平成21年度

慶應義塾大学

修士論文

時間概念を利用したウェブ情報のフィルタリングと検索

政策・メディア研究科修士２年

上野 大樹

インタラクションデザインプロジェクト

2010年1月

修士論文　2009年度（平成21年度)

時間概念を利用したウェブ情報のフィルタリングと検索

論文要旨

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良コンテンツを手軽に取得することが難しくなって来ている。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良コンテンツを取得するために、ソーシャルブックマークのデータを利用した。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、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た。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、実装、評価を行った。

キーワード

Web 検索,　情報フィルタリング，情報レコメンデーション，時間，ソーシャルブックマーク

慶應義塾大学 大学院政策・メディア研究科
上野 大樹

Abstract of Master's Thesis Academic Year 2009

Web information filtering and search using the concept of time

Summary

Summary

Key Word

Web Search, Information Filtering, Recommendation Information, Time, Social Bookmark

Keio University Graduate School of Media and Governance

Taiki Ueno

目 次

第1章　はじめに

概要

本章では，本研究の目的と論文の構成について示す.

1.1. 研究の背景・動機

近年，Web 上の情報やコンテンツの種類は急激に増加してきており，手軽に Web 上から有益な情報を得られることも多い．そのため，Web は情報取得のツールとしてますます身近なものになっており，Web 上から自分が興味のある分野や詳しく知りたい分野に対して，手軽に体系だった知識や有益なコンテンツを取得したいという欲求がある．また，近年の傾向として，Web 情報の更新速度が非常に早くなって来たため，新しい分野などの情報を得るのには，Web 利用が最適である場合もある．

だが，Web 上の情報量が急激に増加してきたことと，コンテンツの種類自体も多種多様化してきたことにより，Web 上から自分の知りたい分野に対して，体系だった知識を手軽に取得することが難しくなってきている．さらに，存在する有益なコンテンツやサービスをなかなか発見できない場合も多い．

1.2. 研究の目的

本研究では，Web 上から利用者が情報を収集したい分野に対して，体系だった知識を手軽に取得できるようにすることを目的とする．また，有益なコンテンツやサービスも手軽に発見できるようにすることも目的とする．

1.3. 論文の構成

第2章　ウェブ検索の背景

概要

本章では，ウェブ検索の背景と現状の問題点，今後の課題について述べる．

2.1. ウェブ検索の背景

2.2. ウェブ検索の問題点

2.2.1. 一般的な問題点

近年，ウェブ上の情報は急激な勢いで増加しており，また，コンテンツの種類も多種多様化して来ている．そのため，既存の Web 検索エンジンの検索手法では，検索者の意図にそぐわないページがひっかかる場合も多い．また，ウェブ上から手軽に調べたい分野に対して，体系的な情報や知識を得ようとしても，そういった情報を集めるためにどのような検索クエリを入力すれば良いかの判断は容易ではない．さらに、ウェブ上には有用なウェブサービスも多々あるが，そういったウェブサービスを発見することも簡単ではない．このため，有用なウェブサービス発見が，知人の口コミによるものだったり，そもそも有用なウェブサービスの存在自体に気付けない場合も多い．

2.2.2. 検索エンジンの問題点

現在，Google などをはじめとした主だった検索エンジンが利用している代表的なウェブページ検索のランキングアルゴリズムでは，あるサイトから他のサイトへのリンクを評価とみなして，ランキングを行っている．代表的なものとしては，Google が利用している PageRank[1] などがある．近年では，ブログやマイクロブログ，wiki などのようにテンプレートから自動生成されるウェブサイトの数が急激に増加してきている．こうしたシステムやサイトはページ間をその品質を人手で判断するわけではなく，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言いがたく，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ウェブページの発見では，高い検索リテラシーが必要となる場合が多い．

第3章　ソーシャルブックマークについて

概要

本章では，ソーシャルブックマーク [1] の概要と日本最大のソーシャルブックマークサービスであるはてなブックマーク [1] のデータを分析した結果を示す.

3.1. 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，有益なウェブページを共有するウェブサービスである．ソーシャルブックマークでは，folksonomy[1] 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．

3.2. ソーシャルブックマークデータ分析

3.2.1. 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図 3.1 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

データ量は，約 70 万 URL，約 2000 万レコードとなった．

3.2.2. ブックマーク数・ユーザ数・タグ数に関する分析

3.2.3. 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 種類のタイプのウェブ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**??**
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**??**
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**??**

図 3.1: はてなブックマーク

以上の 3 種類のタイプのウェブページをさらに，大まかに分類すると，以下の 2 種類のタイプのウェブ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 のウェブページに対して，そのウェブページがどういった種類のウェブページであるかを分析した．その分析結果を以下の図 3.2 と図 3.3 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，且つ，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 3.2，図 3.3 から分かるように，Type1 のウェブページでは，「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ウェブペー

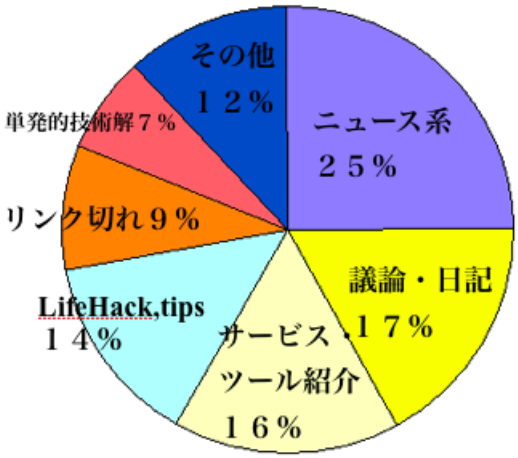

図 3.2: Type1 のウェブページの種類

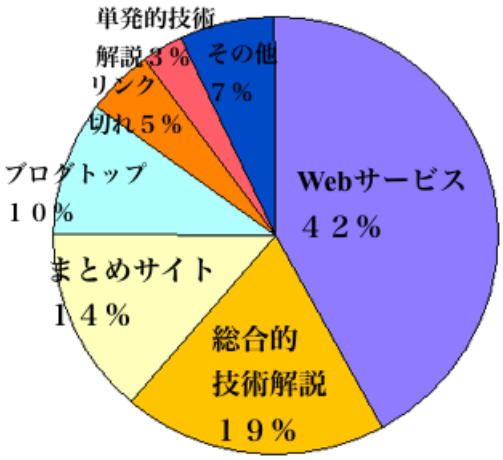

図 3.3: Type2 のウェブページの種類

じが大半を占めている．これに対して，Type2 のウェブページでは,「ウェブサービス」,「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ウェブ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ウェブ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ウェブページをフィルタリングして，いつ見ても有用なウェブページのみを検索できる可能性が高いことが分かった．

第4章　ソーシャルブックマークの時間情報を利用したウェブからの情報収集システムの提案

概要

本章では，第3章のソーシャルブックマークデータの分析結果に基づき提案した，セレクトブクマというウェブからの情報収集システムについて示す．

4.1. セレクトブクマの提案

4.1.1. 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，ウェブ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ウェブサービスを収集できる．

4.1.2. 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
- 有用なウェブサービスを発見すること

情報を収集する際に，上記以外のウェブ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ウェブ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続けるウェブページは，長期間必要とされる種類のウェブページが多いという分析結果に基づいて利用している．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，一時的にしか必要としないウェブページをフィルタリングすることができる．

4.2. セレクトブクマの機能

セレクトブクマは，図 4.1 のような画面構成となっている．図 4.1 は，”java”を検索単語（タグ）として検索した結果である．(1) は検索結果のウェブ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，クリックしたウェブページを表示する．(2) は，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．(3) は検索単語（タグ）を入力するテキストボックスで，検索単語（タグ）を入力し，検索ボタンを押すことにより，指定した検索単語（タグ）で検索を行う．(4) は人気の検索単語（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．

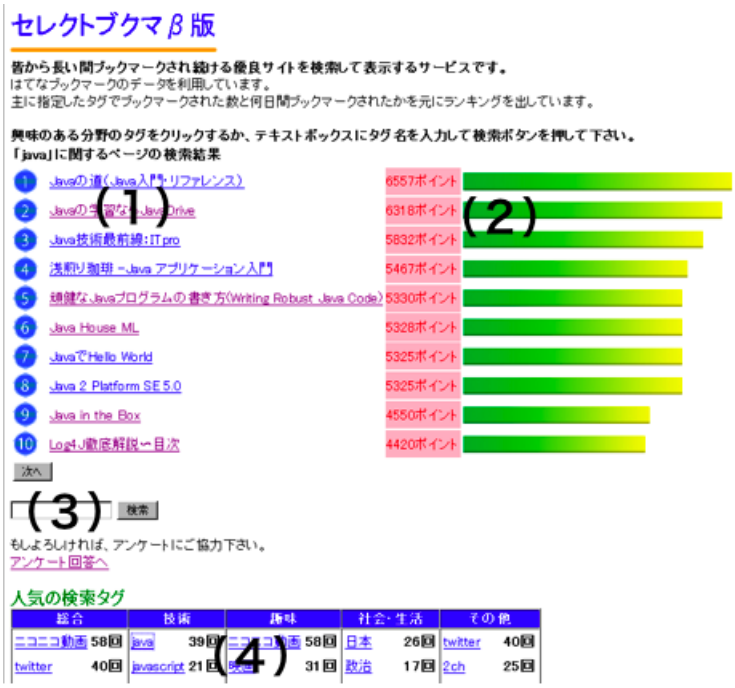

図 4.1: セレクトブクマ画面

4.3. 検索ランキングロジック

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ブックマークされた数に，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，値の大きいものから順にランキングしている．計算式は，以下に示す．

$$Bookmarks \times Days^{\alpha} \tag{4.1}$$

$Bookmarks$：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

第5章　提案システムの実験と評価

概要

本章では，セレクトブクマに対して行った実験とその評価について示す.

5.1. 実験方法

Google 検索結果上位 30 件とセレクトブクマ検索結果上位 30 件を取得し，その中でどのウェブページが良いか被験者に順位をつけてもらう．被験者が選んだウェブページを正解集合として，適合率，再現率を求める．また，同様に時間係数 α の値を変更していって，α の値による適合率，再現率を求め，最適な α の値を求める．

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」,「映画」,「政治」

5.2. 実験結果

5.2.1. 各検索手法による検索結果上位 10 件

Google 検索とはてなブックマークのブックマーク数順にランキングしたものとセレクトブクマでの検索結果を示す．検索単語（タグ）は，”ui”,”java”,”映画”,”政治”の 4 つとする．

5.2.2. Google 検索とブックマーク数順とセレクトブクマの比較

5.2.3. 時間係数 α の最適値

表 5.1: ”ui”で検索した場合の検索結果

順位	Google 検索	ブックマーク数順	セレクトブクマ
1		Life is beautiful:直感的な UI と hande-eye-cordination の話	ユーザーインターフェースデザイン研究室
2		直感的なインタフェースをめざして	Technologies for UI
3		ぼくはまちちゃん！(Hatena) - UI について思うこと	@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
4		ユーザーインターフェースデザイン研究室	ソシオメディア — UI デザインパターン
5		prima materia diary - Google の UI は OK/キャンセルを訊いてこない	アップル ヒューマンインタフェースガイドライン
6		80 年代の Apple に学ぶ UI の部品化とガイドライン？@ IT	Joel on Software - 環境をコントロールできれば楽しく感じるもの
7		naoya のはてなダイアリー - インタフェースの話	Yahoo! Design Pattern Library
8		キャズムを超えろ！ - 家電メーカーよ、今すぐその時代遅れの UI から脱却せよ	ダメなユーザインタフェイス講座
9		Life is beautiful: ユーザー・インターフェイスの設計に大切なのはデザイン・ポリシー	使える GUI デザイン
10		Technologies for UI	80 年代の Apple に学ぶ UI の部品化とガイドライン？@ IT

表 5.2: ”java”で検索した場合の検索結果

順位	Google 検索	ブックマーク数順	セレクトブクマ
1		Java のクラスアンロード (Class Unloading)	Java の道 (Java 入門: リファレンス)
2		Java の道（Java 入門・リファレンス）	Java の学習なら、JavaDrive
3		頑健な Java プログラムの書き方 (Writing Robust Java Code)	Java 技術最前線：ITpro
4		Java 技術最前線：ITpro	浅煎り珈琲-Java アプリケーション入門
5		Java の学習なら JavaDrive	頑健な Java プログラムの書き方
6		浅煎り珈琲-Java アプリケーション入門	Java House ML
7		Java で Hello World	Java で Hello World
8		Java 2 Platform SE	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	Java in the Box
10		Java House ML	Log4J 徹底解説〜目次

表 5.3: ”映画”で検索した場合の検索結果

順位	Google 検索	ブックマーク数順	セレクトブクマ
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

表 5.4: ”政治”で検索した場合の検索結果

順位	Google 検索	ブックマーク数順	セレクトブクマ
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

第6章　関連研究

概要

本章では，本研究の目的と論文の構成について示す.

6.1. ウェブ検索に関する関連研究

6.2. 情報フィルタリングに関する関連研究

6.3. ソーシャルブックマークを利用した関連研究

ソーシャルブックマークデータを利用した関連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. ウェブページの検索
2. ウェブページの推薦
3. ソーシャルブックマークユーザの推薦
4. ソーシャルブックマーク，Folksonomy の分析

この中で本研究は，ウェブページの検索に分類される．ソーシャルブックマークデータを用いたウェブページの検索に関する関連研究として，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 を統合して，ウェブ検索ランキング精度の向上を計っている [1]．また，Takahashi 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，鮮度の高いウェブページを取得する検索手法を提案している [1]．ここでは，ブックマーク日時の散らばりの大きさから，ウェブ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ウェブページを取得している．

ウェブージの推薦に関する研究として，Sasaki らは，タグを表象とするウェブコンテンツ群の類似性に基づいたウェブコンテンツ推薦システムを提案している [1]．また，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ウェブページ全体と対象としたウェブページ推薦システムの構築手法を提案している [1].

第7章　考察と展望

概要

本章では，本研究の目的と論文の構成について示す.

7.1. 考察

7.2. 展望

第8章　おわりに

概要

本章では，本研究の目的と論文の構成について示す.

8.1. 研究の成果

8.2. 総括と結論

本研究に関する発表

1. 上野大樹, 安村通晃 セレクトブクマ：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. 第 50 回プログラミング・シンポジウム pp 9-16, January 2009.

謝辞

本研究を進めるにあたり，主査として研究を基礎から指導して頂き，多くのことを学ばせて頂いた慶應義塾大学 安村通晃教授に深く感謝いたします.

また，副査として，かつ，インタラクションデザインプロジェクトに参加して頂き，本研究に関して多くの的確なコメントとアドバイスをして頂きました慶應義塾大学 増井俊之教授に深く感謝いたします.

また，副査として，本研究に関して，特に評価実験に関してご指導いただきました慶應義塾大学 小川克彦教授に感謝いたします.

さらに，所属する研究会において，多くのコメントとアドバイスや研究に関する相談に乗って頂きました慶應義塾大学 樋口文人先生に感謝いたします.

また，安村研究室およびインタラクションデザインプロジェクトに所属するみなさんには，日頃から本研究に関するアドバイスをして頂いたり，研究に関する議論をして頂き，大変感謝いたします.

２０１０年1月

慶應義塾大学 大学院政策・メディア研究科修士２年

上野 大樹

参考文献

[1] Toshiyuki Masui, Koji Tsukada, and Itiro Siio. Mousefield: A simple and versatile input device for ubiquitous computing. In *UbiComp2004, Springer LNCS3205*, pp. 319–328, 2004.

[2] 石山琢子, 塚田浩二, 安村通晃. 想起将棋の提案と試作. ヒューマンインタフェースシンポジウム 2005 論文集, pp. 483–486, 2005.

付録