ソーシャルブックマークデータの時間情報を用いた情報フィルタリングと検索

慶應義塾大学 大学院
政策・メディア研究科
修士課程1年
上野 大樹

私、発表者について

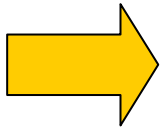
- 修士入学当初からSBM関連の研究に関わる
- 当初、SBMデータを利用したWebページレコメンド及び視覚化の研究を実施
- 現在は、SBMデータを利用したWebページ検索関係の研究を実施

- Twitter：wilfue
- はてなID：kasihara1

目次

SBM研究の種類

- SBM、FolkSonomyの分析
- Webページ推薦
- ブックマーカーの発見
- Webページ検索



本研究の位置づけ

SBM、FolkSonomyの分析

Golder,2005 ”The Structure of Collaborate Tagging System”

- ユーザ、タグ、ブックマークの性質について分析
- 各Webページに対する各タグの出現頻度は一定値に収束

川中,2008 “ソーシャルブックマークにおけるタグの時系列的な依存関係の解析”

- 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
- タグの時系列の関係性をグラフ化

Webページ推薦

Niwa,2006 “Web page recommender system based on folksonomy mining“

- クラスタリングによりタグの表記のゆれの問題を解決
- ユーザのブクマ情報からユーザの趣向に沿ったページを推薦

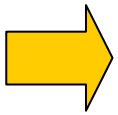
Sasaki,2006 “Web Content Recommendation System based on Similarities among Contents Cluster of Social Bookmark”

- タグではなくコンテンツクラスタを利用したWebページの推薦

※第１回SBM研究会で発表

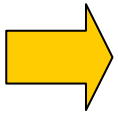
ブックマーカーの発見

Shiratsuchi,2006 “Finding unknown interests utilizing the wisdom of crowds in a social bookmark service”



- 指定したユーザと類似した嗜好を持つブックマーカーの推薦
- 類似した嗜好を持つブックマーカーの関係性のグラフ化

百田,2008 “ソーシャルブックマークに基づく情報発見”



- αブックマーカーの発見
- αブックマーカーを利用したWebページの推薦

Webページ検索

山家,2007 “ソーシャルブックマークの特性を利用したWeb検索ランキング精度の向上”

- PageRankとSBRank(ブクマ数の多さ)を融合し、PageRankの検索精度の向上を計っている

Xu,2008 “Exploring Folksonomy for Personalized Search”

- 質問の期間ベクトルと文書の期間ベクトルのコサイン類似度を利用

目次

背景

- インターネット上の情報爆発が問題となっている
- GoogleなどがPageRank

⇒ blogやwikiなど自動的にページ間を結びつける場合、人々の意思が反映されるとは限らない

- 良サービス、良コンテンツ発見のためには
  - 目的外のページが引っかかり、Webサイト発見が面倒
  - 高い検索リテラシーが必要
  - そもそも良コンテンツ・サービスの存在に気付けない

⇒ 知人からの紹介で気付く場合も・・・

  - RSSリーダー、SBMトップは定期的なチェックが手間

そこで

- 集合知であるSBMデータを用いた情報検索
  - 検索リテラシーを不要としたWebページの発見
  - 不要な情報をフィルタリングして取得

WHY？

- 皆が選んだ情報は、大多数の人にとって有用
- 人手による情報フィルタリングの有用性
- 他人が発見した有益な情報の発見

ただし

- SBMのブクマ数には特徴がある
  - トップに表示される人気のエントリーに入るとブクマ数が急激に上昇
  - 単にブクマ数の多いページほど有用であるか？
  - 一時的にブクマ数が増えたページが後で見て有用かどうか？
- 各SBMによって偏りもある
- 情報検索のためのデータ数は十分か？

目次

ブクマ数と時間の関係をグラフ化したところ・・・

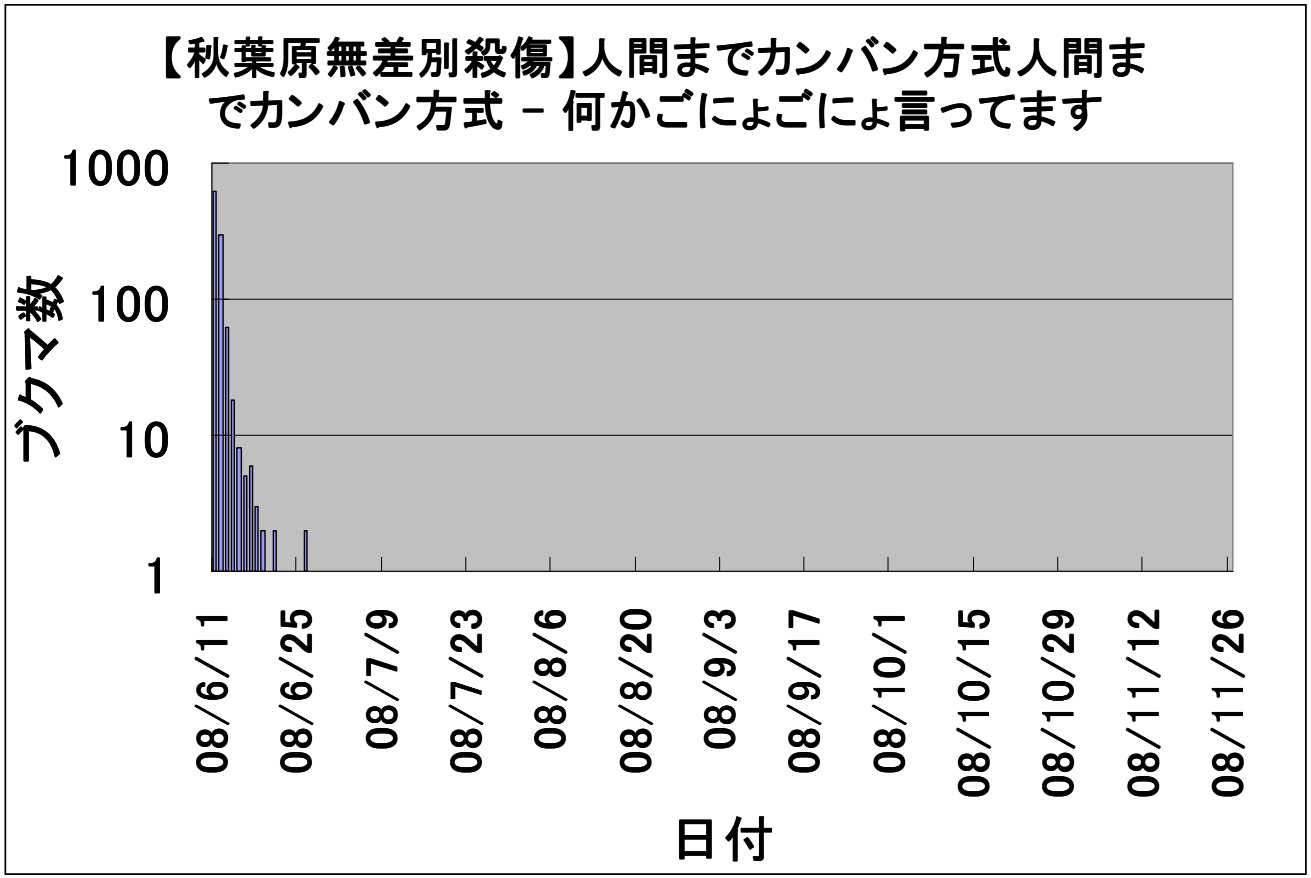
- 以下３つのタイプのページが存在

１．一時的に急激にブクマされ、
その後はあまりブクマされないタイプのページ

２．急激にブクマ数が増える時期はあるが、
長い間ブクマされ続けるページ

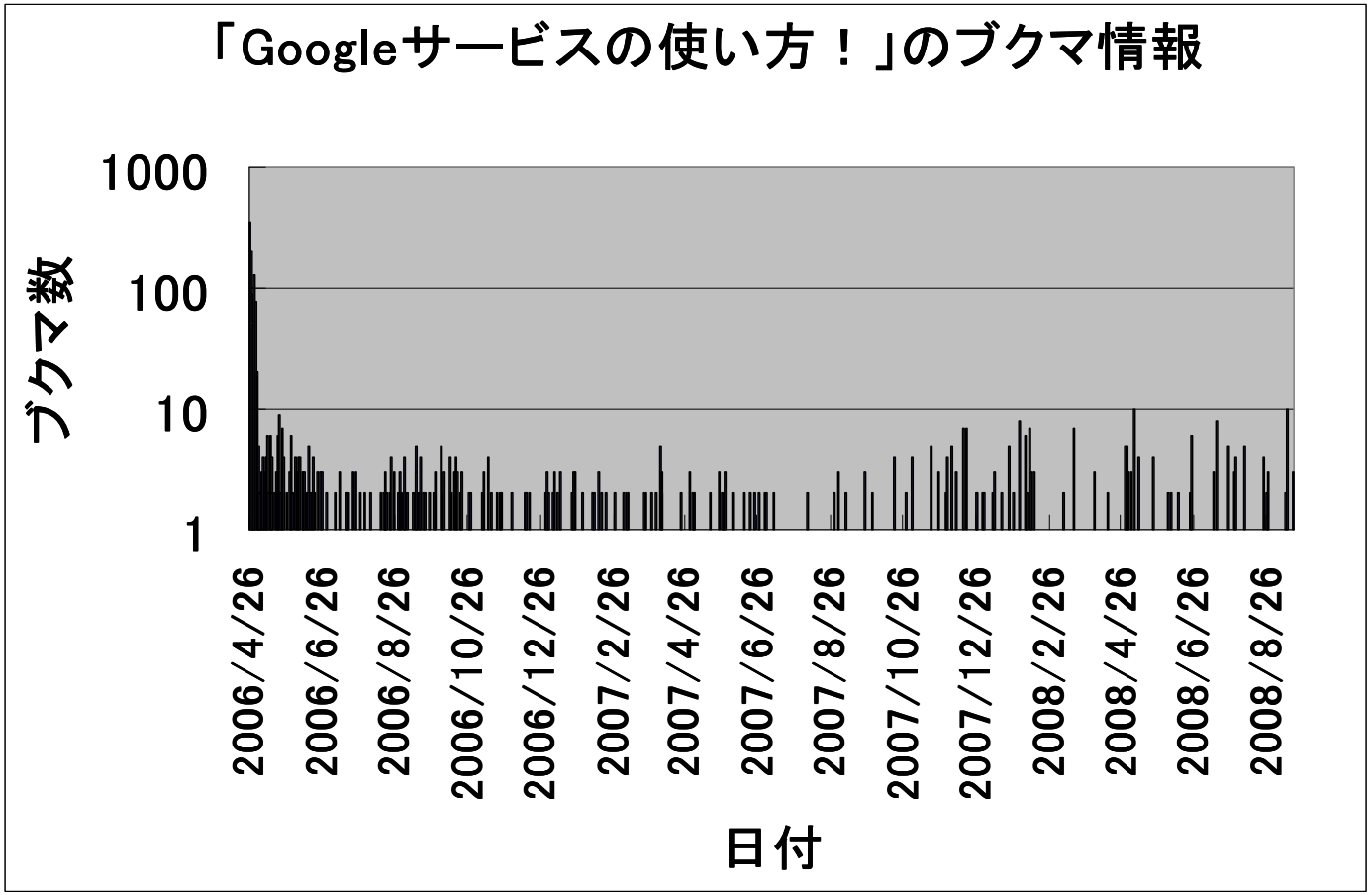
３．急激にブクマ数が増える時期がなく、
長い間ブクマされ続けるページ

ブクマ数と時間の関係1



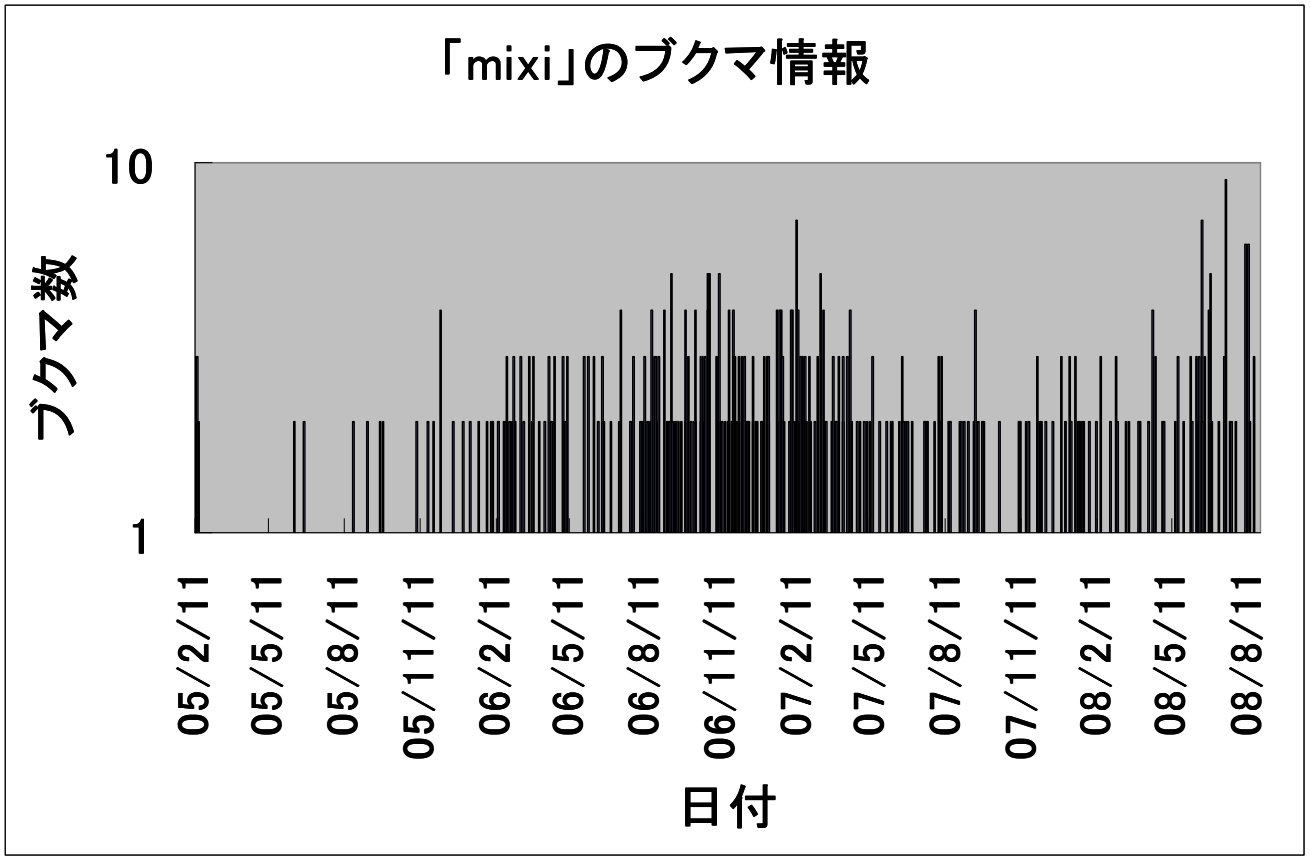
一時的に急激にブクマされ、
その後はあまりブクマされないタイプのページ

ブクマ数と時間の関係2



急激にブクマ数が増える時期はあるが、
長い間ブクマされ続けるページ

ブクマ数と時間の関係3



急激にブクマ数が増える時期がなく、
長い間ブクマされ続けるページ

時間情報に着目した仮説

以上の１～３のタイプのページを2つに集約

- 一時的に急激にブクマされ、その後にはブクマされないタイプのページ（Type Ⅰ）

⇒ 一時的に必要とされる情報

- 長期間に渡ってブクマされるタイプのページ（TypeⅡ）

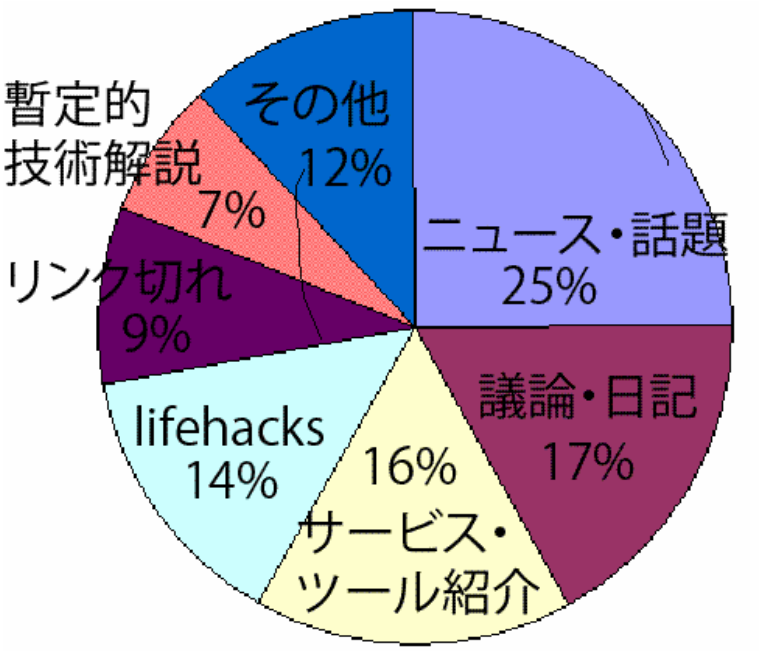
⇒ 長期的に必要とされる情報

時間情報に基づきWebページの種類を分析

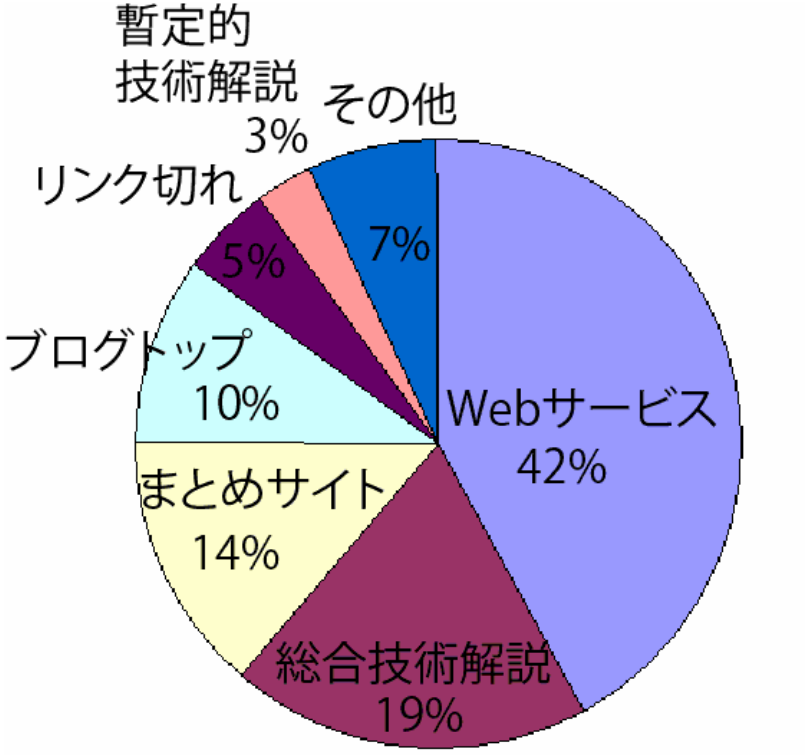
ブクマ数100以上のページをランダムに100ページずつ取得

Type Ⅰ：ブクマされた日数/ブクマ数=0.2以下
Type Ⅱ：ブクマされた日数/ブクマ数=0.8以上

Type Ⅰのページ分析結果



Type Ⅱのページ分析結果



分析結果より

TypeⅠとTypeⅡで全く異なった種類のWebページ

- 一時的に急激にブクマされ、その後にはブクマされないタイプのページ（TypeⅠ）

⇒ 一度読んだら終わりのページが多い

- 長期間に渡ってブクマされるタイプのページ（TypeⅡ）

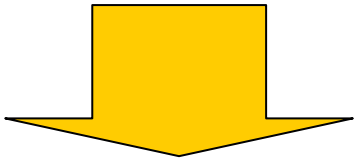
⇒ 何度も利用するタイプのページが多い

TypeⅠとTypeⅡを分類することにより、不要な種類のページをフィルタリングできる

目次

セレクトブクマの開発

- 長期間に渡ってブクマされるタイプのページ（TypeⅡ）を優先的に取得
- タグごとに検索



抽出

- 利用シーンが特定的、一時的、暫定的のページをフィルタリング
- 何度も利用するタイプのページ

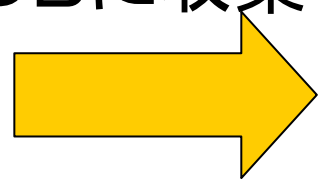
セレクトブクマ（利用データ）

- はてなブックマークのデータを利用
- 2005年2月～2008年8月のデータ
- ブクマ数5以上の全てのページ

はてなの伊藤直也様には
大変お世話になりました！

セレクトブクマ（収集データ概要）

- 以下の6種類のデータを収集
  1. URL
  2. タイトル
  3. ブクマ数
  4. 時間情報（ブクマした日付）
  5. ユーザID（ブクマしたユーザID）
  6. タグ（ブクマしたユーザに紐付くタグ名）



約70万URL
約2000万レコード

セレクトブクマ（ランキングロジック）

指定したタグにおいての
ポイント＝ブクマ数×日数$^{\alpha}$（α=0,1,2で実験)

※ポイントが高いものから順にランキング
※日数とは
例：2006/8/3、2006/8/6,2007/3/5にブクマされたら3日

目次

Google検索との比較（「java」で検索）

順位	Google検索	セレクトブクマ（α=1）
1	Java.com:あなたとJava	Javaの道(Java入門：リファレンス）
2	無料 Java ソフトウェアをダウンロード - Sun Microsystems	Javaの学習なら、JavaDrive
3	テクノロジ - サン・マイクロシステムズ	Java技術最前線：ITpro
4	Java - WikipediaJava	浅煎り珈琲-Java アプリケーション入門
5	ノーブラシ洗車場のJAVA	頑健なJavaプログラムの書き方 (Writing Robust Java Code)
6	Javaの道(Java入門：リファレンス）	Java House ML
7	Java Solution - @IT	JavaでHello World
8	Javaとは-はてなキーワード	Java 2 Platform SE 5.0
9	Javaとは-意味・解説：IT用語辞典	Java in the Box
10	JAVA動物実験の廃止を求める会	Log4J徹底解説～目次

Google検索との比較（「ui」で検索）

順位	Google検索	セレクトブクマ（α=1）
1	ユーザインタフェース - Wikipedia	ユーザーインターフェースデザイン研究室
2	ユーザインターフェースとは【user interface】- 意味・解説：IT用語辞典	Technologies for UI
3	UIとは【ユーザインターフェース】- 意味・解説：IT用語辞典	@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	UI ゼンセン同盟	ソシオメディア	UIデザインパターン
5	UI - Wikipedia	アップル ヒューマンインタフェースガイドライン
6	「使える、使いやすい、使いたい」と思えるUI とは？ - @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	ソシオメディアー UI デザインパターン	Yahoo! Design Pattern Library
8	UI パターンいろいろ- Design-Walker	ダメなユーザインタフェイス講座
9	MOONGIFT: ≫ Prototype.js を用いた UI ライブラリ「PrototypeUI」:オープンソースを毎日紹介	使える GUI デザイン
10	Google Japan Blog: Gmail モバイルの UI を刷新しました	80年代のAppleに学ぶUIの部品化とガイドライン？ @IT

時間情報利用によるフィルタリング効果（「java」タグ）

順位	ブクマ数の多い順（α=0）	セレクトブクマ（α=1）
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(Writing Robust Java Code)
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 5.0	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

時間情報利用によるフィルタリング効果（「ui」タグ）

順位	ブクマ数の多い順（α=0）	セレクトブクマ（α=1）
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？@ITW	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

α=1と2の場合の比較(「java」タグ)

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

α＝1と2の場合の比較（「ui」タグ）

順位	セレクトブクマ（α＝1）	セレクトブクマ（α＝2）
1	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　最新のUIデザインパターン
2	Technologies for UI	UI-patterns.com
3	@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	ユーザーインターフェースデザイン研究室
4	ソシオメディア	UIデザインパターン	@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
5	アップル ヒューマンインタフェースガイドライン	Technologies for UI
6	Joel on Software - 環境をコントロールできれば楽しく感じるもの	ソシオメディア	UIデザインパターン
7	Yahoo! Design Pattern Library	Joel on Software - 環境をコントロールできれば楽しく感じるもの
8	ダメなユーザインタフェイス講座	Yahoo! Design Pattern Library
9	使える GUI デザイン	ダメなユーザインタフェイス講座
10	80年代のAppleに学ぶUIの部品化とガイドライン？@IT	アップル ヒューマンインタフェースガイドライン

セレクトブクマβ版公開中

- セレクトブクマで検索
- URLは以下

http://plazman.chi.mag.keio.ac.jp/sbm/summary.jsp

- 明日（12/7(日)）は、サーバ停止のため利用不可

セレクトブクマ（アンケート結果）

- アンケート調査を実施
- 合計14人が回答
- 有用なページ発見：10人／14人
- 面白いページ発見： 10人／14人

セレクトブクマ（考察）

- ブクマ数の多いタグによる1単語検索で有用
- 利用シーンが特定的、一時的、暫定的なWebページをフィルタリング
- Googleなどの検索では見つけにくいが、有用なWebページの発見も可能
- 検索リテラシーを不要とした検索
- Googleなどと比較して、その単語に強い興味がある場合に有用

セレクトブクマ（今後の課題）

- ロジックの改良
  - 時間の閾値は一日で良いのか？
  - 最適な日付係数αの発見
  - 現時点だと古いページが有利

⇒ ブクマされた時期などによる重み付け

  - タグをつける行為の意味（フィールドワーク）

⇒ タグ率の利用

- データ数、データの偏りの問題

⇒ はてブ以外のSBMデータも調査、利用

- 被験者を利用した評価

セレクトブクマ（今後の課題：評価）

- 以下の4つを比較

1．Google検索

2．タグによるブクマ数順のランキング

3．セレクトブクマ（現時点でのもの）

4．セレクトブクマ（改良版）

・被験者に上記4つでの上位10件のページを比較してもらう

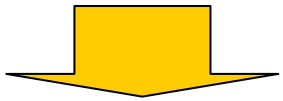
・20～30人ぐらい

・タグは5～10種類程度

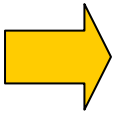
目次

ブクマストリーム概要

- Webページを年間、月間、日間ごとにブクマ数順にランキング化して表示
- タグごとにもランキング可能



- ブクマ数の多いページをまとめてチェック
- 過去のブクマ数の多いページを発見

 近日公開予定

関連サービス

- MixClips

➡複数のSBMの人気ブクマを横断的に表示するサービス

※「はてなブックマーク」「del.icio.us」「@nifty クリップ」「livedoor clip」「buzzurl」「newsing」「あとで新聞」等

- はてなブックマークじわじわ来てるエントリー

➡集中的にブクマされていないため、「人気エントリー」に上がってきていないWeb ページを検出

※総ブクマ数40 以上、同日連続ブクマ数15以下のWeb ページを検出