

セレクトブックマ：ソーシャルブックマークの 時間情報を用いた情報フィルタリング検索

上野 大樹

安村 通晃

{ueno1,yasumura}@sfc.keio.ac.jp
慶應義塾大学 政策・メディア研究科

概要

現在インターネット上の情報爆発が大きな問題となっている。そのため検索によって Web ページを発見するためには、高い検索リテラシーが必要となってしまう。そこで、本研究では膨大な情報を効率よく取得するために、ソーシャルブックマークのデータを利用した検索サービスの提案を行う。まず、ソーシャルブックマークの時間情報とブックマークされるページの種類の関連性を明らかにし、その後、その関連性に基づき効率よく所望の種類のページを取得する検索サービスの提案を行う。

1 はじめに

現在インターネット上の情報爆発が大きな問題となっている。これに対して、Google などが利用している代表的な Web ページ検索のランキングアルゴリズムとして、PageRank [1] が存在する。PageRank は、ページ間のリンク構造に基づいており、あるページへのリンクを一種の投票とみなしている。近年では、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。こうしたシステムやサイトはページ間をその品質に関わらず、自動的に結び付けている。つまり、このようなリンクの多くは人々の意思を反映しているとはいいがたく、PageRank が上手く働いていない。そのため現状の代表的な検索サービスによる Web ページの発見では、高い検索リテラシーが必要となる場合が多い。また、RSS リーダーを利用し

て情報を取得する方法もあるが、RSS リーダーを用いた場合、定期的なチェックが必要となり、取得するページ数が膨大となる可能性が高く効率よく情報を取得できるとは限らない。

このことにより、インターネット上に存在する有用な情報やサービスを発見するために口コミなどに頼る場合もあり、そもそも有用な情報やサービスの存在に気付けない場合も多い。一方近年では、ソーシャルブックマークが注目を浴びている。ソーシャルブックマークとは、インターネット上で自分のブックマークを不特定多数のユーザに公開する Web サービスである。ソーシャルブックマークでは、folksonomy という新しい情報の分類方法を利用しており、ユーザ各々がブックマークしたページに任意のタグをつけることができる。ソーシャルブックマークを用いることにより、被ブックマーク数が急激に増えたページなどから、

人気のブックマークを抽出し、興味深い情報や、最近旬な情報を発見することもできる。有用な情報やサービスを発見するにあたって、大多数の人が選んだ情報は大多数の人にとって有用な情報であり、人手による情報のフィルタリングが有効であると考えられる。

そこで、本研究では、集合知であるソーシャルブックマークデータを利用して、検索リテラシーを必要とせずに、有用な情報やサービスを効率よく発見できる手法を提案する。

2 関連研究

ソーシャルブックマークは比較的新しい Web サービスであるが、既にソーシャルブックマークデータを用いた Web ページを発見するための研究およびサービスが多数存在するため、それらについて紹介する。Sasaki らはタグを表象とする web コンテンツ群の類似性に基づいた web コンテンツ推薦システムを提案している [2]。Niwa らはソーシャルブックマークと Folksonomy を利用して、インターネット上の Web ページ全体を対象とした Web ページ推薦システムの構築手法を提案している [3]。また、Folksonomy のタグの表記のゆれの問題に対して、タグをクラスタリングすることによって解決を計っている。Yanbe らはソーシャルブックマークのブックマーク数を新たな指標 SBRank とし、PageRank と SBRank の組み合わせで、Web 検索ランキング精度の向上を計っている [4]。はてなブックマーク-じわじわ来てるエントリーは、ある程度以上のブックマーク数を集めながらも、集中的にブックマークされたことがないために「人気エントリー」に上がってきていない Web ページを検出する Web サービスである [5]。総ブックマーク数 40 以上、同日連続ブックマーク数 15

以下の Web ページを検出している。

3 ソーシャルブックマークデータ分析

国内最大規模のソーシャルブックマークサービスを提供しているはてなブックマーク [6] のデータを利用して、ブックマークされた Web ページの時間情報とその Web ページの種類に関連性について分析を行った。その結果、大まかに分けて次の 2 種類のタイプの Web ページがあることが分かった。

- (1) 急激にブックマーク数が伸びて、その後はほとんどブックマークされなくなるタイプのページ
- (2) 長い期間に渡ってブックマークされ続けるページ

また、本研究では上記の 2 種類のタイプのページを分析した。データの分析方法として、ブックマーク数 100 以上のページに対して、以下の 2 種類の Web ページをランダムに 100 ページずつ取得した。

- (1) 全日数/全ブックマーク数=0.2 以下
- (2) 全日数/全ブックマーク数=0.8 以上

ここで全日数は、ブックマークされた日数を表す。例えば、2007 年 1 月 3 日と 2007 年 1 月 6 日と 2008 年 5 月 14 日に任意のユーザからブックマークされたとすると、3 日とする。(1) に分類されるページを、急激にブックマーク数が伸びてその後ほとんどブックマークされなくなる TypeI のページとし、(2) に分類されるページを、長い期間に渡ってブックマークされ続ける TypeII のページとした。ここで、全日数/全ブックマーク数=0.2 以下と 0.8 以上で分類した理由は、ブックマーク数 100 以上のページ数が、双方で近い値、かつ、双方とも 100 ページを大きく上回るページ数を確保できたか

らである。

また、TypeI と TypeII のページの特徴についてどのような種類のページが多いか分析を行った。その結果が、以下の図 1,2 である。

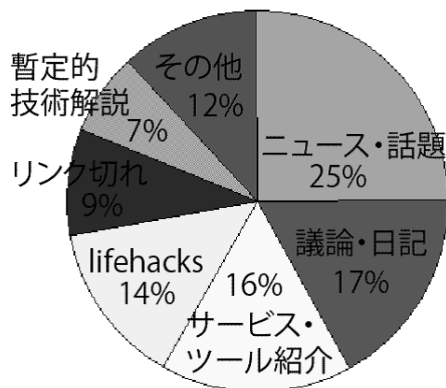


図 1: 全日数/全ブックマーク数=0.2 以下

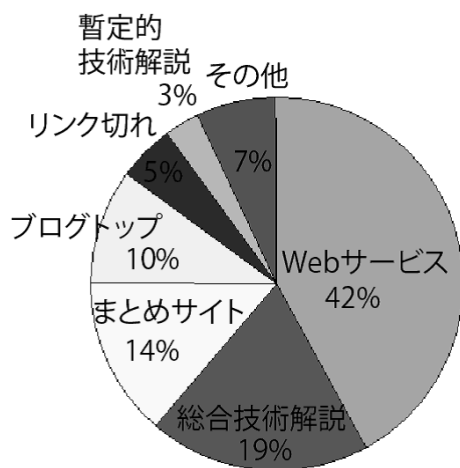


図 2: 全日数/全ブックマーク数=0.8 以上

図 1,2 から分かるように、TypeI の Web ページでは、「ニュース・話題」、「議論・日記」、「サービス・ツール紹介」が上位を占めており、一時的に利用される傾向の強い Web ページが大半を占めている。これに対して、TypeII の Web ページでは、「Web サービス」、「総合的技術解説サイト」、「ま

とめサイト」が上位を占めており、長期間に渡って利用される傾向の強い Web ページが大半を占めていることがわかった。このことから、TypeII のページを優先的に取得することによって、一時的に利用される傾向の強い Web ページ TypeI をフィルタリングして、いつ見ても有用な Web ページのみを検索できる可能性が高いことが分かった。

4 設計

第 3 章で示したように、長い期間に渡ってブックマークされ続けるページを優先的に取得することによって、いつ見ても有用な Web ページのみを検索できる可能性が高い。そこで、本研究では、長い期間に渡ってブックマークされ続ける Web ページを優先的に取得する Web サービス「セレクトブクマ」[7]を開発した。セレクトブクマでは、高い検索リテラシーを必要とせず、調べたい分野について充実した情報が得られる Web ページを、発見可能とすることを目標とした。

4.1 データ収集

2005 年 2 月～2008 年 8 月まではなブックマーク [6] のブックマーク数が 5 以上の Web ページのデータをすべて取得し、データベースに登録した。取得したデータは、以下の 5 種類である。

- Web ページの URL
- Web ページのタイトル
- 誰にブックマークされたか (ユーザ情報)
- いつブックマークされたか (時間情報)

- なんというタグ名でブックマークされたか (タグ情報)

URL 数は約 70 万 URL で、ユーザ情報、時間情報、タグ情報などを含めるとデータベースのレコード数は、約 2000 万レコードとなった。

4.2 セレクトブックマ機能

セレクトブックマは、図 3 のような画面構成となっている。図 3 は、"java" を検索単語 (タグ) として検索した結果である。①は検索結果の Web ページのタイトルを表示したもので、タイトルのリンクをクリックすると、クリックした Web ページを表示する。②は 4.3 に示す計算式に基づいて、ランキング化の際の値を表示したものである。③は検索タグを入力するテキストボックスである。タグを入力し、検索ボタンを押すことにより、指定した検索タグで検索を行う。④は人気の検索タグであり、検索回数の多いタグから順にタグ名と検索回数を表示する。タグ名のリンクをクリックすることにより、指定した検索タグで検索を行う。⑤は⑤に表示されているタグ名でブックマークされた数が一定数以上のタグの一覧のリンクを表示したものである。タグ名のリンクをクリックすることにより、指定した検索タグで検索を行う。

4.3 ランキング手法

ランキングを出すにあたって、検索単語として指定した単語 (タグ) でブックマークされた数に指定した単語 (タグ) でブックマークされた日数で重み付けをして、値の大きいものから順にランキングしている。実際の計算式は、指定したタグでブックマークされたページにおいて、以下の数値となる。

セレクトブックマβ版

昔から長い間ブックマークされ続ける優良サイトを検索して表示するサービスです。はてなブックマークのデータを利用しています。主に指定したタグでブックマークされた数と何日間ブックマークされたかを元にランキングを出しています。

興味のある分野のタグをクリックするか、テキストボックスにタグ名を入力して検索ボタンを押して下さい。

「java」に関するページの検索結果

1	Javaの道 (Java入門・リファレンス)	6557ポイント
2	Javaの学習ならJavaDrive	6318ポイント
3	Java技術最前線「ITPro」	5832ポイント
4	浅煎り珈琲 - Java アプリケーション入門	5467ポイント
5	頑健なJavaプログラムの書き方 (Writing Robust Java Code)	5330ポイント
6	Java House ML	5328ポイント
7	JavaでHello World	5325ポイント
8	Java 2 Platform SE 5.0	5325ポイント
9	Java in the Box	4550ポイント
10	Log4J徹底解説-目次	4420ポイント

検索: java [検索]

もしよろしければ、アンケートにご協力下さい。
アンケート回答へ

人気の検索タグ

debian	2855回
linux	1451回
javascript	1043回
java	779回
mala	755回
ubuntu	743回
ui	717回
ニコニコ動画	708回
映画	694回
twitter	691回
ruby	668回
interface	663回
english	661回
音楽	657回
2ch	645回
写真	638回
動物	617回
1000speakers	616回
youtube	614回
これはずごい	606回

1000speakers 18禁 2ch 2chまとめ 3d 801 @! aa actionscript ad adobe air ajax algorithm amazon america anime apache api app apple art as3 asahi.com atok audio backup baseball bicycle blog blogger bluetooth book bookmark bookmarklet books browser business c# c++ cafe cakephp calendar camera car cd java china chumby cinema clamp classic clip cm cms cnet color column comic command communication community cooking cool copyright CSS culture dankogai debian debug design development dictionary display docomo download dpz dropdb dtm dvd eclipse economics editor education ekken emacs emobile english event

図 3: セレクトブックマ画面

ブックマーク数×日数

ここで言う日数とは、日付ごとにブックマークされた日数であり、例えば、2007年1月3日と2007年1月6日と2008年5月14日に任意のユーザからブックマークされたとすると、3日とする。2007年1月3日に100人のユーザからブックマークされたとすると1日とする。このような計算式にした理由を以下に述べる。ユーザがブックマーク

するという行為は、少なからずブックマークするページに対して興味を持ったという意味を持っている。そこで、ブックマーク数はユーザの Web ページの評価を表していると考えられる。だが、ソーシャルブックマークの特性上、単純にブックマーク数が多いだけでは、タイムリーな情報など、後から見た場合あまり有用ではないページも適用されてしまう。そこで、ブックマーク数に日数で重み付けを行って短期的にしか必要でない Web ページをフィルタリングして、ランキングを出している。

5 実験結果

”java”および”ui”という検索単語に対して、以下の3種類について上位10件の検索結果を比較検討した。

- Google 検索 (2008年11月24日時点)
- 指定したタグでのブックマーク数
- セレクトブックマによる検索

検索結果の上位10件を以下の表1~6に示す。まず、検索単語”java”について検索結果

表 1: Google 検索 (java)

順位	Web ページタイトル
1	java.com: あなたと Java
2	無料 Java ソフトウェアをダウンロード - Sun Microsystems
3	Java テクノロジ - サン・マイクロシステムズ
4	Java - Wikipedia
5	Java の道 (Java 入門・リファレンス)
6	Java とは - 意味・解説 : IT 用語辞典
7	Sun Developer Connection - Java Developer Connection
8	JAVA 動物実験の廃止を求める会
9	Java で Hello World
10	Java Solution - @ IT

表 2: 指定したタグでのブックマーク数 (java)

順位	Web ページタイトル
1	Java のクラスアンロード (Class Unloading)
2	Java の道 (Java 入門・リファレンス)
3	頑健な Java プログラムの書き方 (Writing Robust Java Code)
4	Java 技術最前線 : ITpro
5	Java の学習なら JavaDrive
6	浅煎り珈琲 -Java アプリケーション入門
7	Java で Hello World
8	Java 2 Platform SE 5.0
9	【レポート】Java 初学者には最適!? 解説から実行までブラウザでコンプリート - Javala (MYCOM ジャーナル)
10	Java House ML

表 3: セレクトブックマによる検索 (java)

順位	Web ページタイトル
1	Java の道 (Java 入門・リファレンス)
2	Java の学習なら JavaDrive
3	Java 技術最前線 : ITpro
4	浅煎り珈琲 -Java アプリケーション入門
5	頑健な Java プログラムの書き方 (Writing Robust Java Code)
6	Java House ML
7	Java で Hello World
8	Java 2 Platform SE 5.0
9	Java in the Box
10	Log4J 徹底解説~目次

の比較検討を行う。Google 検索では、Java 関連のダウンロードサイトや Java 自体の意味解説サイト、wiki などが上位にランキングしているのに対して、セレクトブックマによる検索では、Java 全般に関する総合的な技術解説サイトが上位にランキングしていることが分かる。また、指定したタグでのブックマーク数とセレクトブックマによる検索を比較すると、どちらも Java 技術解説サイトが上位にランキングされているが、指

表 4: Google 検索 (ui)

順位	Web ページタイトル
1	ユーザ インタ フェース - Wikipedia
2	UI - Wikipedia
3	ユーザ インターフェースとは【user interface】 - 意味・解説 : IT 用語辞典
4	UI とは【ユーザ インターフェース】 - 意味・解説 : IT 用語辞典
5	「使える、使いやすい、使いたい」と思える UI とは? - @IT
6	UI ゼンセン同盟
7	UI パターンいろいろ - Design-Walker
8	ふくしま UI ターン
9	ソシオメディア — UI デザインパターン
10	MOONGIFT: ≫ Prototype.js を用いた UI ライブラリ「Prototype UI」:オープンソースを毎日紹介

表 5: 指定したタグでのブックマーク数 (ui)

順位	Web ページタイトル
1	UI-patterns.com
2	Life is beautiful: 直感的な UI と hande-eye-cordination の話
3	ソシオメディア — UI デザインパターン 最新の UI デザインパターン
4	直観的なインタフェースをめざして
5	ぼくはまちちゃん!(Hatena) - UI について思うこと
6	ユーザーインターフェースデザイン研究室
7	prima materia diary - Google の UI は OK/キャンセルを訊いてこない
8	80年代の Apple に学ぶ UI の部品化とガイドライン? @ IT
9	naoya のはてなダイアリー - インタフェースの話
10	キャズムを超える! - 家電メーカーよ、今すぐその時代遅れの UI から脱却せよ

定したタグでのブックマーク数の方が Java 全般の総合的な技術サイトではなく、Java に関する一部の技術を解説している暫定的

表 6: セレクトブックマによる検索 (ui)

順位	Web ページタイトル
1	ソシオメディア — UI デザインパターン 最新の UI デザインパターン
2	UI-patterns.com
3	ユーザーインターフェースデザイン研究室
4	Technologies for UI
5	@ IT : Web アプリケーションのユーザーインターフェイス [1] -1
6	ソシオメディア — UI デザインパターン
7	アップル ヒューマンインタフェースガイドライン
8	Joel on Software - 環境をコントロールできれば楽しく感じるもの
9	Yahoo! Design Pattern Library
10	ダメなユーザインタフェイス講座

な技術解説サイトが多いことが分かる。

次に検索単語”ui”について検索結果の比較検討を行う。Google 検索では、UI という用語自体の解説サイトや wiki などが上位にランキングしているのに対して、セレクトブックマによる検索では、UI のデザインパターンのまとめサイトや UI に関する総合的な解説サイトが上位にランキングされている。また、指定したタグでのブックマーク数では、セレクトブックマによる検索と比較して、UI に対する議論や部分的な UI の解説サイトが多くランキングされていることが分かる。

以上により、検索単語自体の意味を調べたい場合は Google 検索が便利であるが、検索単語について詳しく調査したい場合などは、セレクトブックマの方が楽に情報を得られる可能性があることが分かった。また、長い期間ブックマークされる Web ページを優先的にランキングすることにより、特定の場合に必要なような暫定的な Web ページをフィルタリングして、何度も利用できるような総合的な Web ページを取得できる可能性が高いことが分かった。

6 考察

第5章より、セレクトブックマはGoogle検索と比較して、検索単語について総合的に詳しく調査したい場合に、より有用であることが分かった。また、特定の目的でのみ必要となるような、暫定的なWebページをフィルタリングする効果もあることが分かった。だが、セレクトブックマのランキングロジックで、暫定的なWebページを完全にフィルタリングできているとは言えない。暫定的なWebページをフィルタリングするために、より良いランキングロジックが存在する可能性がある。現状では、指定したタグでのブックマーク数に日数をかけているが、日数に係数をかけて、より日数に重きを置いたほうが良いのかもしれない。また、現状では日数の閾値を1日としているが、この閾値の値も変えたほうが良いのかもしれない。

そのため、今後の課題としては、さらに詳しくソーシャルブックマークの時間情報とWebページの種類の関係を分析する必要がある。さらに、日数の係数の値や日数の閾値の値を変化させて実験を行い、最適なロジックを探る。最終的には、Google検索とブックマーク数の多い順にランキングした場合とランキングロジックを最適化したセレクトブックマの3種類の比較を、多数のタグおよび多数の被験者を利用して行い、本研究の有用性を、より明らかにしていく。

7 結論

ソーシャルブックマークに登録されるWebページは、短期間のみブックマークされるWebページと長期間に渡ってブックマークされ続けるWebページが存在する。これらについて分析を行ったところ、短期間のみブックマークされるWebページは一時的に必要とされるWebページである傾向

が強く、長期間に渡ってブックマークされ続けるWebページは、何度も必要とされるWebページである傾向が強いことが分かった。このことから、ソーシャルブックマークの時間情報を利用して一時的に必要となるようなWebページをフィルタリングすることが可能であると考えられた。そこで、長い期間に渡ってブックマークされ続けるWebページを優先的に取得するWebサービスであるセレクトブックマを開発し、一時的に必要となるようなWebページがフィルタリングされていることを確認した。さらに、Google検索とセレクトブックマによる検索の比較も行った。その結果、検索単語について詳しく調査したい場合、セレクトブックマによる検索の方が、高い検索リテラシーを必要とせず有用なWebページを発見できることが分かった。

参考文献

- [1] 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.
- [2] A. Sasaki, T. Miyata, Y. Inazumi, A. Kobayashi and Y. Sakai : Web Content Recommendation System based on Similarities among Contents Cluster of Social Bookmark, *DBWeb*, pp.59-66, 2006.
- [3] S. Niwa, T. Doi and S. Honiden : Web Page Recommender System based on Folksonomy Mining, *Proc. 3rd International Conference on Information Technology : New Generations (ITNG'06)*, pp.388-393, 2006.
- [4] Yusuke Yanbe, Satoshi Nakamura, Adam Jatowt, Katsumi Tanaka : Uti-

lizing Social Bookmark Characteristics to Enhance Ranking in Web Search, *DBSJ Letters*, Vols.6, No.1, 2007.

[5] <http://k52.org/jwjw/>

[6] <http://b.hatena.ne.jp/>

[7] <http://plazman.chi.mag.keio.ac.jp/sbm/summary.jsp>