

ソーシャルブックマークの特性を利用した Web 検索のランキング精度の向上

Utilizing Social Bookmark Characteristics to Enhance Ranking in Web Search

山家 雄介[†] 中村 聡史[‡]
 アダム ヤトフト[‡] 田中 克己[‡]

Yusuke YANBE Satoshi NAKAMURA
 Adam JATOWT Katsumi TANAKA

ソーシャルブックマークは、Web 利用者がブラウジング中に興味や関心を持ったページを記録、共有、分類そして発見することを支援する Web サービスである。ソーシャルブックマークにおいて、あるページをブックマークしたユーザの数は、しばしばそのページの人気度として利用される。一方、PageRankによるPageRankは、コンテンツ製作者が記述したページ間のリンク構造をマクロ的に分析することで、各ページの重要度を算出する。本稿ではこれら2つの尺度を統合した、Web 検索の精度を向上させるための新しいランキング手法を提案し、また実験によって試験的な評価を行う。

Social bookmarking is an emergent Web service that helps users to record, share, classify and discover interesting Web pages. In social bookmarking systems, the popularity of each Web page is usually calculated by a total number of its bookmarks, that is, by the sum of explicit evaluations made by users. On the other hand, PageRank measures the attention and interest of authors in pages using the link structure of the Web. This paper describes, demonstrates, and evaluates our method for hybrid Web search that combines both ranking measures.

1. はじめに

90年代後半にPageRank[9]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登場した。PageRank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るが、その際、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。これは暗黙的に、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90年代のWebの環境に上手く適合しており、GoogleのWeb検索エンジンとしての成功もあって、PageRankはページの人気を推定するにあたって、最も有名な手法となった。

近年、blogやwikiなどのようにテンプレートから自動生成されるWebサイトの数が飛躍的に増加している。こうしたシステムはサイトやページ間をその品質に関わらず、自動的に

密に結び付けている。つまり、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言いがたく、PageRankが上手く働く前提に合致しない。さらに、spam trackbackや、splogなどのように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って、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。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のリンク構造をもと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きている。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。本論文ではこの被ブックマーク数をWebにおける一種の社会的受容度とみなし、SBRank値と表記する。ここでPageRankとSBRank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。直観的には、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(閲覧者)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」と見なすことが出来るだろう。

ソーシャルブックマークにまつわる研究はすでにいくつかの議論や分析が行われている[4][7][8][10][11][12]ものの、ソーシャルブックマーク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言いがたい。これまでの研究では主に、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[8][10][11][12]。一方で、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や、それらを組み合わせるといった検討はなされていない。本論文は、PageRankとSBRankを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検索のランキング精度向上のために応用することを検討するものである。

そこで、まずPageRankとSBRank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、SBRankとPageRankを統合したランキングにより、元の検索結果のランキング結果を向上させる方法を提案し、予備実験によってその評価を行う。

2. 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。さらに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、それらの分析結果のまとめを行う。

2.1 データセットの特性

我々は分析を行うためのデータセットの拠り所としてdel.icio.us[3]を利用した。del.icio.usは2006年12月時点で最もユーザ数が多いソーシャルブックマークである。なお、データの収集は、2006年12月6日に実施した。以下にデータ取得の手順を示す。

del.icio.usではPopular tagsという、ユーザが最近使用する頻度が高いタグの集合を(<http://del.icio.us/tag/>)から取得可能である。この機能を使用して、まず我々は135

[†] 学生会員 京都大学大学院 情報学研究所 博士前期課程
yanbe@dl.kuis.kyoto-u.ac.jp

[‡] 正会員 京都大学大学院 情報学研究所 社会情報学専攻
{nakamura,adam,tanaka}@dl.kuis.kyoto-u.ac.jp

種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(最大 20 件)の URL は (<http://del.icio.us/A/popular>) から取得することができる。そこで、高頻度で利用されているこの 135 種のタグそれぞれについて、人気のあるページを取得し、合計 2,673 件のデータを収集した。収集したページ情報はそれぞれ [Tag, URL, firstDate, SBRank 値] という属性をもつ。ただし、firstDate はその URL が del.icio.us で最初にブックマークされた日時を示す。また、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る。del.icio.us の仕様上、あるページが 2 つ以上のタグにおいて人気のあるページと判定される場合がある。そこで、同じ URL はひとつを残して削除することで重複を解消し、最終的に 1,290 のユニークな URL と、それぞれに対応する属性をデータセットとして得た。

次に、我々は Google 社の提供する Google ツールバーを利用することで各 URL の PageRank 値を取得した。なおこれは Web ブラウザのインタフェースを拡張するものであり、現在アクセスしているページの PageRank 値を表示する機能をもつ。

2.2 PageRank 値および SBRank 値の分布の特徴

図 1 はデータセットにおける PageRank 値の分布である。調査の結果、半数を超える (56.1%) ページについて 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である。しかし興味深いことに、多くの del.icio.us ユーザが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。このことより、こうしたページは通常の Web 検索エンジン以外の情報源、おそらくは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人気のコンテンツとなったと推測できる。

これらのページに現れる SBRank 値と PageRank 値とのギャップについてはおよそ 3 つの原因があると考えられる。

- ページが作成されてから日が浅く、まだ PageRank 値が与えられていない
- ページが比較的最近作成されたため、少数の被リンクしか持たない
- ページが作成されてから十分に時間が経過しているにもかかわらず、PageRank の手法では評価されていない

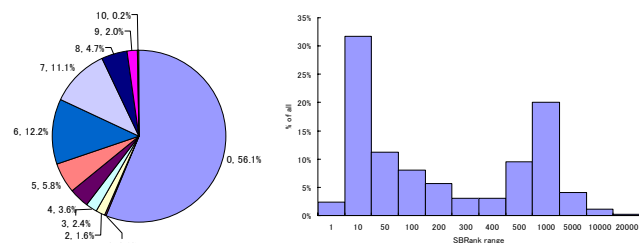


図 1 PageRank 値の分布

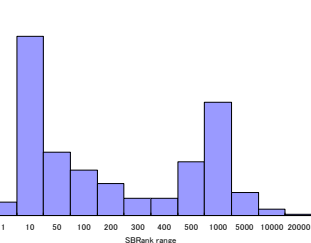


図 2 SBRank 値の分布

Fig. 1 Distribution of PageRank values

Fig. 2 Histogram of SBRank

図 2 は SBRank 値、つまりデータセット中の各ページにおける被ブックマーク数の分布である。一部のページ(上位 10%)が数千人と非常に多くのユーザからブックマークされている一方で、残りは比較的少数のユーザからしかブックマークされていないという特徴を示している。なお、SBRank

値の中央値が 144 であるのに対して平均は 1115 であった。この偏りは、「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、結果として参照されること機会が多くなるため、その順位を保持し続ける傾向がある」という Cho らによる報告[2]と関係する。

2.3 PageRank 値と SBRank 値との間の相関

図 3 は各ページの PageRank 値と SBRank 値をプロットしたものであり、PageRank 値と SBRank 値の関係を提示している。この図から、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取れる。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 ($r=0.53$) を観測した。SBRank 値と PageRank 値の相関がそれほど高くなく、どちらかが他方から従属するというわけではないため、SBRank と PageRank はお互いに補完できる可能性を秘めているといえる。

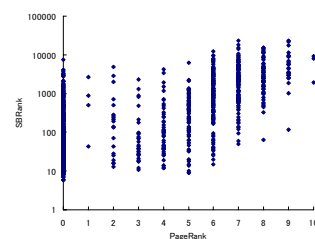


図 3 PageRank 値と SBRank 値の散布図

Fig. 3 Scatter plot of PageRank and SBRank

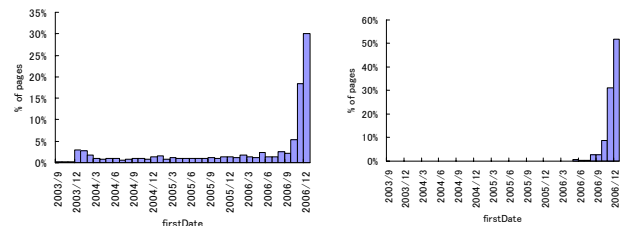


図 4 各ページの firstDate の分布 (左:全てのページ, 右:PageRank=0 のページのみ)

Fig. 4 Histogram of FirstDate of page (left: all pages, right: PageRank=0 pages only)

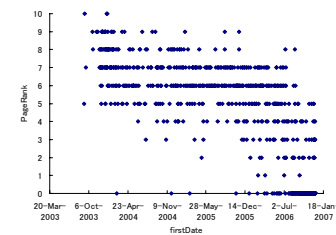


図 5 firstDate と PageRank 値の散布図

Fig. 5 Scatter plot of FirstDate and PageRank

2.4 時間軸による分析

図 4 左のグラフは、各ページが del.icio.us で初めてブックマークされた日付(firstDate)を 1 ヶ月ごとにグループ化し、ヒストグラムで表記したものである。この図から、今回データソースとしたページの半数が、最初に del.icio.us でブックマークされてから 3 ヶ月を経過していないということがわかる。このことから、ソーシャルブックマークのユーザ

は、新しく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ーの典型的な振る舞いとして興味深い。

図 4 右のグラフは、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 5 は図 4 と比べて、プロットされた点が実験日から見て最近の日付に非常に偏っている。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。一般に、この結果は SBRank がリンク構造をもとにしたランキングアルゴリズムよりも有用な側面を示している。例えば、リンク構造に基づいたページランキングのアプローチは新鮮な情報の検索の観点から見て効果的とはいえない。これは、あるページが一定の被リンクを獲得するのに比較的長い時間を必要とするからである。結論として、新しいページは、たとえその品質が非常に高かったとしても一般的な検索エンジンで探し出すことは非常に難しくなる。そしてこれは、そのページに対する他のページからのリンク数の成長速度[1]に比べ、非常に短い。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は、現時点では最適な戦略であると思われる。

図 5 は横軸に firstDate、縦軸に PageRank 値をとった散布図である。この図は、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 値との関係を示している。この図から、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ほど一般に PageRank 値が高い、という傾向がわかる。これに関連して firstDate と PageRank 値の間の相関係数を調べたところ、高い負の相関($r=-0.85$)を観測した。

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勝っているといえる。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を秘めているといえる。

2.5 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、それにもとづく我々の推測について列挙する。

- del.icio.us 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

3. PageRank と SBRank を統合した検索

本節では、両ランキング尺度を組み合わせた統合型 Web 検索手法の将来性を調査する。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提案する。

$$\alpha * SBRank + (1 - \alpha) * PageRank \quad (1)$$

実際のランキングは、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 SBRank 値と PageRank 値をもとにして行う。具体的には個々のページの 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。

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 以外にも色々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、“nintendo wii”，“social network”，“iphone”，“gardening”という 4 つの検索クエリを選択した。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。さらに、検索結果中で高品質なページを手作業で判定した。この判定は、a. 検索クエリに対するページの関連性、b. ページに記述された情報の新鮮さ、および c. Mandl によるクエリ独立なページの品質を測る尺度[6]という 3 つの尺度に基づいて行った。Mandl による尺度には、ページのテキストの量、ページに含まれるリンクの数などの項目が含まれる。

次に、式(1)を適用した際、 α の値を変化させたときに検索結果が上位 k 件 ($k=\{10, 20, 30, 40, 50\}$) において再現率-適合率曲線がどのように変化するかを調査した (図 6 から図 8)。なお、この実験は検索結果のページの品質を手作業で判定するのに作業量的な限界があることから、解ページが検索結果の上位 50 件に全て含まれていると仮定している。 $k=50$ で全ての再現率-適合率曲線の再現率(Recall)が 1 に収束しているのはそのためである。

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、“nintendo wii”というクエリにおいて α を 0.25 に設定した場合であった(図 7)。これは、del.icio.us では技術的で新しいことに興味をもつユーザが多く、そのような内容のブログのエントリなどは SBRank 値が高くなりやすいと共に、高品質なページと判定されやすいためだと考えられる。

一方で、“social network”というクエリでは、 $k=\{10, 20\}$ において我々の手法が低い再現率と適合率を示した(図 7)。この結果は、検索結果に数件含まれていたソーシャルネットワークサービスのログイン用ページが、比較的高い SBRank 値をもつ反面、評価項目のひとつである Mandl による尺度で低く評価されたことによる。

図 8 はすべてのクエリの平均をとったものである。この図から、我々の手法は $k=10$ において $\alpha=0.25$ が、 $k \leq 20$ においては $\alpha=0.75$ が、PageRank 値のみのランキング($\alpha=0$)と同程度の適合率および再現率を示すことがわかる。

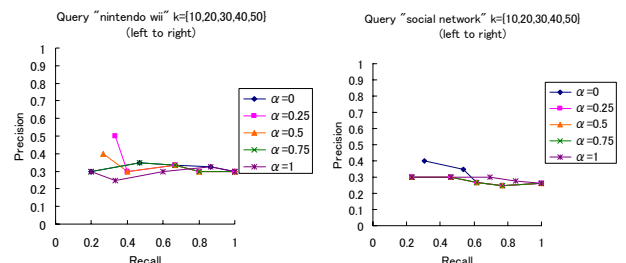


図 6, 図 7 クエリ “nintendo wii” および “social network” における再現率-適合率曲線

Fig. 6, Fig. 7 Precision-recall curves for query “nintendo wii” and “social network”

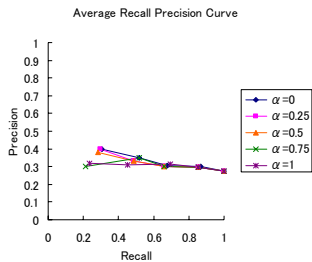


図 8 クエリ “nintendo wii”, “social network”, “iphone” および “gardening” の平均適合率・再現率曲線
Fig. 8. Average precision-recall curves for query “nintendo wii”, “social network”, “iphone”, and “gardening”

4. まとめ

ソーシャルブックマークは Web 2.0 の基盤のひとつである。既存の Web 検索を補完するためにソーシャルブックマーク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。本論文では、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ことが分かった。以上のことから、両者を統合した 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。次に、単純な SBRank 値と値を重み付けた上で合成という予備実験を通じて、SBRank と PageRank を統合するアプローチが有望であることを確認した。

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、ページ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。

【謝辞】

本研究の一部は、文部科学省 21 世紀 COE 拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」（リーダー：田中克己，平成 14～18 年度），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するものとします。

【文献】

- [1] Baeza-Yates, R., Castillo, C. and Saint-Jean, F.: Web Dynamics, Structure and Page Quality. In M. Levene and A.

Poulovassilis (eds.) "Web Dynamics", Springer, pp. 93-109, 2004.

- [2] Cho, J., Roy S., and Adams R.: "Page Quality: In Search of an Unbiased Web Ranking," In Proceedings of SIGMOD Conference 2005, pp. 551-562
- [3] del.icio.us, <http://del.icio.us>
- [4] Golder, S.A. and Huberman, B.A.: The Structure of Collaborative Tagging Systems, Journal of Information Science, 2006
- [5] Haveliwala, T. H.: Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Transactions on Knowledge and Data Engineering, 2003
- [6] Mandl, T.: Implementation and evaluation of a quality-based search engine. Hypertext2006, 73-84.
- [7] Marlow, C., Naaman, M., Boyd, D. and Davis, M.: "HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read", Proceedings of ACM HyperText 2006 Conference, 2006.
- [8] Mathes, A.: Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Computer Mediated Communication, LIS590CMC (Doctoral Seminar), 2004.
- [9] Page, L., Brin S., Motwaniand, R. and Winograd, T.: The pagerank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.
- [10] Strutz, D. N.: Communal Categorization: The Folksonomy. INFO622: Content Representation, 2004.
- [11] Wu, H., Zubair, M. and Maly, K.: Harvesting Social Knowledge from Folksonomies. Proceedings of ACM HyperText 2006 Conference, 2006.
- [12] Zhang, L., Wu, X. and Yu, Y.: Emergent Semantics from Folksonomies: A Quantitative Study Journal on Data Semantics VI, LNCS 4090, 2006, pp.168-186.

山家 雄介 Yusuke YANBE

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。日本データベース学会学生会員。

中村 聡史 Satoshi NAKAMURA

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。情報処理学会，日本データベース学会会員。

アダム ヤトフト Adam JATOWT

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過程修了。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 Katsumi TANAKA

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，情報処理学会，日本データベース学会各会員。