ソーシャルブックマークの時間スケールに着目した
長期間利用するWebページ収集支援システムの研究

Research on Web Information Gathering Support System focusing Time Scale using Social Bookmark

上野 大樹（うえの たいき・Taiki Ueno）[1]・樋口 文人（ひぐち ふみと・Higuchi Fumito）[2]・安村 通晃（やすむら みちあき・Yasumura Michiaki）[3]

[1]慶應義塾大学大学院政策・メディア研究科　博士課程　・　[2]慶應義塾大学環境情報学部　非常勤講師　・　[3]慶應義塾大学環境情報学部　教授

［Abstract］
Recently the contents on Web become massive and diversified. As the result, it is not always easy for us to get Web pages and Web services that are able to use for long periods of time only by utilizing general search engines. Thus, in this research, we used social bookmark data as base data in order to get contents utilized for long periods of time by filtering a huge variety of information. First we analyzed social bookmark data, and we found that Web pages that are bookmarked merely temporarily tend to have temporal information. On the other hand, Web pages that are continuously bookmarked for long periods of time tend to have information useful for long periods of time. Based on this characteristic of time information of social bookmark, we developed the Web system called “SelectBukuma” that can pick up contents utilized for long periods of time from a huge variety of information. Further, we evaluated the system whether or not it can find and gather Web pages of the long term information. The results of evaluation show that ”SelectBukuma” is more effective than general search engines in order to get Web pages and Web services that are used long period of time. The results also show, using bookmark time span is useful information for filtering Web pages.

［キーワード］
Web検索、情報フィルタリング、ソーシャルブックマーク、Web2.0、長期的利用、情報収集

１．はじめに

Webは集合知の宝庫であり，必要な情報にアクセスさえできれば非常に有用である．一方で，Webは混沌としており，あらゆる情報で溢れているために，望んだ情報にアクセスすることが困難である場合も多い．このアクセスの問題を解決することがWeb技術において非常に重要なテーマである．

近年では，Web上には最新の情報や今流行の情報が増大しており，むしろ，長期的に利用する情報を発見することが難しくなってきている．一般的なGoogleやYahooなどの検索エンジンは，長期的に利用する情報を中心に取得するサービスではなく，またこういった長期的に利用する情報へのアクセスについての研究事例も少ない．そのため，Web上から長期的に利用するWebサイトを発見・収集することは手軽にはできず，知人からの口コミによってようやくそのWebサイトの存在に気付かされる場合もある．

しかし，ユーザの興味のある分野に対して，Web上の長期的に利用する体系だった情報やWebサービスを知っておくことは重要である．例えば，コンピュータ関連の技術情報のように何度も利用する長期的に必要な情報が載っているWebページや便利なWebサービスを知っていることは，現代の情報化社会における技術開発などにとって不可欠なものである．

こうしたことから，本研究では，長期間利用する有益なWebページとWebサービスを手軽に取得・発見する手法を提案することを目的とする．以上の目的を達成するために，まず長期的に利用される情報は，長期的に多くのユーザからアクセスされたり，ブックマークされたりする可能性が高いという仮説を立てた．次に，ソーシャルブックマークデータを利用・分析して，この仮説を実証した．そして，どれほど長期にわたってブックマークされ続けるかという時間情報を利用して情報をフィルタリングする，情報検索・収集システム「セレクトブクマ」

を開発し，評価した．

２． 背景

２．１．Web 検索の背景

かつて，Web 検索エンジンは，人手で検索結果を作成するディレクトリ型の検索エンジンが主流であった．例えば，初期の頃の Yahoo の検索エンジンなどがこれにあたる．だが，Web ページの急激な増加に伴い，完全な人手での Web ページの選択・分類が難しくなってきた．そこで，近年では，ロボットが自動的に Web ページを巡回して，検索結果を生成するロボット型の検索エンジンが主流となっている．例えば， Google や現在の Yahoo の検索エンジンなどがこれにあたる．この中の代表的ランキングアルゴリズムとして，PageRank[1]がある．

だが，現在のロボット型の検索エンジンでは，検索結果が大量に取得されるため，検索キーワードの選択が重要であるが，その適切な選択は容易でない．また，検索者にとって不要な情報も大量に検索結果としてひっかかってしまい，検索結果の中から所望の Web ページを見つけるのも容易ではない場合がある．

そこで，近年では，情報フィルタリングや情報レコメンデーションの重要性が問われており，人にやさしい検索エンジンの研究がさかんになってきている．さらに，そのために人手を使ったソーシャルな検索手法も多数提案されている．

２．２．近年の Web 上の情報収集システム

近年では，一般的な Web 検索エンジンからの情報取得以外にも，RSS リーダー，ソーシャルブックマーク，さらには twitter[2]などのマイクロブログサービスを用いて情報を取得することも可能である．以上のようなサービスは，非常に有用な情報取得ツールであるが，これらは主に最新の情報を取得するために利用される場合が多く，どちらかと言えば短期的に利用する情報を収集するのに向いている．さらに，こういったサービスでは日々大量の情報を取得することが可能だが，中には自分にとって不要な情報も大量に存在し，情報のフィルタリングがうまくなされていない場合も多い．

だが，これらのサービスもサービス開始からの年数が経過してきて，多くのユーザが利用することによって有益な情報が日々蓄積されてきている．蓄積されたデータをうまく有効活用することによって，ユーザが興味のある分野に対して，うまく情報をフィルタリングして，もっと手軽に有益な情報を発見できる可能性がある．

２．３．ソーシャルブックマークとは

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し，有益な Web ページを共有する 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．付与したタグを利用して自分や他人が過去にブックマークした Web ページを見つけやすくしている．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．日本国内では，主なソーシャルブックマークサービスとして，はてなブックマーク[3]，livedoor クリップ[4]，Yahoo!ブックマーク[5]などがある．海外での主なソーシャルブックマークサービスとして，del.icio.us[6]，Digg[7]などがある．

ソーシャルブックマークのデータは，ユーザが気に入った Web ページをブックマークし，かつ，ブックマークした Web ページにタグやコメントを付与することが可能である．さらに，ブックマークした時間情報も保持されている．ソーシャルブックマークによって，多くの Web ページには， ブックマーク数，ユーザ ID，タグ，コメント，ブックマークされた時間などのメタ情報が付与されている．

本研究では，ソーシャルブックマークデータを大量に利用し，上述のメタ情報を有効活用することによって，一般の Web 検索エンジンではできない，情報収集手法や情報フィルタリング手法を提案できないか考えた．そのために，まずソーシャルブックマークデータを大量に収集し，分析した．

３．関連研究

ソーシャルブックマークデータを利用した先行研究について紹介する．ソーシャルブックマークデータを利用した先行研究としては，主に以下の４種類の研究に分類できる．

(1) ソーシャルブックマーク，Folksonomy の分析
(2) Web ページの検索
(3) Web ページの推薦
(4) ソーシャルブックマークユーザの推薦

上述の中で本研究は，Web ページの検索に分類されるが，Web ページの推薦という側面も持っている．

(1) ソーシャルブックマーク，Folksonomy の分析

Golder らは，ソーシャルブックマークのユーザやタグ，ブックマークの性質について分析し，各 Web ページに対する各タグの出現頻度は一定値に収束することを証明している[8]．

Paul らは，del.icio.us のデータを収集して，ソーシャルブックマークが Web 検索において大きな改革を起こせるかどうか検討している．その結果，現状ではソーシャルブックマークのデータ量不足の問題やタグのゆらぎの問題から，現時点のデータでは，Web 検索に関して劇的な改革は起こせないが，今後ソーシャルブックマークのデータ量が急激に増えたりした場合は，Web 検索において改革を起こせる可能性があると結論づけている[9]．

川中らは，あるタグと共起関係の強いタグを取得し，出現時期の早いほうを親タグとする手法を用いることによって，タグの時系列の関係性をグラフ化している[10]．

(2) Web ページの検索

Xu らは，ソーシャルブックマークデータを利用したパーソナライズド検索手法を提案している[11]．さらに，ソーシャルブックマークのタグ情報を利用することによって，パーソナライズド検索を自動的に評価する手法も提案している．

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 を提案し，PageRank と SBRank を統合して，Web 検索ランキング精度の向上を計っている[12]．

高橋らは，ソーシャルブックマークデータの時間データを利用して，鮮度の高い Web ページを取得する検索手法を提案している[13]．ここでは，ブックマーク日時の散らばりの大きさから，Web ページの賞味期限を判定して，賞味期限を過ぎていない鮮度の高い Web ページを取得している．この研究は，ソーシャルブックマークの時間データに着目した情報検索手法を提案しているため，本研究と近いが，本研究との違いは，この研究では，鮮度の高い Web ページを取得することを目的としているのに対して，本件研究では，長期に渡って役立つ Web ページを取得することを目的としている点である．そのため，時間データの利用の仕方も異なっている．

(3) Web ページの推薦

Niwa らは，タグのクラスタリングをおこなうことによりタグの表記ゆれの問題の解決をはかり，ユーザのブックマーク情報からユーザの趣向に沿った Web ページの推薦をおこなう手法を提案している[14]．

佐々木らは，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[15]．

(4) ソーシャルブックマークユーザの推薦

白土らは，ソーシャルブックマークユーザのブックマーク情報からユーザの関連度を解析した結果から興味の類似したユーザを推薦し，ネットワーク図として表示するシステムを構築した[16]．

大力らは，ソーシャルブックマークユーザの中のイノベータ，いわばαブックマーカーに注目した情報推薦手法を提案している[17]．

４．ソーシャルブックマークデータ分析

４．１．ソーシャルブックマークデータ収集

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のデータを収集した．はてなブックマークは，2010 年 1 月現在，約 30 万人，ブックマーク数は約 5000 万ブックマーク程の規模がある．その中から，2005 年 5 月～2008 年 9 月までにブックマークされたデータの中でブックマーク数 5 以上のページの以下のデータをすべてデータベースに収集した．

- URL

- タイトル
- ユーザ ID
- ブックマークされた日付
- 付与されたタグ

はてなブックマークのデータ収集には，はてなブックマーク API を利用した．データベースに収集したデータの量は，表-1 に示す．

表-1　収集したデータ量

データ名	データ量
URL 数	762,239URL
ブックマーク数	12,751,661 ブックマーク
ユーザ数	87,898 人
タグ数	17,168,666 タグ
タグ数（種類）	252,512 種類
レコード数	21,686,536 レコード

収集したデータから計算すると，ひとりが 1 つのブックマークをするときに平均して約 1.35 個のタグを付与していることがわかる．

４．２．時間情報に関する分析

ソーシャルブックマークデータにおいて，いつ，どれくらいブックマークされるか，ブックマーク数と時間の関係について分析を行った．その結果，大まかに分類すると，以下の 2 種類のタイプの Web ページに分類できることがわかった．

Type1 ： 短期間しかブックマークされないページ（図-1）
Type2 ： 長期間ブックマークされ続けるページ（図-2）

以上の Type1 と Type2 の Web ページに対して，その Web ページがどういった種類の Web ページであるかを分析した．その分析結果を以下の図-1 と図-2 に示す．分析対象のページは，以下の条件とした．

- Type1 は，ブックマークされた日数/全ブックマーク数=0.2 以下
- Type2 は，ブックマークされた日数/全ブックマーク数=0.8 以上
- Type1 と Type2 に対して，ブックマーク数 100 以上のページをランダムに 100 ページずつ取得

ここでブックマークされた日数とは，ユーザから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 2 月 10 日と 2008 年 10 月 10 日にブックマークされた場合，3 日とする．ここで，ブックマークされた日数/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，双方ともブックマーク数100以上のページの数が1000ページ以上取得できたからである．ブックマーク数とは，そのページが何回ブックマークされたかを表す．

図-1，図-2 から分かるように，Type1 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，Type2 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，長期間にわた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．

このことから，Type2 のような長期間にわたってブックマークされ続けるような Web ページを優先的に取得することによって，Type1 のような一時的に利用される傾向の強い Web ページの優先順位を下げて，いつ見ても有用な Web ページを優先的に検索できる可能性が高いことがわかった．

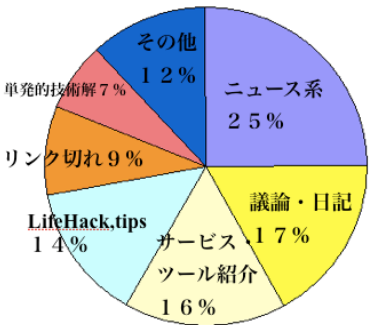

図-1　Type1 の Web ページの種類

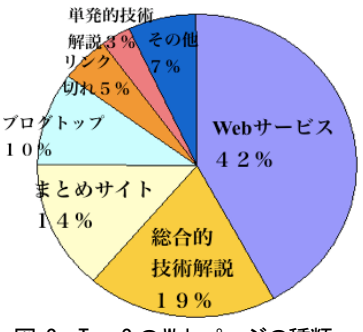

図-2　Type2 の Web ページの種類

5. セレクトブクマの提案と試作

5.1. セレクトブクマの提案

5.1.1. セレクトブクマの概要

本研究では，ソーシャルブックマークのデータを利用した情報収集システム「セレクトブクマ」を提案・実装した．セレクトブクマでは，調べたい分野に対して，ソーシャルブックマークのブックマーク数とブックマークされた日数という二つの指標を利用して，Web ページをランキング化している．セレクトブクマを利用することによって，手軽に興味ある分野に関して長期的に利用する Web ページや Web サービスを発見・収集する．

5.1.2. セレクトブクマの設計思想

セレクトブクマでは，長期間利用される Web ページや Web サービスを発見・収集することを目的としている．また，情報収集の際の手軽さを重視している．

そのために，1 つ目の指標として，ソーシャルブックマークのブックマーク数という指標を利用している．これは，ユーザがブックマークするという行為が Web ページへの評価であるという考えに基づいている．2 つ目の指標として，ブックマークされた日数を利用している．これは，第 3 章の分析結果に基づき，長い間ブックマークされ続ける Web ページは，長期間必要とされる種類の Web ページが多いことを利用している．このブックマークされた日数という指標を利用し，長い間ブックマークされ続ける Web ページを優先的に取得することによって，一時的にしか利用しない Web ページを検索結果から排除することができると考えた．

5.2. セレクトブクマの機能

セレクトブクマの検索前の画面と検索後の画面を図-3 に示す．

図-3 の(1)～(4)の説明を以下に示す．

(1) 検索単語（タグ）入力ボックス

検索単語(タグ)を入力するテキストボックスで，タグを入力し，検索ボタンを押すことにより，指定したタグで検索を行う．

(2) 検索回数の多いタグ

人気のタグであり，検索回数の多い順に並べたものである．すべてを総合して検索回数の多い順に並べた「総合」と，「技術」，「趣味」，「社会・生活」，「その他」のカテゴリごとに検索回数の多い順に並べたものがある．

(3) 検索結果の Web ページのタイトル

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．

(4) ランキングの値

後に示すランキングの計算式を用いて計算した値とその値を棒グラフで可視化したものである．

図-3　セレクトブクマの画面

5.3. 検索ランキングロジック

セレクトブクマでは，検索結果のランキングを出すにあたって，検索単語（タグ）として指定したタグでのブ

ックマーク数に，指定したタグでブックマークされた日数で重み付けをして，値の大きいものほど順位が高くなるようにランキングを行っている．

ランキングの計算式を以下に示す．以下の式の日数にかける係数αの値を大きくすればするほど、時間情報の影響が強くなる．

$$\text{Bookmarks} \times \text{Days}^{\alpha} \cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{(式1)}$$

Bookmarks：指定したタグでのブックマーク数
Days：指定したタグでブックマークされた日数
α：重み付けのパラメータ

本研究では，検索結果のランキング手法について，他にもさまざまな手法を考案し，試作した．ひとつひとつの手法についての詳細な評価は行っていないが，上記の手法が現状ではもっとも効果的であった．今回，上述の手法を採用した理由は，ブックマーク数と日数という指標を利用し，かつ，自由に日数での重み付けをしたかったからである．ブックマーク数は， Web ページ人気を表している場合が多い．そのため，ある程度以上人気のある記事の中から，どれだけ長期的に利用されている Web ページかどうかという指標の重みを自由に決定できる．

５．４．システム構成・運用

セレクトブクマは，Web サービスとして実装した．画面の表示部分は，HTML，JSP，JavaScript を利用し，計算などやデータベースとの連携は，主に Java を利用している．Java と JavaScript の連携は，Ajax 方式を用いて，JSON 形式でデータの受け渡しをしている．ユーザが検索を行う場合，主に以下の手順でシステムが動作する．

(1) ユーザが検索する
(2) 検索単語がサーバへ送られる
(3) 検索単語でデータベースを検索する
(4) データベースの検索結果からランキングを計算する
(5) ランキングに基づき，ユーザに検索結果を返す

セレクトブクマのシステム構成図を図-4 に示す．

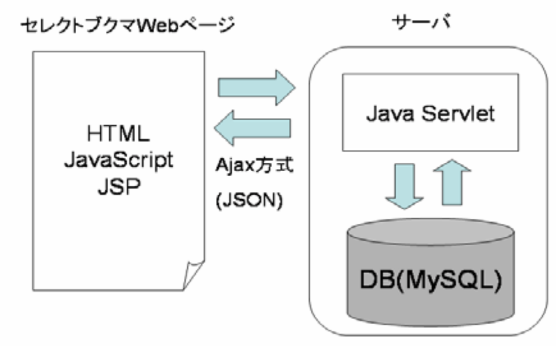

図-4　システム構成図

セレクトブクマは，2008 年 10 月 22 日から Web 公開を開始し，運用している．URL は以下の通りである．
http://plazman.chi.mag.keio.ac.jp/sbm/summary.jsp

６． セレクトブクマの評価

６．１．実験概要

本実験の目的は，主に以下の 2 点である．

(1) セレクトブクマと既存の検索エンジンにおいて，どちらがより長期間利用する Web ページを手軽に取得できるか評価する
(2) セレクトブクマで利用している手法である，時間情報による重み付けによって，一時的に必要とされる情報の優先度を下げることができているか評価する

そのため，以下 3 種類の手法において比較実験をおこなった．

- Google 検索
- 取得したはてなブックマークのデータの中で，タグで検索した場合のブックマーク数が多いものから順にランキングしたもの（以降，ブックマーク数順）
- セレクトブクマ（ただし，α=1）

６．２．実験方法

上述の 3 種類の手法で検索を行い，検索結果上位各 30 件を取得した．取得した合計 90 件の Web ページの中で

重複したものを除いたWebページを、順番をランダムにしたリストとして示した．そのリストのから被験者に今後も利用したいと思うWebページを１位〜10位まで選んでもらった．

被験者が選んだWebページを適合文書として，情報検索手法の一般的な評価方法である適合率，再現率を求めた．実験条件は表−2に示す．

表−2　実験条件

検索単語	「java」，「健康」，「映画」の３つ
「java」の被験者	10名（男性６名，女性４名）
「健康」の被験者	10名（男性５名，女性５名）
「映画」の被験者	10名（男性３名，女性７名）
適合文書１（人数から作成）	３人以上の被験者が選んだページ
適合文書２（得点から作成）	８点以上の合計得点が付与されたページ 被験者がつけた順位によって得点を付与する １位：10点，２位：９点，．．．，10位：１点

検索単語（タグ）は，技術系の分野から「java」，生活系の分野から「健康」，娯楽系の分野から「映画」と３つの異なった分野から１つずつ選んだ．それぞれ，技術系，生活系，娯楽系の中でも，特別セレクトブクマが有利になるような単語ではなく，できるだけ一般的な単語を選んだ．また，Google検索結果は，2009年12月23日に検索した検索結果を利用した．

６．３．実験結果

各手法による検索結果は割愛するが，検索結果を見ると，Google検索の場合は，権威あるWebサイトが上位にランキングされやすい傾向にあるのに対して，逆にセレクトブクマでは，個人で作成したようなWebサイトが上位にランキングされやすい傾向にあることがわかる．また，ブックマーク数順とセレクトブクマでは，ある程度ランキングされるWebページが近い傾向にあることがわかる．これは，セレクトブクマのロジックが，ブックマーク数に対して日付で重みを付けていることに起因している．

ここで全適合文書の件数を表−3に示す．

表−3　全適合文書の件数

単語名	java	健康	映画
適合文書１の全適合文書数[件]	13	16	19
適合文書２の全適合文書数[件]	26	30	29

各検索手法においての検索結果上位10件と上位30件の適合率と検索結果上位30件の再現率の値を表−4〜表−6に示す．ここで，適合率と再現率の式を以下に示す．

$$\text{適合率} = \frac{\text{検索結果中の適合文書の数}}{\text{検索結果の文書の数}} \quad \cdots\cdots\cdots\cdots \text{(式2)}$$

$$\text{再現率} = \frac{\text{検索結果中の適合文書の数}}{\text{全適合文書の数}} \quad \cdots\cdots\cdots\cdots \text{(式3)}$$

表−4　「java」で検索した場合の適合率・再現率

	セレクトブクマ	ブックマーク数順	Google検索
適合文書１を利用した適合率（上位10件）[%]※	40	40	10
適合文書１を利用した適合率（上位30件）[%]※	33	17	17
適合文書２を利用した適合率（上位10件）[%]※	70	60	10
適合文書２を利用した適合率（上位30件）[%]※	57	43	30
適合文書１を利用した再現率（上位30件）[%]※	77	38	38
適合文書２を利用した再現率（上位30件）[%]※	65	50	35

表-5 「健康」で検索した場合の適合率・再現率

	セレクトブクマ	ブックマーク数順	Google 検索
適合文書１を利用した適合率（上位 10 件）[%]※	40	30	0
適合文書１を利用した適合率（上位 30 件）[%]※	40	37	0
適合文書２を利用した適合率（上位 10 件）[%]※	70	60	30
適合文書２を利用した適合率（上位 30 件）[%]※	63	60	20
適合文書１を利用した再現率（上位 30 件）[%]※	81	69	0
適合文書２を利用した再現率（上位 30 件）[%]※	63	60	20

表-6 「映画」で検索した場合の適合率・再現率

	セレクトブクマ	ブックマーク数順	Google 検索
適合文書１を利用した適合率（上位 10 件）[%]※	50	40	60
適合文書１を利用した適合率（上位 30 件）[%]※	17	27	47
適合文書２を利用した適合率（上位 10 件）[%]※	50	40	80
適合文書２を利用した適合率（上位 30 件）[%]※	27	30	67
適合文書１を利用した再現率（上位 30 件）[%]※	26	42	74
適合文書２を利用した再現率（上位 30 件）[%]※	28	31	69

※値は，小数点第一位以下を四捨五入している．

全体的に 3 人以上を適合文書とした場合の適合率が低いのは，表-3 からわかるように，全適合文書の件数が少ないからである．逆に，再現率は，8 点以上を適合文書としたときより，3 人以上を適合文書とした場合の方が全体的に高い値となっている．これは，一般的に再現率は，全適合文書の数が少ないと高い値になる傾向があるためである．

「java」に関しては，全体的にセレクトブクマにおいてもっとも高い適合率となっている．また，再現率も人数で適合文書を作成した場合，得点から適合文書を作った場合の両方において，セレクトブクマでもっとも高い値となった．

「健康」に関しても，「java」と類似した結果となり，全体的にセレクトブクマにおいてもっとも高い適合率となっている．さらに，高い順位においてその傾向が強い．また，再現率も人数で適合文書を作った場合，得点から適合文書を作成した場合の両方において，セレクトブクマでもっとも高い値となった．

「映画」に関しては，「java」，「健康」とはまったく異なる結果となり，Google 検索においてもっとも高い適合率となった．また，再現率も人数で適合文書を作成した場合，得点から適合文書を作った場合の両方において，Google 検索でもっとも高い値となった．

７．考察と課題，展望

７．１．考察

７．１．１．実験結果からの考察

実験結果より，セレクトブクマにおいて，「java」，「健康」という単語に関しては，Google 検索やブックマーク数順に並べたものと比較して，もっとも高い適合率，再現率を得ることができた．このことから，セレクトブクマにおいて，今後も見たいような長期的に利用される Web ページや Web サービスを手軽に収集できる可能性が高いと考えられる．だが，逆に「映画」という単語においては，Google 検索と比較して適合率，再現率は大幅に低い値となった．この原因の１つとして，Google などの一般の検索エンジンでは，権威の高い Web ページが検索結果に出やすいという特性の影響が考えられる．「映画」という分野においては，そういった大手企業の Web サイトが多数あり，内容も充実しているため，被験者がそういった Web サイトを選択することが多かった．個人で作成している映画関連の Web サイトも，いくつか有益なレビューサイトやまとめサイトなどが存在するが，その数自体が少ない．そのため，「映画」という単語においては，Google 検索において，高い適合率・再現率となり，

セレクトブクマで低い適合率・再現率となったと考えられる.

７．１．２. 時間情報を利用することの有効性

セレクトブクマとブックマーク数順を比較した場合，セレクトブクマにおいてある程度高い適合率・再現率を得ることができた．また，特に上位 10 件の適合率が，ブックマーク数順と比較してセレクトブクマで高い値になっていることからも時間情報を利用することによって，一時的に利用される情報をフィルタリングできている可能性が高いと考えられる.

今回被験者実験をおこなった検索単語は，技術系，生活系，娯楽系と 3 つの異なった分野から 1 つずつ，できるだけ一般的な単語を選んだ．他の単語で検索した場合，もっとセレクトブクマとブックマーク数順で検索結果に相違がでるものも多い．例えば，「ui」で検索した場合は，実験をおこなった 3 つの単語と比較して，検索結果の相違が大きいため，時間情報による影響がもっと強いといえる.

さらに，現状セレクトブクマのランキングロジックでは，ブックマーク数順は，日付にかける係数αの値が 0 のときと同義である．今回の実験は，α=1 として実験をおこなったが，αの値をもっと大きくすれば，時間の影響が強くなり，セレクトブクマとブックマーク数順の検索結果の違いを大きくすることができる．αの値を変化させ，実験・評価をおこなうと時間情報を利用する効果についてもっと明確にできると考えられる.

７．１．３. 本研究の有効性

検索結果や実験結果から，Google 検索と比較してセレクトブクマでは，権威ある Web サイトだけでユーザの満足が得られない分野において，有効性が高いと考えられる．個人で作成している Web ページは，数多くあり玉石混合である. そのため, 既存の PageRank などの手法では, Web コンテンツ作成者がリンクを貼った場合に, PageRank があがるため，コンテンツ作成者側しか Web ページの評価をすることができない．また，SEO 対策がしっかりとなされている，権威ある Web サイトが検索結果の上位に表示されやすい傾向もある.

これに対して，ソーシャルブックマークを利用した場合，コンテンツ消費者がブックマークをするという簡単な行為によって，コンテンツが評価される．ブックマークされた Web ページの中で長い間多くのユーザからブックマークされるという指標を利用することにより，玉石混合の Web ページ群の中から，特に有益で長期間利用できる Web ページや Web サービスを発見するのに有効であると考えられる.

７．２. 課題と展望

本研究の課題として現在収集したはてなブックマークのデータ量不足とデータの偏りの問題がある．はてなブックマークのデータを今後も収集し続けることによって，データ量や利用できる時間の期間が増えていくため，今後この課題は軽減していくと思われる．また，他のソーシャルブックマークデータやアクセスログデータを利用することも可能である.

また，セレクトブクマのランキングロジックは，ブックマーク数にブックマークされた日数で重み付けをする手法を用いている．このような時間情報を利用したランキングの手法としては，本手法以外にも様々な手法があり，手法ごとの特性と利用法を確認する必要がある.

さらに，セレクトブクマでは，ユーザが付与したタグデータを利用しているため，タグの揺らぎの問題が発生している．タグの揺らぎの問題を解決するために同義語を収集する辞書を作るなどの対策を行うことによって，検索精度があがる可能性が高い.

今回の実験では，セレクトブクマがインターネットのライトユーザかコアユーザかなど，どういったユーザ層にとって有用なのか明確にはできなかった．セレクトブクマは，既存の検索エンジンと利用シーンや利用目的が少し異なるため，どういったユーザ層がどういったシーンで利用すると有用であるか，明確にする実験を行う必要がある.

筆者は，短期的に必要とされる情報と長期的に必要とされる情報というテーマは，ソーシャルブックマークだけに関わらず，もっと大きなテーマとして見て，非常に興味深いテーマであると考える．例えば，文献，アプリケーション，ファイル，メモなどへのアクセス情報に本研究の時間理論を適用することで興味深い研究結果を得ることができるかもしれない.

８．おわりに

近年 Web 上のコンテンツは多種多様になってきており，一般の検索エンジンでは，Web 上に存在する長期間利用する Web ページや Web サービスを手軽に取得することが容易ではなくなってきている．特に個人が作成しているような Web ページは大量に存在し，その内容も玉石混合であるが，その中から玉にあたる優れた Web ページを手軽に発見することは，容易ではない．そのため，Web 上の情報フィルタリングの重要性がますます問われるようになってきている．

本研究では，長い期間多くのユーザからアクセスされたりブックマークされたりする情報は，いつ見ても有用な情報である可能性が高いという点に着目した．この特性をソーシャルブックマークの時間情報に適用し，Web 上の情報収集支援システム「セレクトブクマ」を提案し，実装・評価をおこなった．その結果，セレクトブクマは，玉石混合の Web ページの中から権威が高くはないが，有益な Web ページを発見・収集するために有効であることがわかった．また，時間情報を利用することによって，一時的に利用される Web ページの優先度を下げることできる可能性があることがわかった．

本研究により，長期間利用する Web ページを発見するためのひとつの情報フィルタリング手法として，時間情報を利用することの有効性を示すことができた．また，時間情報を利用した情報フィルタリングは，さまざまな情報検索手法や情報フィルタリング手法を組み合わせて利用することも可能であると考えられる．さらに，今後も Web ページは急激な勢いで増加していくことが予想されるため，時間情報を利用した Web ページのフィルタリングは，今後ますます有効性が高まっていく可能性があると考えている．

［参考文献］

[1] Page L, Brin S, Motwaniand R, and Winograd T. The pagerank citation ranking: Bringing order to the web. In *Technical Report, Stanford Digital Library Technologies Project*, 1998.

[2] Twitter. Twitter. https://twitter.com/, （参照 2010-01-11）.

[3] 株式会社はてな. はてなブックマーク. http://b.hatena.ne.jp/, （参照 2010-01-11）.

[4] 株式会社ライブドア. livedoor クリップ- ソーシャルブックマーク. http://clip.livedoor.com/, （参照 2010-01-11）.

[5] ヤフー株式会社. Yahoo!ブックマーク. http://bookmarks.yahoo.co.jp/all, （参照 2010-01-11）.

[6] Yahoo! Delicious. http://delicious.com/, （参照 2010-01-11）.

[7] Digg. Digg. http://digg.com/, （参照 2010-01-11）.

[8] S.A. Golder and B.A. Huberman. The structure of collaborative tagging system. In *Information Dynamics Laboratory, HP Labs*, 2008.

[9] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? In *WSDM'08*, 2008.

[10] 川中翔, 佐藤周行. ソーシャルブックマークにおけるタグの派生関係の抽出. 東京大学修士論文, 2009.

[11] Shengliang Xu, Shenghua Bao, Ben Fei, Zhong Su, and Yong Yu. Exploring folksonomy for personalized search. In *Proc. 31st ACM SIGIR*. pp.155-162, 2008.

[12] Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka. Can social bookmarking enhance search in the web? In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries(JCDL 2007)*, pp.107-116, 2007.

[13] 高橋翼, 北川博之. ソーシャルブックマークによる情報の鮮度を考慮した Web ページ評価手法. Web とデータベースに関するフォーラム *(WebDB Forum 2008)*, 2008.

[14] S. Niwa, T. Doi, and S. Honiden. Web page recommender system based on folksonomy mining. In *Proc. 3rd International Conference on Information Technology：New Generations(ITNG '06)*, pp.388-393, 2006.

[15] 佐々木祥, 宮田高道, 稲積泰宏, 小林亜樹, 酒井善則. Social bookmark におけるコンテンツクラスタ間の類似度を用いた web コンテンツ推薦システム. 情報処理学会論文誌データベース 48, pp.14-27, 2007.

[16] 白土慧, 吉井伸一郎, 古川正志. ソーシャルブックマークサービスを利用した情報レコメンデーション. 情報処理学会研究報告. ICS, 知能と複雑系, pp.15-20, 2006.

[17] 大力慶祐, 大向一輝, 武田英明. ソーシャルブックマークにおけるイノベータに注目した情報推薦手法の提案. 人工知能学会第 22 回全国大会(JSAI2008), 2008.