が増加することになり好ましい．2 つのクエリ集

	クエリの例
goo	14才の母，AAA，bird，DADA FLORA，F-1GP，Google Earth，HERO，IKEA，JRA，KANON，LEXUS，MHF，NANA，nanaco，NEWS，ONE PIECE 46巻，PASMO，R-1グランプリ，SEO，TOTO，USEN，WBC(ワールド・ベースボール・クラシック)，Xbox360，YouTube，あいのり，イチロー，ウオッカ，エウレカセブン，おいしいプロポーズ，かもめーる，きっこのブログ，くまぇり(平田恵里香)，ゲゲゲの鬼太郎，こうちゃん レシピ，サイレン，しそ酢，スキー場，セクシーボイスアンドロボ，そのまんま東，ダイエット，ちゃお，ディズニー，トイザらス，なかやまきんに君，ニコニコ動画，ねこやん，のじぎく兵庫国体，はいだしょうこ，ひぐらしのなく頃に，ビックカメラ，ファースト・キス，ペルセウス座流星群，ポケットモンスター，マイドコモ，みうらじゅん，メタンガス，もってけ！セーラーふく，やくみつる，ユニクロ，ヨドバシカメラ，ライアーゲーム，リア・ディゾン，レミオロメン，ロト6，ワザップ，エヴァンゲリヲン新劇場版:序，阿部典史，伊勢丹，宇多田ヒカル，英検，押切もえ，仮面ライダー電王，偽りの花園，熊田曜子，芸能ニュース，古瀬絵理，戸田恵梨香，佐賀北高校，四元奈生美，水うちわ，世界バレー，倉田麗華，体内メーカー，地震，辻希美 結婚，鉄道博物館，杜仲茶，内田有紀，日テレ，熱中症，納豆ダイエット，梅雨明け，美しい罠，不信のとき，平井理央，母の日，魔法先生ネギま!，民主党，無双OROCHI，無料ゲーム，冥王星，木村カエラ，夜王，優香，浴衣，嵐，里田まい，恋空，六星占術，和田義彦
はてな	2007，.net，*css，2ch，aa，bicycle，c，dankogai，eclipse，acebook，gadget，hatena，ibm，japan，language，mac，nagios，office，pc，rails，sbm，tech，ui，video，web，xampp，yahoo，アイコン，いい話，エロ，オタク，カスタマイズ，キャッシング，クリスマス，ケータイ，ことば，サイト，ジェネレータ，スポーツ，セキュリティ，ソフト，ダイエット，チュートリアル，ツール，テクスチャ，トイレ，ニコニコ，ネタ，パソコン，ビジネス，ファッション，ペット，ポイント，マジック，メディア，モバイル，ユーザビリティ，ライブ，リファレンス，レイアウト，ロゴ，医療，宇宙，映画，音楽，科学，企画，軍事，携帯，言葉，裁判，仕事，手作り，心理，図書館，政治，素材，大学，知識，痛いニュース(ノ∀`)，鉄道，投資，日記，猫，配色，比較，不動産，便利，萌え，漫画，無料，野球，欲しい，旅行，歴史，労働，和風，京都

図 13　両検索クエリ集合のクエリの例
Fig. 13　Example of search query in query sets.

⋆1 http://www.goo.ne.jp
⋆2 http://ranking.goo.ne.jp/service/001/
⋆3 http://b.hatena.ne.jp/t

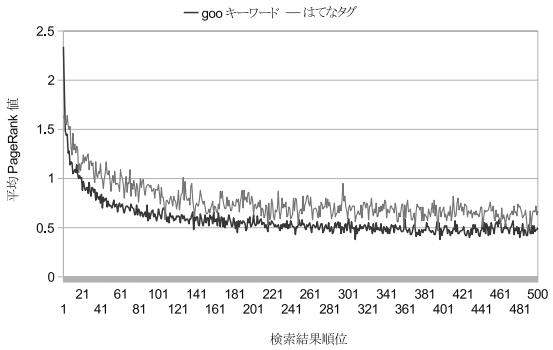

図 14　クエリ集合における Yahoo!検索結果順位の平均 PageRank 値
Fig. 14　Average PageRank values on top 50 Yahoo! search results.

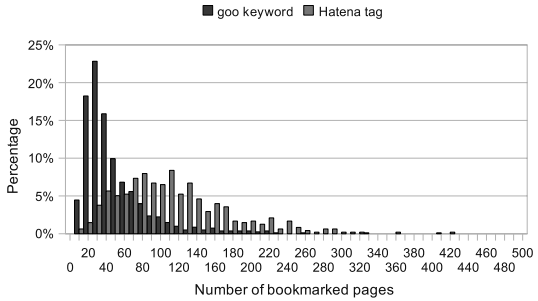

図 15　クエリ集合におけるブックマーク済みページ数の分布
Fig. 15　Distribution of number of bookmarked pages for each query sets.

合を比較すると，本手法ははてなタグに基づいたものの方が適していることが見てとれる．

次に，提案手法で $\alpha = 1$ を設定した状態，つまりページのブックマーク数が大きい順に再ランキングを行ったときの，元の検索結果順位からの変化の度合いを，Spearman 順位相関係数を用いて評価した．Spearman 順位相関係数とは，2 つの順位データの相関関係を調べるための指標であり，2 つの順位データが完全に同一である場合に 1，正反対である場合に -1 をとる．**図 16** は各クエリにおいて提案手法で再ランキングを行った際の Spearman 相関順位係数の分布である．図 15 で示したとおり，検索結果があまりブックマークされていないような検索クエリが存在する．このような検索クエリは，一部のページのランキング

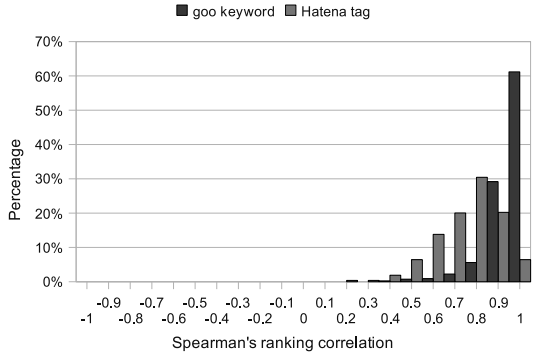

図 16　提案手法を適用した際の Spearman 順位相関係数の分布
Fig. 16　Distribution of Spearman correlation after applied our reranking method for query sets.

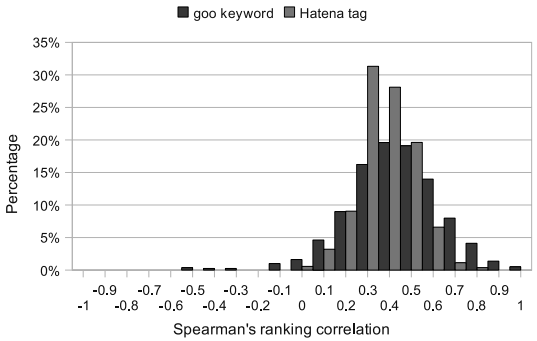

図 17　提案手法を適用した際の Spearman 順位相関係数の分布（ブックマーク済みページのみ）
Fig. 17　Distribution of Spearman correlation after applied our reranking method for query sets (bookmarked pages only).

が大きく変動するものの，ほとんどのページは元の順位からの相対的位置は変化せず，結果としてグラフから提案手法に起因するページ間の順位の入れ替わりが読み取りにくくなってしまっている．そこで，Spearman 相関順位係数を算出する際に，検索結果中のブックマーク済みのページのみを対象にしたものが **図 17** である．

図 17 から，ランキングの順位変動の分布は，両クエリ集合とも 0.4 を中心として同じような傾向があることが分かる．つまりブックマークされているページのみを対象にした場合，元のランキングとは正反対になるような大幅な順位の変動するクエリや，まったく順位

	Spearman係数	
	低い	高い
goo	ブログ, 検索, 株, NEWS, 旅行, SEO, SEO対策, ライフ, 沖縄, サッカー, 日本語, ライブカメラ, mixi, YOUTUBE, YouTube, ミッチェルの「中小企業診断士のお勉強」ブログ, ニコニコ動画, Youtube, Google Earth, iPod, 民主党, Xbox360, 任天堂, Winny, Wii, PS3, Windows Vista, ソフトバンク, PSP, ニンテンドーDS, チュートリアル	梅雨明け, 狂犬病, 寒中見舞い, 安全標語, 飯島茜, 都市対抗野球, MINMI, チャーガ, 日本女子プロゴルフ, 探偵学園Q, ナルト, ファースト・キス, 河合塾, 表久禎, 三宅恵美, 統一地方選挙, 有馬記念, 中江里香, パロマ, 大山千穂, どろろ, 倉田麗華, ファーベストファイバー, 叶晴栄, 柿崎順一, 若宮優子, 春のワルツ, マイドコモ, DADA, FLORA, 滝浪愛
はてな	サイト, *web, api, web, japan, wiki, ブログ, blog, インターネット, 2007, ネット, ダウンロード, 日本, *css, css, ゲーム, ブックマーク, 検索エンジン, 無料, flash, 会社, 情報, ruby, rails, google, はてなブックマーク, はてブ, twitter, gmail, ニコニコ動画	創作, economy, mad, asia, 司法, copyright, event, sex, education, いい話, webdesign, branding, literature, おやつ, キャッシング, 年金, ベトナム株口座開設, ベトナムファンド, scalability, brush, amf, ベトナム株, 男女, politics, food, military, ブラシ, pipes, はてなハイク, mtos

図 18　検索クエリの分類
Fig. 18　Search query categorization.

変動しないようなクエリはごく一部であり，多くの検索クエリではある程度ランキングが入れ替わるといえる．

4.2　Spearman 相関順位係数の値別の特徴

次に，Spearman 順位相関係数が低いクエリと高い検索クエリがそれぞれどのような特徴を持っているのかを調査した．Broder [2] による検索質問の分類に基づくと，検索クエリ集合はユーザの意図に応じて調査型の検索質問（informational query）と誘導型の検索質問（navigational query）に大きく分けることができる．ここで両クエリ集合において，Spearman 順位相関係数が最も低かったグループと最も高かったグループ，つまり図 16 の分布の両端部分に位置する検索クエリをそれぞれ 30 件ずつ抽出したものが，**図 18** である．以降は図 18 のそれぞれの領域の特徴について考察を述べる．

まず，goo のクエリ集合に着目すると，Spearman 順位相関係数が低いグループでは，高いグループに比べて informational なクエリの割合が高くなる．具体的には，ブログ，検索，株，旅行，SEO，沖縄，サッカー，ライブカメラなどである．また，Spearman 順位相関係数が高いグループで多く見られる人名やイベント名が含まれないという特徴がある．次にはてなのクエリ集合に着目しても同様に，Spearman 順位相関係数が低いグループにおいて informational なクエリの割合が高いことが分かる．

以上をまとめると，Spearman 順位相関係数が低いグループ，すなわち提案手法が強く作用する検索クエリ集合は，同係数が高いグループに比べ informational なクエリの割合が高く，navigational なクエリの割合が低いという特徴が見出せる．つまり我々のシステムは informational なクエリにおいて効果を発揮できる可能性を秘めているといえる．

4.3　適合率および再現率による評価

実際のランキングに関する実験では，任意の検索クエリに対する Google の検索結果に対して，個々のページの PageRank 値とデータセット A（del.icio.us）における SBRank 値を基にして行った．具体的には，あるクエリに基づく検索結果の個々のページにおいて，SBRank 値と PageRank 値に対して式 (1) を適用し，得られた値が大きい順に並べたものが新たな検索結果のランキングである．なお，正解集合は手動で各ページにアクセスして，そのページが検索クエリと関連性が強いかを，我々が主観で判定した．また，その作業量的な問題から，正解集合は Google の検索結果の上位 50 件にすべて含まれる，検索クエリと関連性のあるページと仮定した．

式 (1) を適用するに先立って，SBRank 値と PageRank 値は正規化する必要がある．そこで，各ページの SBRank 値と PageRank 値は，それぞれ検索結果における最も大きい値で除算し，最大値が 1 になるように正規化する．なお，Google が検索結果のランキングを決定する際には，PageRank 以外にもいろいろな尺度を用いていると予想されるが，本稿では評価のために，ランキングに影響するのは PageRank のみであると見なし，単純化する．

この手法を評価するために我々は “graphic design”，“gardening”，“java”，“apple”，“computer vision” および “stamp collecting” という 6 つの検索クエリを用意した．これらの検索クエリは主に文献 7) で用いられているものから抜粋したものであり，一般的に informational なクエリとして利用されるものである．次に，それぞれの検索クエリについて，Google 検索エンジンから上位 50 件の URL を取得した．

次に，各クエリに対する検索結果に式 (1) を適用し，α の値を変化させたときに検索結果

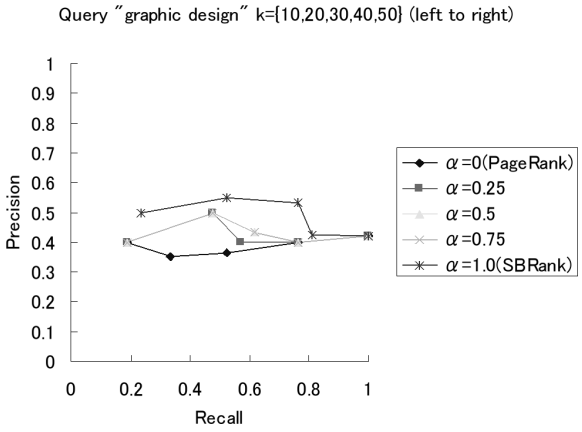

図 19　クエリ “graphic design” における再現率–適合率曲線
Fig. 19　Recall-Precision curve for query “grapahic design”.

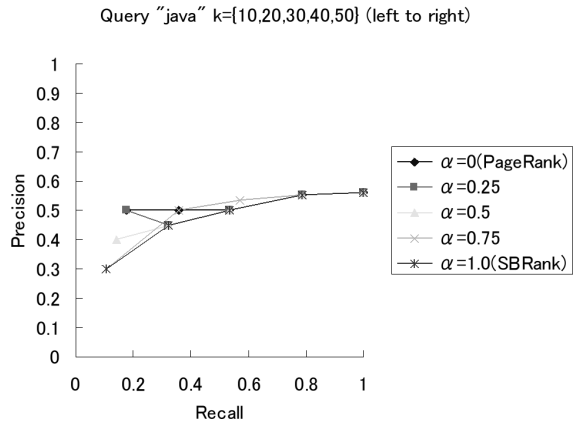

図 20　クエリ “java” における再現率–適合率曲線
Fig. 20　Recall-Precision curve for query “java”.

上位 k 件（$k = 10$，20，30，40，50）において再現率と適合率がどのように変化するのかを調査した（**図 19** および **図 20**）．

ここでいう再現率とは，再ランキングされた検索結果の上位 n 件に正解ページが何%含まれているかを示す尺度である．また，適合率とは，再ランキングが行われた後の上位 n 件の検索結果のうち，正解ページが占める割合を指す．

$k = 50$ ですべての再現率–適合率曲線の再現率（Recall）が 1 に収束しているのは，前述のとおり上位 50 件に正解集合がすべて含まれると仮定したためである．

本来は，あるクエリに対してページ間のリンク関係を利用しないアルゴリズムに基づいた検索結果をランキングの元データとし，その検索結果に対して SBRank と PageRank を利用してランキングを行うのが理想的な評価方法であるが，そのような検索結果を用意するのは現段階では難しいと判断したため，現状の条件設定・評価方法を採用した．今回の実験によって少なくとも「現在主流となっているリンクベースのランキングに，ソーシャルブックマークにおけるユーザベースの社会的受容度というもう 1 つの軸を段階的に加えた場合の，検索精度の変化」は，評価できていると考える．

観測した結果のうち最も高い再現率と適合率を示したのは，クエリ “graphic design” において α を 1 に設定した場合であった（図 19）．このクエリでは，$k = 20$，30 においては，α の値に比例して再現率と適合率が高くなるという特性を示した．

一方，“java” というクエリでは，特に $k = 10$ のときに低い再現率と適合率を観測した（図 20）．この結果は，検索結果は Java ではなく JavaScript に関するページが相当数含まれていたことによる．PageRank はページに対するクエリ独立な評価尺度であると同時に，SBRank も同様にクエリ独立な評価尺度であるため，このような問題には無力であるが，将来的にはタグを用いてページのトピックを自動的に抽出することで，このような問題に対処することが考えられる．

なお，適合率–再現率曲線には表れなかった現象として，リランキングによるのページの順位変動があげられる．クエリ “java” は特にそれが顕著であり，$\alpha = 1$（SBRank）に設定して再ランキングを行った場合の上位 10 件の元の順位は {42，3，24，25，14，9，36，18，6，34} であった．また，検索結果における SBRank の最大値はクエリごとにかなりばらつきがあり，最も大きかったのがクエリ “java” で 1,823，最も小さかったのがクエリ “stamp collecting” で 27 であった．このように，SBRank はページで扱われている内容によって偏りが大きいという特性がある．これは，現在ソーシャルブックマークを利用しているユーザ層の興味の偏りが原因であると考えられる．

図 21 はすべてのクエリの平均をとったものである．この図を見ると α の値にかかわらず重なっている部分が多く見られることから，本手法は全体的な適合率や再現率を底上げするものではないといえる．しかし，クエリ “graphic design” の例や，4 章で述べたとおり，探索型クエリでなおかつ技術的なトピックに関しては，ある程度有効であることが推測さ

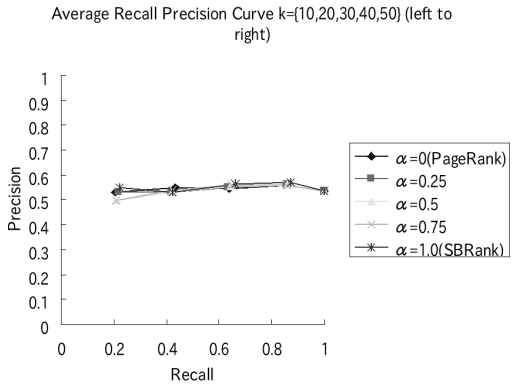

図 21　クエリ “graphic design”，“gardening”，“java”，“apple”，“computer vision” および “stamp collecting” における平均再現率–適合率曲線
Fig. 21　Average Recall-Precision curve for query “graphic design”, “gardening”, “java”, “apple”, “computer vision” and “stamp collecting”.

れる．

現在ソーシャルブックマークを利用しているユーザ層は主に技術的なことに興味がある人々であり，ブックマークされるページも技術的なものに偏っているのが現状である．そのためクエリによってはほとんどブックマークされていないこともあり，そのような場合は効果的にランキングを行うことができず，結果として適合率・再現率が低くなるという現象が見られた．しかし，将来的にソーシャルブックマークがいろいろなユーザによって利用されることが予想されることから，将来的にはこのような統合型ランキングのアプローチの適用可能範囲は拡大していくと思われる．

今回の実験では，ページに関連付けられたタグは考慮しなかった．この情報を考慮することは今後の課題である．Haveliwala による Topic-Sensitive PageRank[7] は，ページの重要性をトピックごとに測定することで通常の PageRank よりも良い性能を出すことに成功している．そして，これに類似したアプローチはソーシャルブックマークにも適用できると考える．このアプローチの技術的に困難な点は，元となるトピックをどこから取得するか，ということにある．Topic-Sensitive PageRank ではページのトピックを決定するために，Open Directory Project（ODP）[⋆1]という，ボランティアベースの Web ディレクトリサービスを間接的に利用している．これと同じように，ODP の代わりにソーシャルブックマークのタグを使うことは考えられる．より適切な SBRank 値の算出には，ブックマークユーザと彼らがブックマークしたページの二項関係から重要なページとユーザを導くような，HITS[9] 的なアプローチが有効であると考えられる．

⋆1 http://dmoz.org/

5. ま と め

ソーシャルブックマークは Web 2.0 の基盤の 1 つである．既存の Web 検索を補完するためにソーシャルレコメンデーションを利用するアプローチは，ページの品質が実際の人間のブックマークという行為を通して担保されていることから，信頼性を増すものだと考えられる．本稿では，リンク構造分析によるランキングとソーシャルブックマークを統合することの可能性について分析および実験を行い，統合により性能が向上することを確かめた．

SBRank と PageRank の比較分析の結果，両者はある程度の相関関係を持つことが判明した．そして，SBRank は PageRank よりも即時性に優れている可能性があることが分かった．また，PageRank 値の時間的な側面に言及する過去の研究[1),4)] の追認をすることができた．以上のことから，統合型 Web 検索は有効である可能性が高いという考察が得られた．また，統合型 Web 検索の振舞いについて goo およびはてなブックマークのホットキーワードを用いて検索結果をソースとして調査することにより，informational なクエリの方が navigational なクエリに比べ，有効に働くであろう可能性を示した．さらに，単純な SBRank 値と PageRank 値を重み付けしたうえで合成するという実験を通じて，SBRank と PageRank を統合することで検索結果のランキング精度が多少向上することを確認した．我々は今後，異なるデータセットを使った大規模な実験を行うことを計画している．また，検索結果のより効果的なランキング修正のために，より複雑なアルゴリズムについて検討する予定である．具体的には，本稿では PageRank と SBRank の統合方法に単純な重み付けパラメータを利用したが，このランキングに確率モデルを導入することでより多くのクエリにおいてページのランキング精度を向上させるアプローチが考えられる．

謝辞　本研究の一部は，文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」，異メディア・アーカイブの横断的検索・統合ソフトウェア開発（研究代表者：田中克己），文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」（研究代表者：田中克己，A01-00-02，課題番号 18049041），文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研

究」，計画研究「情報爆発に対応する新 IT 基盤研究支援プラットフォームの構築」（研究代表者：安達淳，Y00-01，課題番号：18049073），および文部科学省科学研究費補助金若手研究（B）「情報検索とウェブアーカイブにおけるマイニング」（研究代表者：Adam Jatowt，課題番号：18700111）によるものです．ここに記して謝意を表します．

参 考 文 献

1) Baeza-Yates, R., Castillo, C. and Saint-Jean, F.: Web Dynamics, Structure and Page Quality, Web Dynamics, Levene, M. and Poulovassilis, A. (Eds.), pp.93–109, Springer (2004).
2) Broder, A.: A taxonomy of web search, *ACM SIGIR Forum*, Vol.36, No.2, p.3 (2002).
3) Bry, F. and Wagner, H.: Collaborative Categorization on the Web: Approach, Prototype and Experience Report, Forschungsbericht/research report (2003).
4) Cho, J., Roy, S. and Adams, R.: Page Quality: In Search of an Unbiased Web Ranking, *Proc. SIGMOD Conference 2005*, pp.551–562 (2005).
5) Dwork, C., Kumar, R., Naor, N. and Sivakumar, D.: Rank Aggregation Methods for the Web, *Proc. 10th World Wide Conference 2001* (2001).
6) Golder, S.A. and Huberman, B.A.: The Structure of Collaborative Tagging Systems, *Journal of Information Science* (2006).
7) Haveliwala, T.H.: Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search, *IEEE Trans. Knowledge and Data Engineering* (2003).
8) Keller, R.M., Wolfe, R.R., Chen, J.R., Labinowitz, J.L. and Mathe, N.: A Bookmarking Service for Organizing and Sharing URLs, *Proc. 6th Intl. WWW Conference*, Santa Clara, CA (1997).
9) Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment, *J. ACM* (1999).
10) Lawrence, S. and Giles, C.L.: Inquirus, the NECI meta search engine, *Proc. 7th International World Wide Web Conference*, Brisbane, Australia, pp.95–105 (1998).
11) Lu, Y., Meng, W., Shu, L., Yu, C. and Liu, K.: Evaluation of Result Merging Strategies for Metasearch Engines, *Proc. VLDB'05*, pp.141–150 (2006).
12) Marlow, C., Naaman, M., Boyd, D. and Davis, M.: HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read, *Proc. ACM HyperText 2006 Conference* (2006).
13) Meng, W., Yu, C. and Liu, K.-L.: Building efficient and effective metasearch engines, *ACM Computing Surveys*, Vol.34, No.1, pp.48–84 (2002).
14) Page, L., Brin, S., Motwani, R. and Winograd, T.: The pagerank citation ranking: Bringing order to the Web, Technical report, Stanford Digital Library Technologies Project (1998).
15) Strutz, D.N.: Communal Categorization: The Folksonomy, INFO622: Content Representation (2004).
16) Wu, X., Zhang, L. and Yu, Y.: Exploring Social Annotations for the Semantic Web, World Wide Web Conference 2006 (2006).
17) Zhang, L., Wu, X. and Yu, Y.: Emergent Semantics from Folksonomies: A Quantitative Study, *Journal on Data Semantics*, VI, LNCS 4090, pp.168–186 (2006).
18) PageRank — Wikipedia, the free encyclopedia. http://en.wikipedia.org/PageRank

(平成 19 年 12 月 20 日受付)
(平成 20 年 4 月 10 日採録)

(担当編集委員　井上 潮)

山家　雄介（学生会員）
京都大学大学院情報学研究科博士前期課程在学中．2006 年宮城大学事業構想学部デザイン情報学科卒業．ソーシャルブックマークとメタサーチに関する研究・開発に従事．電子情報通信学会，日本データベース学会各学生会員．

中村　聡史（正会員）
京都大学大学院情報学研究科社会情報学専攻特任助教．2004 年大阪大学大学院情報学研究科博士後期課程修了．博士（工学）．主にヒューマンコンピュータインタラクション，ウェブ検索の研究に従事．日本データベース学会会員．

アダム　ヤトフト（正会員）

京都大学大学院情報学研究科社会情報学専攻特任助教．2005 年東京大学大学院情報理工学系研究科電子情報学博士後期課程修了．博士（情報学）．主にウェブ検索，ウェブアーカイブマイニングの研究に従事．ACM 会員．

田中　克己（正会員）

京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院博士前期課程修了．博士（工学）．主にデータベース，マルチメディアコンテンツ処理，ウェブ検索の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，日本データベース学会各会員．