第1章　長期的なWebページの検索

概要

本章では，長期的なWebページを検索するシステム「セレクトブクマ」について述べる。

1.1. 背景

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 上に存在する体系だった知識や有益なコンテンツを手軽に取得することが容易ではなくなってきている．これに対して，筆者は Web 上に存在する多種多様で玉石混合の Web コンテンツの中でも，いつ見ても有用な体系だった知識を得られる Web ページや有益なコンテンツは，長い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と考えた．

そこで，本研究では膨大で多種多様な情報の中から手軽に有益なコンテンツを取得するために，皆のブックマーク情報を共有できるソーシャルブックマークのデータの活用方法を考案した．本研究ではまず，ソーシャルブックマークデータを分析した．データの分析に基づき，一時期に限りブックマークされる Web ページは，一時的に必要とされる種類の Web ページが多いのに対し，長い間多くのユーザからブックマークされ続ける Web ページは，いつ見ても有用な情報を得られる種類の Web ページが多いことを示した．この特性に基づき，効率よく体系だった知識の得られる Web ページや有益な Web コンテンツを発見・収集する情報収集支援システム「セレクトブクマ」を提案，実装し，評価実験を行った．

1.2. 分析

1.2.1. ソーシャルブックマークデータの分析

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 [?] 図 1.1 のデータを収集した．はてなブックマークはユーザ数が約 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
- ブックマークしたユーザ ID
- ブックマークした日時
- ブックマークしたユーザが付与したタグ名

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，以下の表 1.1 に示す．

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．

1.2.2. 時間情報に関する分析

ユーザからいつ，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，大まかに分けて次の 3 種類のタイプの Web ページがあることが分かった．

1. 一時的にブックマークされ，その後ほとんどブックマークされなくなるタイプのページ（図 1.2）

表 1.1: 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

図 1.1: はてなブックマーク

2. 一時的に大量にブックマークされ，その後も長い間ブックマークされ続けるタイプのページ（図 1.3）

3. 大量にブックマークされる時期はないが，長い間ブックマークされ続けるタイプのページ（図 1.4）

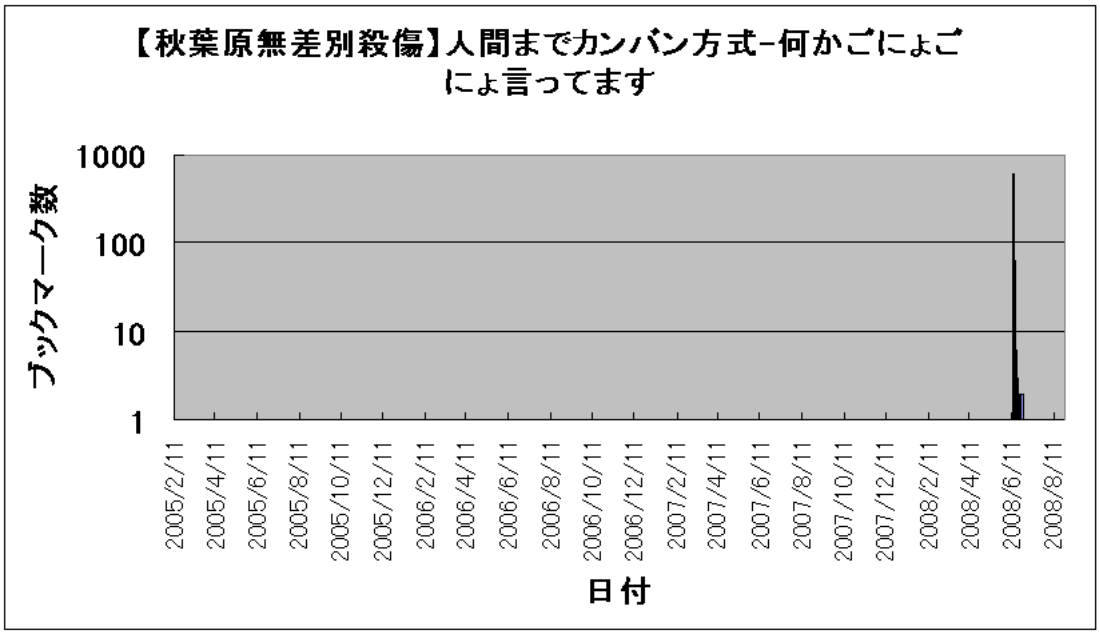

図 1.2: 一時的にブックマークされ，その後ほとんどブックマークされなくなるタイプのページ

以上の 3 種類のタイプの Web ページをさらに，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．

- Type1：一時期しかユーザからブックマークされないページ
- Type2：長い間ユーザからブックマークされ続けるページ

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 1.5 と図 1.6 に示す．分析対象のページは，以下の条件とした．

- Type1 は，全日数/全ブックマーク数=0.2 以下
- Type2 は，全日数/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

ここで全日数とは，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 日にそれぞれ異なったユーザからブックマークされた場合，3 日とする．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．

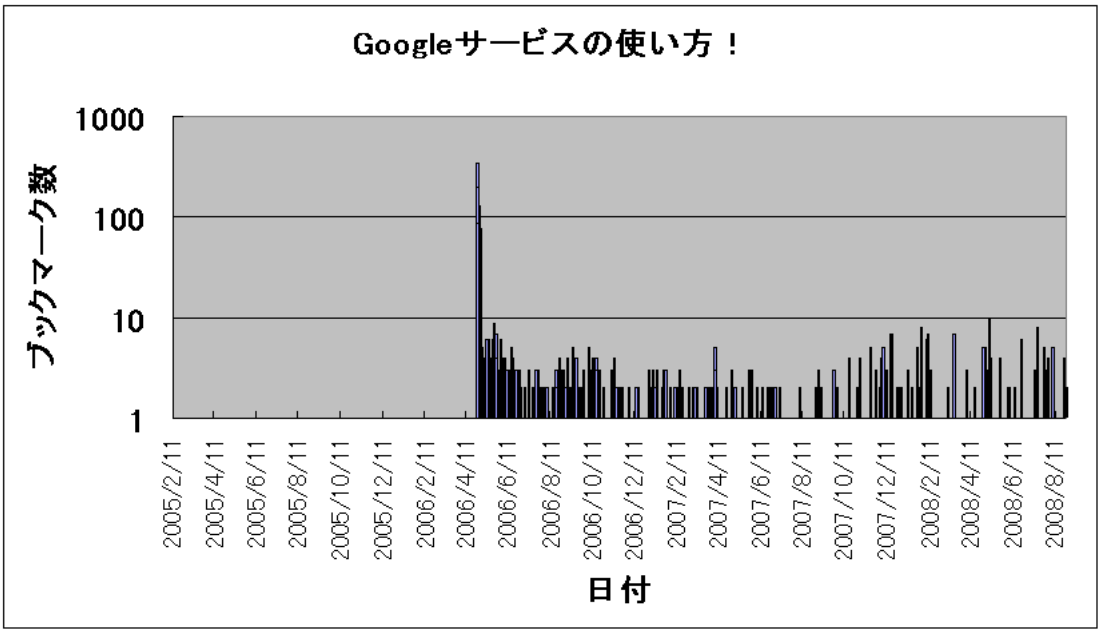

図 1.3: 一時的に大量にブックマークされ，その後も長い間ブックマークされ続けるタイプのページ

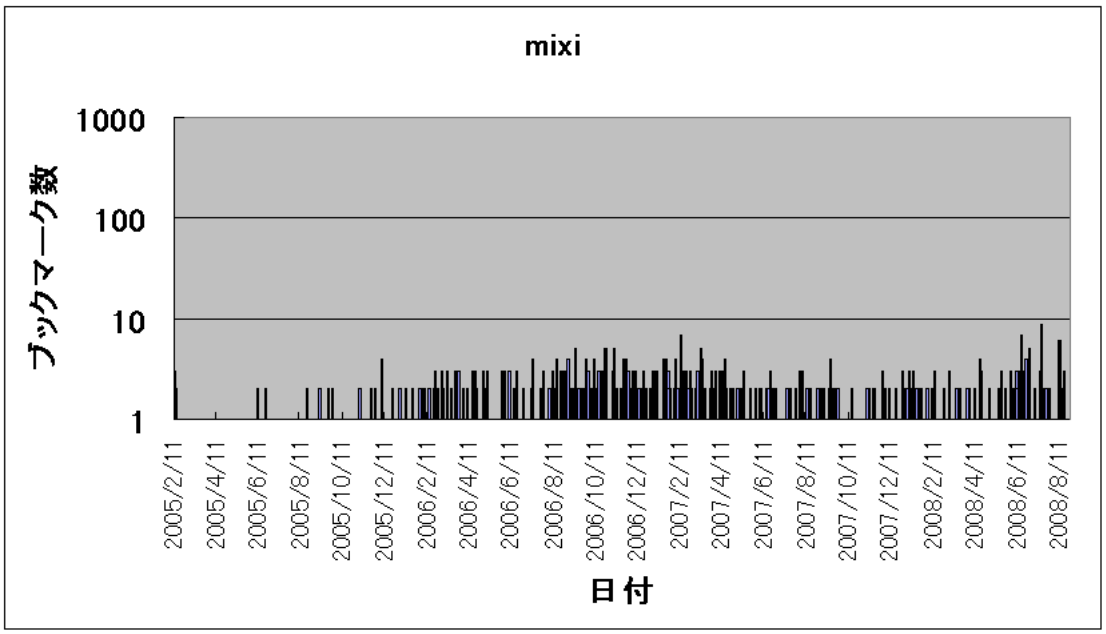

図 1.4: 大量にブックマークされる時期はないが，長い間ブックマークされ続けるタイプのページ

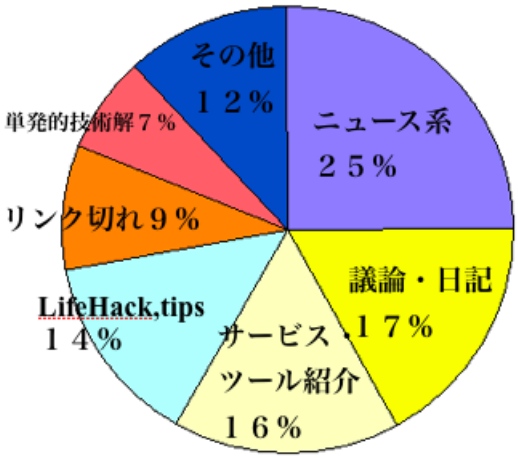

図 1.5: Type1 の Web ページの種類

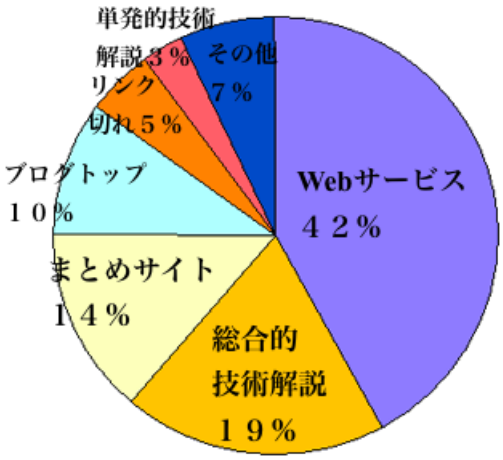

図 1.6: Type2 の Web ページの種類

図 1.5，図 1.6 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とが分かった．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わかった．

1.3. セレクトブクマの提案

1.3.1. 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，手軽に体系だった知識や有用な Web サービスを収集できる．

1.3.2. セレクトブクマの設計思想

セレクトブクマでは，特に情報収集の手軽さを重視している．また，調べたい分野に対して，以下 2 つのことを目的としている．

- 体系だった知識を得ること
- 有用な Web サービスを発見すること

情報を収集する際に，上記以外の Web ページが表示されないように，情報フィルタリングを行うことに注力している．

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，ユーザの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することによって，一時的にしか利用しない Web ページをフィルタリングすることができる．

1.3.3. セレクトブクマ画面構成

セレクトブクマの検索前の画面と検索後の画面を図 1.7 に示す．

図 1.7 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．

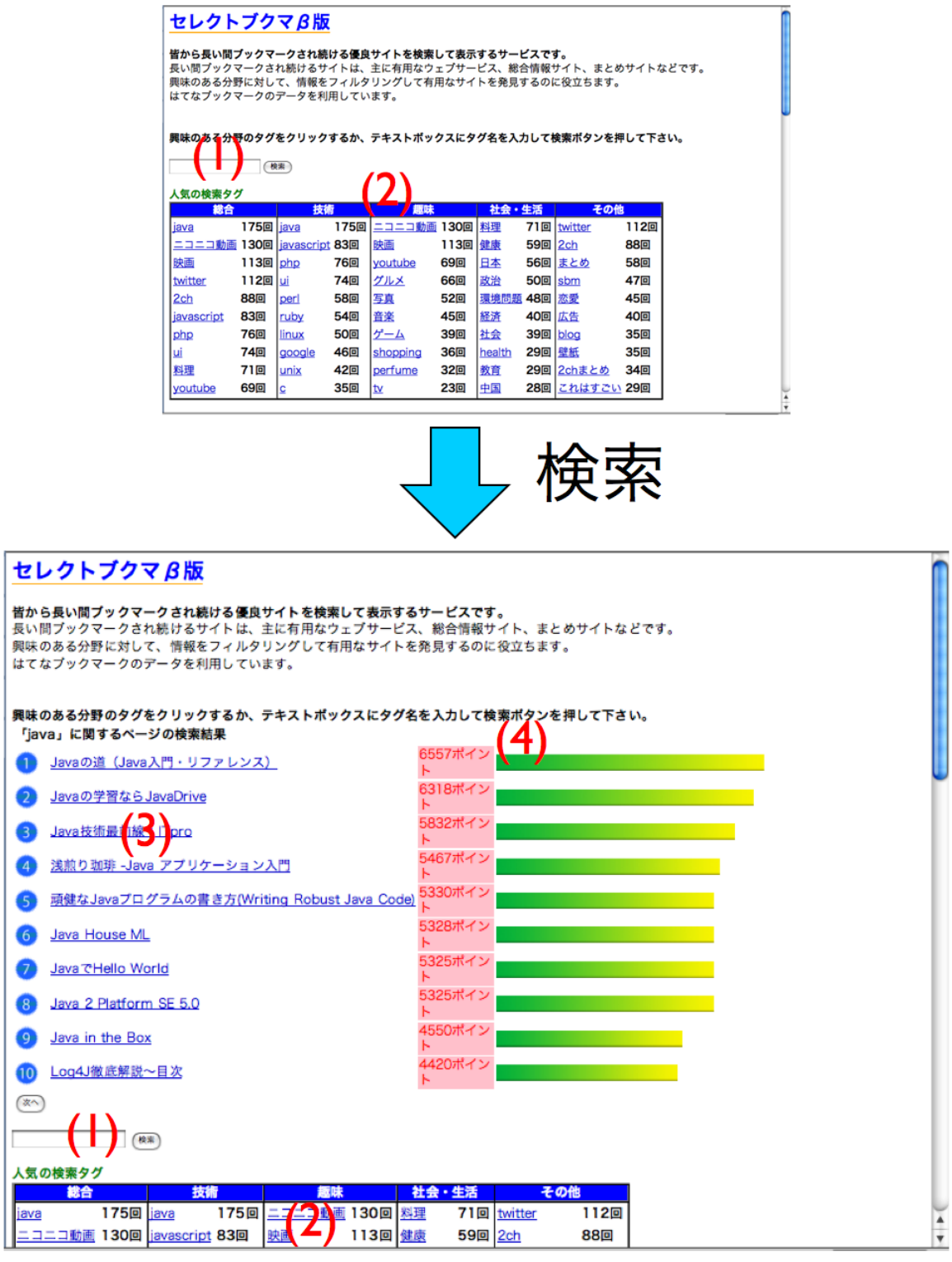

図 1.7: セレクトブクマ画面

(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．

1.3.4. セレクトブクマインタフェース

セレクトブクマの基本的なインタフェースは，検索窓になにか単語を入力し，検索ボタンを押すことによって検索をおこなう．検索結果は，1 ページに上位 10 件表示され,「次へ」ボタンを押すと，11 位〜20 位までが表示される．このあたりは，一般的な検索エンジンと同じである．
上記に加えて，補助的な機能として，セレクトブクマを利用したユーザの検索回数の多い単語をカテゴリごとに表示し，そのリンクをクリックすることによっても検索可能としている．セレクトブクマは，検索エンジンというより，情報収集支援システムという位置付けで，運用をしている．ユーザには，ある分野ごとに情報を収集するという目的で利用してもらう想定である．そのため，検索単語の選び方が少し一般的な検索エンジンと異なり，検索単語が複数例示されていると，新規ユーザにとっても取り付きやすいのではないかと考えた．
検索回数の多い単語は，興味のある分野を見つけやすいように,「技術」,「趣味」,「社会・生活」,「その他」の 4 分野にカテゴリわけを行っている．さらに，すべてのカテゴリを含めた検索回数に対して順位を付け、それを「総合」として表示している．

1.3.5. 検索ランキングロジック

本研究では，検索結果のランキング手法について，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，以下に示す手法が現状ではもっとも効果的であった．
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．

$$Bookmarks \times Days^{\alpha} \tag{1.1}$$

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：任意の係数

1.3.6. システム構成

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以下の図 1.8 に示す．

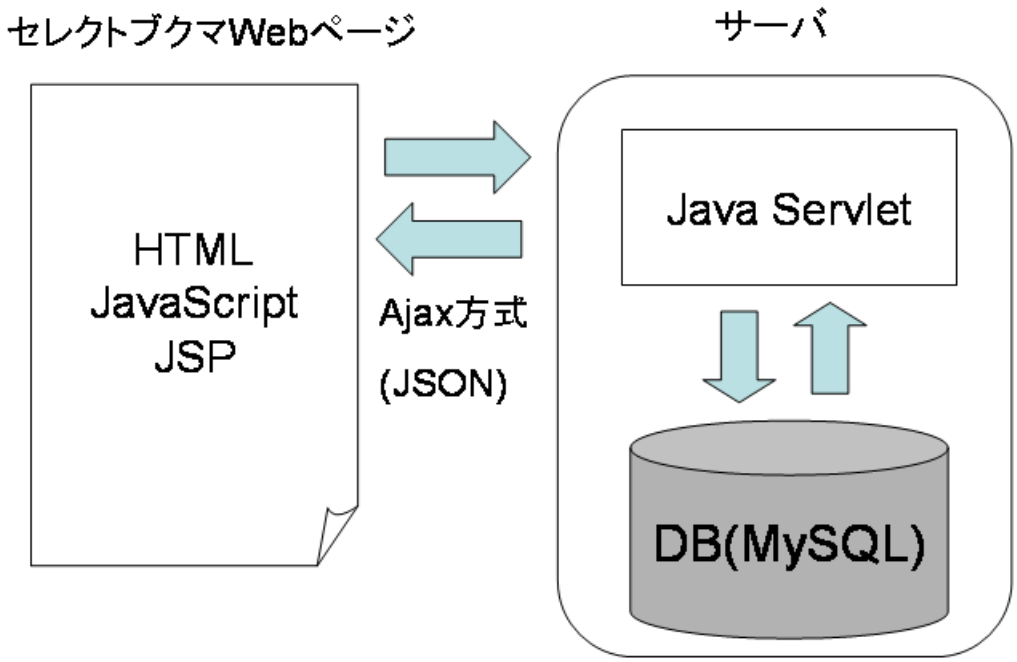

図 1.8: システム構成図

また，DB のテーブル構造は，以下の図 1.9 に示す．
http://plazman.chi.mag.keio.ac.jp/sbm/summary.jsp

テーブル 1

カラム名	id	url
説明	ID	URL
例	1000000	http://xxx.xxx

テーブル 2

カラム名	url	title	count	user	time	tag
説明	URL	タイトル	ブックマーク数	ブックマークしたユーザのID	ブックマークされた時間	ブックマークしたタグ名
例	http://xxx.xxx	ページA	1232	wilfue	20070730	java

図 1.9: DB のテーブル構成図

1.4. 評価実験

1.4.1. 実験概要

セレクトブクマの有用性を評価するため，以下 3 種類の手法において比較実験を行った．

- 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（α=1）

1.4.2. 実験目的

本実験の目的としては，本研究の目的である，Web 上から以下の情報を手軽に取得できるかどうかについて評価することである．

- 体系だった知識を得ること
- 有益な Web コンテンツを発見すること

そのため，以下のようなことを目的とした実験をおこなった．

- セレクトブクマと既存の検索エンジンにおいて，どちらがより今後も利用するような体系だった知識を得られる Web ページを手軽に取得できるか評価する
- 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をフィルタリングできていることを実証する

1.4.3. 実験方法

以下，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 サイトの中で重複したものを除いた Web サイトを、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 サイトを 1 位〜10 位まで選んでもらう．

- Google 検索
- ブックマーク数順
- セレクトブクマ（α=1）

被験者が選んだ Web サイトを適合文書として，適合率 [*8]，再現率 [*9] を求める．
実験条件は以下とする．

- 被験者数：30 人（1 単語につき 10 人× 3 単語）
- 検索単語（タグ）:「java」,「健康」,「映画」
- 「java」: 男性 6 人，女性 4 人
- 「健康」: 男性 5 人，女性 5 人
- 「映画」: 男性 3 人，女性 7 人

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，特別はてなブックマークデータを利用した場合に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．被験者の負担を軽減するため，1 人 1 単語を目安に実験をおこなった．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．

1.4.4. 実験結果

各検索手法による検索結果上位 10 件

Google 検索とブックマーク数順とセレクトブクマでの検索結果上位 10 件のタイトルを以下の表 1.2〜1.4 に示す．すべての検索結果上位 30 件のタイトルと URL は，付録 A に示す．

検索結果を見ると，Google 検索の場合は，大手企業が作成した 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．

被験者が選んだ Web ページ

Google 検索、ブックマーク数順、セレクトブクマの検索結果から，被験者が今後も利用したいと思う Web サイトを 1 位〜10 位までを被験者に選んでもらった結果を以下の表 1.5〜表 1.10 に示す.「java」,「健康」,「映画」という各単語に対して，以下 2 種類の結果を示す.

1. 人数順
   被験者が何人が選んだか人数順に示したものである．3 人以上選んだ Web ページのみ記載する．

表 1.2: 「java」で検索した場合の検索結果上位 10 件

順位	セレクトブクマ	ブックマーク数順	Google 検索
1	Java の道 (Java 入門: リファレンス)	Java のクラスアンロード (Class Unloading)	java.com: あなたと Java
2	Java の学習なら、JavaDrive	Java の道（Java 入門・リファレンス）	無料 Java ソフトウェアをダウンロード - Sun Microsystems
3	Java 技術最前線：ITpro	頑健な Java プログラムの書き方 (Writing Robust Java Code)	Java - Wikipedia
4	浅煎り珈琲-Java アプリケーション入門	Java 技術最前線：ITpro	Java テクノロジ - サン・マイクロシステムズ
5	頑健な Java プログラムの書き方	Java の学習なら JavaDrive	サン・マイクロシステムズ
6	Java House ML	浅煎り珈琲-Java アプリケーション入門	Java とは - 意味/解説/説明/定義 ： IT 用語辞典
7	Java で Hello World	Java で Hello World	日本 Java ユーザグループ
8	Java 2 Platform SE 5.0	Java 2 Platform SE	Java とは - はてなキーワード
9	Java in the Box	【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	Sun Developer Connection - Java Developer Connection
10	Log4J 徹底解説〜目次	Java House ML	Java の道（Java 入門・リファレンス）

表 1.3: 「健康」で検索した場合の検索結果上位 10 件

順位	セレクトブクマ	ブックマーク数順	Google 検索
1	health クリック｜健康 生活習慣病 サプリメント	ゲンダイネット - 目の疲れをためない３大作戦	健康 - Wikipedia
2	基礎代謝を高めるための９９の技法	体の歪みを治したい:アルファルファモザイク	家庭の医学・健康 - goo ヘルスケア
3	ゲンダイネット - 目の疲れをためない３大作戦	重い、痛いを吹き飛ばせ！肩コリ解消法 — Web 担当者 Forum	健康ネット
4	視力回復マッサージは本当に効き目があるのだ（映像付） — ｉｄｅａ＊ｉｄｅａ	眼精疲労を治すには:アルファルファモザイク	病院情報、家庭の医学、病気の検索、薬の情報、健康情報 - Yahoo ...
5	重い、痛いを吹き飛ばせ！肩コリ解消法 — Web 担当者 Forum	金も時間も掛からない花粉症予防の仕方:アルファルファモザイク	厚生労働省：健康
6	体の歪みを治したい:アルファルファモザイク	health クリック｜健康 生活習慣病 サプリメント	健康 ON-LINE
7	病院検索ならここカラダ	蒸しタオルを使うと酷使した目の疲れが取れる - GIGAZINE	asahi.com (朝日新聞社):医療・健康・ニュース
8	意外に効き目のある視力回復マッサージ	疲れ目：パソコン作業で肩こり、視力低下など眼精疲労　対策は…（上） - 毎日ｊｐ (毎日新聞)	いきいき健康　NIKKEI NET
9	蒸しタオルを使うと酷使した目の疲れが取れる - GIGAZINE	ビジネスマンの不死身力：「夜食は太る」の科学 (1/2) - ITmedia エンタープライズ	[健康管理] All About
10	MouRa｜Net 現代｜巷にはびこる「健康情報」50 のウソ・ホント	基礎代謝を高めるための９９の技法	ケンコーコム - 健康メガショップ

表 1.4: 「映画」で検索した場合の検索結果上位 10 件

順位	セレクトブクマ	ブックマーク数順	Google 検索
1	★前田有一の超映画批評★	★前田有一の超映画批評★	Yahoo!映画 - 映画情報
2	みんなのシネマレビュー	みんなのシネマレビュー	映画情報 - goo 映画
3	あの映画のココがわからない　まとめサイト	あの映画のココがわからない　まとめサイト	映画・DVD と映画館の上映時間を完全網羅 — Movie Walker
4	allcinema ONLINE 映画データベース	allcinema ONLINE 映画データベース	映画のことなら eiga.com
5	映画生活 - 新作映画情報	超映画批評	映画「サマーウォーズ」公式サイト
6	超映画批評	段ボールで『2001 年宇宙の旅』をリメイクできるか? - ベイエリア在住町山智浩アメリカ日記	映画 - Wikipedia
7	CinemaScape?映画批評空間?	映画生活 - 新作映画情報	TOHO シネマズ
8	cinemacafe.net シネマカフェ ?映画ファンによる、映画ファンのための、映画的ライフスタイル・エンターテイメント・サイト?	CinemaScape?映画批評空間?	ワーナー・マイカル・シネマズ－映画館、映画情報、上映スケジュール ...
9	eiga.com	痛いニュース (∀ ‘):もっとも感動した映画トップ 10 発表	大ブーイング！ぐだぐだ最終回「JIN」映画化?TBS 意外な反応 ...
10	いのちの食べかた	404 Blog Not Found:アマとプロとが選んだディストピア映画歴代トップ 26 - 1	新作映画情報「ぴあ映画生活」

2. 得点順
   被験者が選んだ順位によって得点を付与し，得点順に並べたものの 2 種類を示す．得点は，1 位 10 点，2 位 9 点，3 位 8 点 ⋯ 10 位 1 点のようにつけた．得点の上位 30 件のみを記載する．

「java」に関しては，表 1.5，表 1.6,「健康」に関しては，表 1.7，表 1.8,「映画」に関しては，表 1.9，表 1.10，に示す．

表 1.5: 「java」で検索した場合の検索結果の中から被験者が選んだ Web ページ（人数）

人数	**タイトル**	**セレクトブクマ順位**	**ブックマーク数順位**	**Google 検索順位**
7	Java の道（Java 入門・リファレンス）	**1**	**2**	**10**
7	とほほの Java 入門			14
7	Java の学習なら JavaDrive	**2**	**5**	
5	初心者が Java を“ 超高速 ”で学ぶためのコツ:ITpro	12	22	
4	Java 入門	27		
4	TECHSCORE(テックスコア) ?C 言語/JAVA/デザインパターン/CORBA/XML/SQL/UML を基礎から丁寧に解説します?	26		
4	Java FAQ: よくある質問とその回答集			21
4	Java で Hello World	**7**	**7**	13
4	MYCOM ジャーナル - エンタープライズ - コラム - ライトニング Java	20		
3	Java を JavaScript に変換するグーグルのツールを使ってみよう ? @ IT	25		
3	Java -TECHSCORE-			19
3	頑健な Java プログラムの書き方 (Writing Robust Java Code)	**5**	**3**	
3	@ IT：Hibernate で理解する O/R マッピング（1）	30		

表 1.6: 「java」で検索した場合の検索結果の中から被験者が選んだ Web ページ（得点）

得点	タイトル	セレクトブクマ順位	ブックマーク数順位	Google 検索順位
55	とほほの Java 入門			14
45	Java の道（Java 入門・リファレンス）	**1**	**2**	**10**
33	Java の学習なら JavaDrive	**2**	**5**	
27	Java FAQ: よくある質問とその回答集			21
23	Java 入門	27		
20	MYCOM ジャーナル - エンタープライズ - コラム - ライトニング Java	20		
19	Java で Hello World	**7**	**7**	13
19	Java 技術最前線：ITpro	**3**	**4**	
18	Java -TECHSCORE-			19
18	頑健な Java プログラムの書き方 (Writing Robust Java Code)	**5**	**3**	
18	JavaA2Z	22		
15	TECHSCORE(テックスコア) ?C 言語/JAVA/デザインパターン/CORBA/XML/SQL/UML を基礎から丁寧に解説します?	26		
15	ITmedia エンタープライズ：矛盾を抱えつつ進化する“ Java ”??黒船となった Ruby on Rails (1/2)		29	
14	初心者が Java を“ 超高速 ”で学ぶためのコツ:ITpro	12	22	
14	Java 2 Platform SE 5.0	**8**	**8**	
13	Java を JavaScript に変換するグーグルのツールを使ってみよう ? @ IT	25		
13	Java におけるコード進化パターン (Code Evolution Patterns in Java)		14	
12	C/C いっさいなし、Java だけで開発された OS - JNode (MYCOM PC WEB)		19	
11	@ IT：Hibernate で理解する O/R マッピング（1）	30		
10	Java アプレット - Wikipedia			20
9	Java Solution ? @ IT	28		12
9	IBM developerWorks Japan : Resources for Java developers			17
9	Java でゲーム作りますが何か？	16	17	
8	Ruby よりも Java が好きな理由		11	
8	Java in the Box	**9**	13	23
8	Ja-Jakarta Project	21		

表 1.7: 「健康」で検索した場合の検索結果の中から被験者が選んだ Web ページ（人数）

人数	タイトル	セレクトブクマ順位	ブックマーク数順位	Google 検索順位
6	意外に効き目のある視力回復マッサージ	**8**	28	
6	視力検査 - あなたの視力 今いくつ?: 視力回復のアイポータル	30		
5	蒸しタオルを使うと酷使した目の疲れが取れる - GIGAZINE	**9**	**7**	
5	なかなか眠れない人のための簡単に眠る 10 の方法 - GIGAZINE	20		
4	第 1 回　眼の疲れを取る：ITpro	11	12	
3	基礎代謝を高めるための９９の技法	**2**	**10**	
3	視力回復とレーシックのアイポータル	14	16	
3	ビジネスマンの不死身力：「夜食は太る」の科学 (1/2) - ITmedia エンタープライズ		**9**	
3	5 時間以下の睡眠続け死亡率 1.7 倍に　7 時間寝よう｜Ameba News	13	18	
3	視力回復マッサージは本当に効き目があるのだ（映像付） — ｉｄｅａ＊ｉｄｅａ	**4**	21	
3	睡眠時間を記録するサイト — ねむログ	25		
3	なんでも評点:空腹は幸福？？ ストレスで腹が減ったときは何も食べずに我慢した方がストレスに打ち克てる...		19	
3	シゴタノ！ - 睡眠時間を短くする１４のコツ＜前編＞	26		
3	「うつ」にならない、繰り返さない？ @ IT 自分戦略研究所	27		
3	スラッシュドット・ジャパン — 睡眠不足だと仕事がはかどらない理由、科学的に明らかになる		26	
3	姿勢をよくするための運動 - Tech Mom from Silicon Valley	28	13	

表 1.8: 「健康」で検索した場合の検索結果の中から被験者が選んだ Web ページ（得点）

得点	タイトル	セレクトブクマ順位	ブックマーク数順位	Google 検索順位
36	視力検査 - あなたの視力 今いくつ?: 視力回復のアイポータル	30		
33	基礎代謝を高めるための９９の技法	**2**	**10**	
30	意外に効き目のある視力回復マッサージ	**8**	28	
30	なかなか眠れない人のための簡単に眠る 10 の方法 - GIGAZINE	20		
24	蒸しタオルを使うと酷使した目の疲れが取れる - GIGAZINE	**9**	**7**	
23	睡眠時間を記録するサイト — ねむログ	25		
21	第 1 回　眼の疲れを取る：ITpro	11	12	
21	視力回復マッサージは本当に効き目があるのだ（映像付） — ｉｄｅａ＊ｉｄｅａ	**4**	21	
19	Medical Tribune あなたの健康百科			16
18	ビジネスマンの不死身力：「夜食は太る」の科学 (1/2) - ITmedia エンタープライズ		**9**	
18	スラッシュドット・ジャパン — 睡眠不足だと仕事がはかどらない理由、科学的に明らかになる		26	
17	5 時間以下の睡眠続け死亡率 1.7 倍に　7 時間寝よう｜Ameba News	13	18	
17	「うつ」にならない、繰り返さない？ @ IT 自分戦略研究所	27		
14	視力回復とレーシックのアイポータル	14	16	
14	シゴタノ！ - 睡眠時間を短くする１４のコツ＜前編＞	26		
14	精神状態の健康がピンチになった時の３つの対処 - koe だめ		27	
14	重い、痛いを吹き飛ばせ！ 肩コリ解消法 — Web 担当者 Forum	**5**	**3**	
14	割れた腹筋を手に入れるトレーニングを教えてもらいました。 — その他 (ライフ) — とりあえず...	17		
13	[健康管理] All About			19
12	医学都市伝説: 暗いところで本を読んでも目は悪くならない	19	22	
11	asahi.com（朝日新聞社）：医療・健康・ニュース			**7**
10	なんでも評点:空腹は幸福？？ ストレスで腹が減ったときは何も食べずに我慢した方がストレスに打ち克てる...		19	
10	健康 - Wikipedia			**1**
10	病院情報、家庭の医学、病気の検索、薬の情報、健康情報 - Yahoo ...			**4**
10	PC で眼が疲れない方法 - 萌え理論 Blog	12	17	
9	指をポキポキ鳴らすと太くなるって本当？		20	
9	姿勢をよくするための運動 - Tech Mom from Silicon Valley	28	13	
9	NHK 健康ホームページ：トップページ			12
8	health クリック｜健康 生活習慣病 サプリメント	**1**	**6**	
8	体の歪みを治したい:アルファルファモザイク	**6**	**2**	

表 1.9: 「映画」で検索した場合の検索結果の中から被験者が選んだ Web ページ（人数）

人数	タイトル	セレクトブクマ順位	ブックマーク数順位	Google 検索順位
5	Yahoo!映画 - 映画情報			**1**
4	映画情報 - goo 映画			**2**
4	シネマぴあ			26
4	TOHO シネマズ			**7**
4	映画、映画館検索 - TSUTAYA online			16
4	@nifty 映画 - 映画情報			27
4	ワーナー・マイカル・シネマズ－映画館、映画情報、上映スケジュール ...			**8**
3	あの映画のココがわからない　まとめサイト	**3**	**3**	
3	【2ch】ニュー速クオリティ:一生のうちに一度は見ておくべき映画		11	
3	もっと知られていい映画:アルファルファモザイク		14	
3	★前田有一の超映画批評★	**1**	**1**	25
3	みんなのシネマレビュー	**2**	**2**	
3	シネマスクランブル 映画予告編・映画ランキング・映画上映 ...			13
3	映画なら GyaO! 【映画】 無料映画の映像視聴 | GyaO![ギャオ] 映画			30
3	【2ch】ニュー速クオリティ:観るまでバカにしてたのに観たら面白かった映画		27	
3	allcinema ONLINE 映画データベース	**4**	**4**	12
3	eiga.com	**9**	18	**4**
3	109 シネマズ公式ホームページ			24
3	映画・DVD と映画館の上映時間を完全網羅 — Movie Walker			**3**

表 1.10: 「映画」で検索した場合の検索結果の中から被験者が選んだ Web ページ（得点）

得点	タイトル	セレクトブクマ順位	ブックマーク数順位	Google 検索順位
29	Yahoo!映画 - 映画情報			**1**
28	ワーナー・マイカル・シネマズ−映画館、映画情報、上映スケジュール ...			**8**
26	TOHO シネマズ			**7**
25	映画、映画館検索 - TSUTAYA online			16
25	【2ch】ニュー速クオリティ:観るまでバカにしてたのに観たら面白かった映画		27	
23	あの映画のココがわからない　まとめサイト	**3**	**3**	
21	allcinema ONLINE 映画データベース	**4**	**4**	12
20	109 シネマズ公式ホームページ			24
19	映画・DVD と映画館の上映時間を完全網羅 — Movie Walker			**3**
18	eiga.com	**9**	18	**4**
18	著作権の切れた過去の名作映画やドキュメンタリーをダウンロード - GIGAZINE	27		
17	みんなのシネマレビュー	**2**	**2**	
17	痛いニュース (∀ ‘):大人が選ぶ泣ける洋画ベスト 30　1「タイタニック」、2「アルマゲドン」		28	
17	映画、ビデオ - Yahoo!カテゴリ			17
16	シネマぴあ			26
15	@nifty 映画 - 映画情報			27
15	【2ch】ニュー速クオリティ:一生のうちに一度は見ておくべき映画		11	
15	映画「サマーウォーズ」公式サイト			**5**
15	映画館・シネコンの【ムービックス-MOVIX】			11
14	映画なら GyaO! 【映画】 無料映画の映像視聴 | GyaO![ギャオ] 映画			30
12	映画情報 - goo 映画			**2**
12	もっと知られていい映画:アルファルファモザイク		14	
12	シネマスクランブル 映画予告編・映画ランキング・映画上映 ...			13
12	無料映画館			28
11	新作映画情報「ぴあ映画生活」			**10**
10	★前田有一の超映画批評★	**1**	**1**	25
10	『アバター』は映画の未来に iPhone 登場なみの衝撃を与える			19
9	時をかける少女	12		
8	flowerwild.net ー 蓮實重彦インタビュー——リアルタイム批評のすすめ vol.1	25		

適合率・再現率による比較

各手法においての，適合率[*8]・再現率[*9] を比較した．適合率[*8]・再現率[*9] の意味に関しては，1章の用語定義に示した．

適合率・再現率を出すために必要な適合文書は，以下 2 種類作成した．

1. 3 人以上の被験者が上位 10 位以内に選択した Web ページ
2. 被験者が選んだ Web ページの中で，得点の合計値が 8 点以上の Web ページ

全適合文書の件数を以下の表 1.11 に示す．

表 1.11: 全適合文書の件数

単語名	java	健康	映画
人数 [件]	13	16	19
得点 [件]	26	30	29

人数から全適合文書を作成した場合の適合率を図 1.10，図 1.12，図 1.14 に，得点から全適合文書を作成した場合の適合率を図 1.11，図 1.13，図 1.15 に示す．また，人数から全適合文書を作成した場合と得点から全適合文書を作成した場合の再現率を表 1.12〜表 1.14 に示す．適合率の値は，セレクトブクマ，ブックマーク数順，Google 検索での検索結果 1 位〜30 位までに対する以下の値を示している．

$$\text{適合率} = \frac{\text{適合文書の数}}{\text{検索結果の文書の数}}$$

ここでいう検索結果の文書の数とは, 1 位の場合 1, 2 位の場合 2,..., 30 位の場合 30 となる．

また，再現率の値は，以下の値を示している．

$$\text{再現率} = \frac{\text{検索結果中の適合文書の数}}{\text{全適合文書の数}}$$

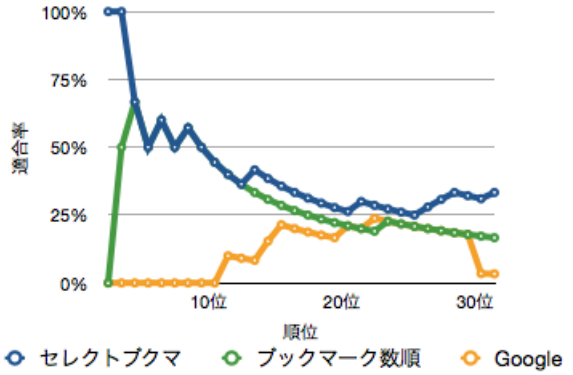

図 1.10: 3 人以上を適合文書とした場合の適合率 (java)

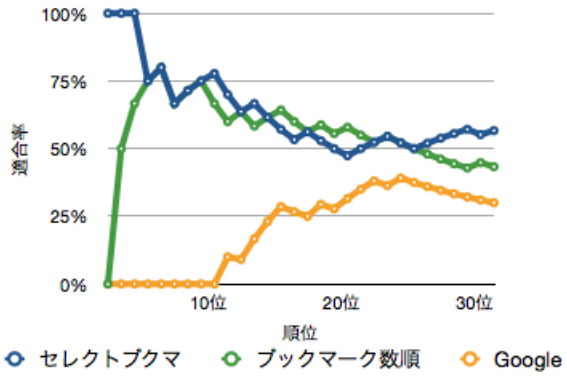

図 1.11: 8 点以上を適合文書とした場合の適合率 (java)

表 1.12: 再現率 (java)

	セレクトブクマ	**ブックマーク数順**	**Google 検索**
再現率※（人数）	77 %	38 %	38 %
再現率※（得点）	65 %	50 %	35 %

※再現率の値は，小数点第一位以下を四捨五入している．

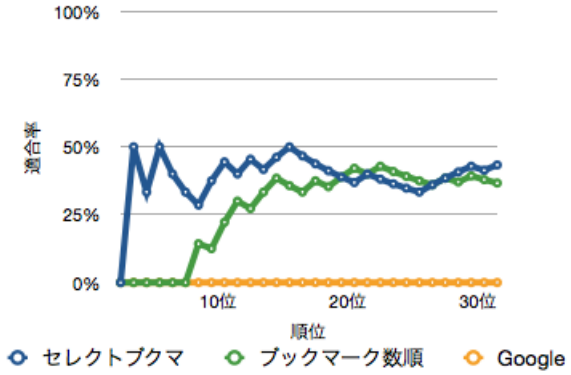

図 1.12: 3 人以上を適合文書とした場合の適合率 (健康)

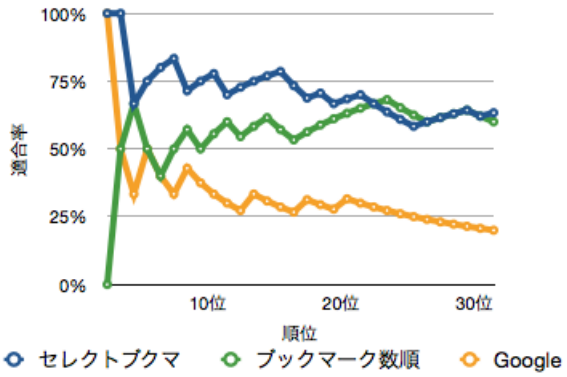

図 1.13: 8 点以上を適合文書とした場合の適合率 (健康)

表 1.13: 再現率 (健康)

	セレクトブクマ	**ブックマーク数順**	**Google 検索**
再現率※（人数）	81 %	69 %	0 %
再現率※（得点）	63 %	60 %	20 %

※再現率の値は，小数点第一位以下を四捨五入している．

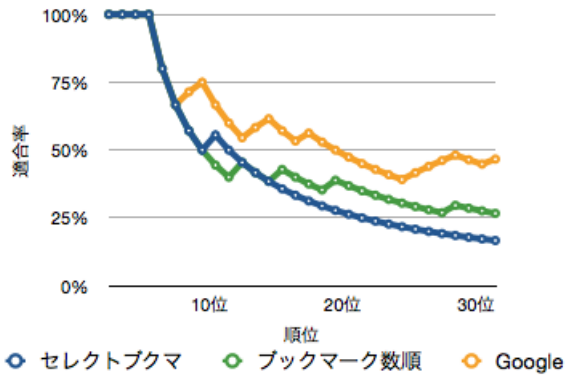

図 1.14: 3 人以上を適合文書とした場合の適合率 (映画)

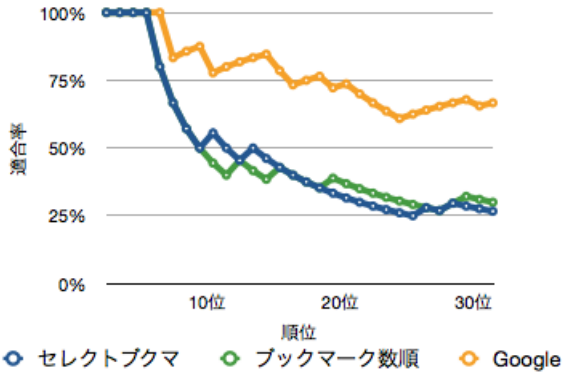

図 1.15: 8 点以上を適合文書とした場合の適合率 (映画)

表 1.14: 再現率 (映画)

	セレクトブクマ	**ブックマーク数順**	**Google 検索**
再現率※（人数）	26 %	42 %	74 %
再現率※（得点）	28 %	31 %	69 %

※再現率の値は，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 1.11 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が，特に高い順位においてその傾向は強い．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．

「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った場合，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．

得点の合計値による比較

被験者が選んだ Web ページには，順位によって得点がついている．その得点がセレクトブクマ，ブックマーク数順，Google 検索の検索結果の Web ページに合計何点入っているか，比較する．セレクトブクマ，ブックマーク数順，Google 検索において，検索結果上位 10 件，上位 20 件，上位 30 件の 3 つの得点の合計値を以下の表 1.15〜表 1.17 に示す．

表 1.15: 「java」での検索結果の得点の合計値

	セレクトブクマ	**ブックマーク数順**	**Google 検索**
上位 10 件の合計値 [点]	162	152	23
上位 20 件の合計値 [点]	225	222	148
上位 30 件の合計値 [点]	322	279	195

表 1.16: 「健康」での検索結果の得点の合計値

	セレクトブクマ	**ブックマーク数順**	**Google 検索**
上位 10 件の合計値 [点]	147	119	48
上位 20 件の合計値 [点]	269	208	92
上位 30 件の合計値 [点]	378	311	95

表 1.17: 「映画」での検索結果の得点の合計値

	セレクトブクマ	**ブックマーク数順**	**Google 検索**
上位 10 件の合計値 [点]	89	77	163
上位 20 件の合計値 [点]	114	125	276
上位 30 件の合計値 [点]	153	169	371

結果を見ると「java」と「健康」においては，セレクトブクマで上位 10 件，上位 20 件，上位 30 すべてにおいて，もっとも高い合計得点となった．これに対して,「映画」においては，Google 検索で上位 10 件，上位 20 件，上位 30 すべてにおいて，もっとも高い合計得点となった．

1.4.5. 時間情報を利用する効果について

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，特別セレクトブクマが有利になるような単語を選ばずに，できるだけ一般的な単語を選んだ．そのため,「java」,「健康」,「映画」という単語を選ぶこととなったが，これらの単語はセレクトブクマとブックマーク数順の結果において，上位 10 件の中だけ見ても，その順位に変化はあるが，半数程度同一の結果が含まれる．

これに対して、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，上位 10 件の中に 3 件しか同一の結果が含まれない．そのため,「ui」の場合は，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は，以下のようになっている．

$$Bookmarks \times Days^{\alpha} \tag{1.2}$$

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：任意の係数
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．

1.4.6. 各実験結果からの考察

適合率からの考察

「java」で検索した場合，興味深いのは，Google 検索の検索結果の順位が下がるにつれて，適合率が上昇傾向にあることである．このことから「java」という分野においては，Google 検索の検索結果の上位が，一時的にしか利用しない Web ページが多く，ユーザにとって今後も利用したいような Web ページが少ないということがわかる．この傾向は，他の一般的な検索エンジンを用いて「java」で検索しても Google 検索と類似した結果が出るため，一般的な検索エンジン全体にいえるのではないかと考えられる．このため，少なくとも「java」という分野においては，一般的な検索エンジンが体系だった知識を得られる Web ページや有益な Web サービスを手軽に発見・収集するためには，向かないのではないかと考えられる．

これに対して,「java」で検索した場合，セレクトブクマでもっとも高い適合率が得られており，さらに高い順位程高い適合率が得られている．そのため，セレクトブクマが体系だった知識を得られる Web ページや有益な Web サービスを手軽に発見するために有用であることを確認できた．ブックマーク数順と比較しても，高い順位程適合率の差が大きい傾向にあるため，時間情報を利用することによるフィルタリング効果も確認できた．

「健康」や「映画」で検索した場合は，Google 検索の検索結果の順位が高い程，適合率が高い傾向にあるので，検索結果上位の方に今後も利用したいような Web ページが多いことがわかる．特に「映画」といったような，大手企業が Web サイトを作ることにより利益を見込めるような分野では，多くの大手企業が Web サイトや Web サービスを作成し，内容も充実している場合が多い．こういった Web ページが Google 検索において上位にランキングしたため,「映画」における適合率が高い値になったのではないかと考えられる．そのため，このような分野においては，Google 検索で有益な Web コンテンツを手軽に発見することができると考えられる．

セレクトブクマにおいても，同様に検索結果の順位が高い程，高い適合率が得られているため，検索結果上位に今後も利用したいような Web ページが多いことがわかる．ただし，Google 検索とは，検索結果の傾向が異なり，検索結果には大手企業が作成したような Web サイトはほとんど出てこない．そのため，個人で作成したような Web サイトの中から，体系だった知識を得られる Web ページや有益な Web サービスを発見したい場合，セレクトブクマが有効なのではないかと考えられる．

再現率，得点の合計値からの考察

「java」,「健康」においては，人数から適合文書を作成した場合も得点から適合文書を作成した場合も，セレクトブクマの再現率がもっとも高い．さらに，得点の合計値も上位 10 件，上位 20 件，上位 30 件においてセレクトブクマがもっとも高い得点となっている．このことから，セレクトブクマにおいて，検索結果上位 30 件中には，Google 検索と比較してもブックマーク数順と比較しても，体系だった知識を得られる Web ページや有益な Web サービスの数が多いのではないかと考えられる．

「映画」においては，Google 検索において，すべての再現率，得点の合計値がもっとも高い値となっている．このことから,「映画」においては，Google 検索において，検索結果上位 30 件中に体系だった知識を得られる Web ページや有益な Web サービスの数がもっとも多いと考えられる．

1.5. 考察

1.5.1. 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体系だった知識を得られる Web ページや有益な Web コンテンツ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，大手企業の Web サイトの満足度が依存していると考えられる.「映画」という分野においては，大手企業の Web サイトが多数あり，内容も充実している．そのため，Google 検索でそういった大手企業の Web サイトが検索結果の上位に表示される傾向にあり，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 ページでも，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．

1.5.2. 時間情報を利用することの有効性

セレクトブクマでは，ソーシャルブックマークが長い間ブックマークされてるかどうかという指標を利用することによって，一時的に面白い情報をフィルタリングし，体系だった知識を得られる Web ページやいつ見ても有益な Web コンテンツを中心に取得することを目的としている．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，一時的に必要とされる情報をフィルタリングできていることがわかった．

1.5.3. 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，大手企業が作成している 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，コンテンツ作成者しか Web ページの評価をすることができない上，コンテンツ作成者が良いと思った Web ページにどんどんリンクを貼って行くような Web サイトやコンテンツ作成者の数はそれほど多くない．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．そのため，個人が作成した玉石混合の Web ページ群の中から，特に有益な Web ページを発見するのに有用であると考えられる．

1.5.4. 課題

現時点のセレクトブクマでは，収集したデータ量不足の問題やはてなブックマークユーザのデータの偏りの問題がある．また，タグ数というものを指標にしているため，はてなブックマークユーザが，タグをつけにくい分野や単語での検索は，十分な結果が得られない場合が多い．同様にマイナーな分野では十分な結果が得られない場合が多く，万能な情報収集支援ツールとはなっていない．

メジャーな分野においても，現在利用しているセレクトブクマのランキングロジックでは，体系だった知識を得られる Web ページや有益な Web サービス以外を完全にフィルタリングできているは言えない．セレクトブクマの検索結果や被験者の選んだ Web ページやを見てもわかるように，上位 10 件の Web ページの中にも体系だった知識を得られる Web ページや有益な Web サービスではないものが存在する．さらに，上位 10 件の中にあるような Web ページよりも，あきらかに被験者にとって人気の高く有益である可能性の高い Web サービスなどが，上位 20 件以降になっている場合もある．

また，セレクトブクマは，Web 検索システムというより，どちらかというと，Web からの情報収集支援システムという位置付けとなっている．そのため，現在のようにセレクトブクマの Web サイトが一つあり，そこで検索するというアーキテクチャが良いとは限らない．情報収集支援システムとして，もっとよいインタフェースやアーキテクチャを模索していく必要がある．

1.6. まとめ

本章では，ソーシャルブックマークの以下の特性を利用し，それに基づいた長期的に利用できる情報を検索するシステムの提案・実装・評価を行った．

- 短い期間しかブックマークされない Web ページ
  →一時的に必要とされる種類の Web ページが多い.

- 長い期間多くのユーザからブックマークされる Web ページ
  →いつ見ても有用な情報を得られる種類の Web ページが多い.

上述の特性に基づき，ソーシャルブックマークにおいて，ブックマークされる期間の長短という指標を利用した，情報収集支援システム「セレクトブクマ」を提案した.「セレクトブクマ」の評価実験として,「セレクトブクマ」,「Google 検索」,「はてなブックマークのデータをブックマーク数順に並べたもの」の 3 つを比較した．その結果,「Google 検索」と比較した場合，特に大手企業の作成した Web サイトでユーザが満足できず，個人が作成したような Web ページをユーザが利用している場合において「セレクトブクマ」が有用であることがわかった．さらに,「はてなブックマークのデータをブックマーク数順に並べたもの」と比較した場合にも，特に検索結果上位 10 件において，より適合率の高い値を得ることができた．ゆえに，時間情報を利用することによって，一時的に利用する Web ページをフィルタリングし，今後も利用したいと思うような体系だった知識を得られる Web ページや有益な Web コンテンツを取得することができることがわかった.

以上の結果から，本研究のソーシャルブックマークの時間情報を利用した情報フィルタリング手法により，個人が作成している大量の Web コンテンツ群の中から，体系だった知識を得られる Web ページや有益な Web サービスを中心に取得する一つの手法を提案できた.