社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

ソーシャルブックマークにおける文書解析を利用した類似文書および類似ユーザの推薦方法の提案

矢島 健太郎† 井上 潮‡

† 東京電機大学 工学部情報通信工学科 〒105-8457 東京都千代田区神田錦町 2-2
E-mail: † yjm@de.c.dendai.ac.jp ‡ inoue@c.dendai.ac.jp

あらまし ユーザ間でブックマーク資源を共有し，有効活用する試みがソーシャルブックマーク（SBM）である．SBM においてブックマーク資源をより有効に活用するためには，ユーザ間での優れた共有手段が不可欠である．すでに多くの SBM が存在し，さまざまなアプローチで共有を図っているが，ブックマークされた文書自体の情報が十分活用されていない．本稿では，ブックマークされた文書の内容について構文解析を行い，その特徴を抽出することにより，類似ブックマークおよび類似ユーザの推薦を行うための手法を提案し，プロトタイプの実装と有効性の検証結果について述べる．

キーワード ソーシャルブックマーク，情報推薦，文書解析

Recommendation of Similar Documents and Users by Document Analysis on Social Bookmark

Kentaro YAJIMA† and Ushio INOUE‡

† Tokyo Denki University 2-2 Kanda-Nishiki-cho, Chiyoda-ku, Tokyo 105-8457 Japan
E-mail: † yjm@de.c.dendai.ac.jp, ‡ inoue@c.dendai.ac.jp

Abstract Social Bookmark (SBM) shares bookmarks among users and uses them for common resource of the community. In order to use the resource effectively, a good mechanism for the sharing is required. Many approaches have been proposed for the sharing, but the contents of the bookmarked documents are not fully utilized. This paper proposes a new approach that analyzes the contents of bookmarks, extracts the features of the bookmarks and recommends similar documents and users, and describes a prototype system and feasibility studies.

Keyword social bookmark, information recommendation, document analysis

1. はじめに

近年，Web の普及に伴い個人の持つブックマークが増加している．ブックマークは Web 上の多くの文書から個人が選択したものであり，その嗜好をはじめ多くの意味を含む資源である．しかし，標準的なブラウザのブックマーク機能においては，ブックマークは個人のものに過ぎず，ユーザ間での相互利用がなされていない．

これらを共有し有効活用しようという試みがソーシャルブックマーク（SBM）である．SBM はブックマーク資源をユーザ間で相互利用する．共有による有効活用とはすなわち，あるユーザにとって有用な文書を他ユーザのブックマークの中から抽出することである．文書の評価はユーザの主観であるので，有用であるという判断をプログラムで完全に行うことは非常に困難である．よって有用である可能性が高いコンテンツを推薦することが重要である．

すでに多くの SBM が存在し，さまざまな手段でブックマークの共有を図っている[1,3,4,7,8,9]．しかし，既存の手段ではブックマークされた文書の内容（本文）に目が向けられていない．本文にはユーザの嗜好が反映されているため，この情報を用いることによって，より精度の高い推薦が行えると考えられる．

ブックマークされた文書の本文を文書解析することによってその特徴を抽出，比較することで，内容が類似した文書（類似文書）を検出することができる[5,6]．ユーザにとって有用である文書と類似した文書もまたユーザにとって有用である可能性が高い[10]．

また，ユーザが所有するブックマーク集合の各々の文書について構文解析を行うことで，集合の特徴を抽出することができる．これはユーザの嗜好を反映しており，これを比較することで嗜好の近いユーザ（類似

ユーザ）を検出することができる．類似ユーザが所有する文書はユーザの嗜好に合致している可能性が高い．

本研究では，類似文書の検出と類似ユーザの検出手法について提案する．以下 2 章で既存 SBM における共有の手段について述べる．3 章で具体的な手法を提案し，4 章で実装したプロトタイプについて説明する．5 章で提案手法の有効性を検証し，6 章でまとめ，7 章で今後の課題について述べる．

2. 既存 SBM における共有手法

既存 SBM では，ブックマーク共有手法として以下の機能が実現されている[1,3,4].

(1) 検索

ユーザ自身のブックマークに加えて他ユーザのブックマークに対しての検索を行う．興味のある語によって検索して絞り込むことで有用な文書を発見する．

(2) 分類による絞込み

タグと呼ばれるラベルによるブックマークの分類が行われている．タグもユーザ間で共有することにより，タグによって特定分野の文書を抽出できる．嗜好に沿ったタグによって絞り込むことで，有用な文書を発見する．

(3) ランキング

多くのユーザにブックマークされている文書を一定期間ごとにランキングで表示する．多くのユーザにブックマークされていることはその文書の価値が高いことを意味し，個々のユーザにとっても有用である可能性が高い．

(4) 類似ユーザ

同じ文書をブックマークしているユーザは，互いに嗜好が近い可能性がある．そのような他ユーザのブックマークにはユーザの嗜好に沿った有用な文書が存在する可能性が高い．

3. 文書解析を利用した類似検出

既存の共有手段はブックマークされた文書の URL 情報のみを扱い，内容（本文）については利用されていない．本研究では本文に対し構文解析を行うことによりその特徴を抽出し，類似文書および類似ユーザを検出する．

3.1. 類似文書

類似文書の検出に用いる文書の特徴の算出には TF・IDF 法を用いる．文書における名詞の出現回数を TF(Term Frequency)とし，すべてのユーザのブックマーク集合における名詞の出現頻度を DF(Document Frequency)とする．IDF(Inverse Document Frequency)は DF の逆数の対数を取り 1 を足したものとする．TF･IDF は TF と IDF の積とする．これらの定義を表 1 にまとめる。

表 1　TF・IDF の定義

$TF(t) =$ 個々の文書における t の出現回数
$DF(t) = \dfrac{\text{t が出現する文書数}}{\text{ブックマークされたすべての文書数}}$
$IDF(t) = 1 + \log\left(\dfrac{1}{DF(t)}\right)$
$TF・IDF(t) = TF(t) \times IDF(t)$

文書に存在するすべての語について TF･IDF を求めると，TF･IDF 群は文書固有の特徴ベクトルになる．2 つの文書の特徴ベクトルがなす角を求めることで類似ユーザを検出することができ，この角が小さいほど文書の類似度が高いといえる．

3.2. 類似ユーザ

SBM においてはあるユーザの所有するブックマークがすべてのブックマークの部分集合になっている（図 1）.

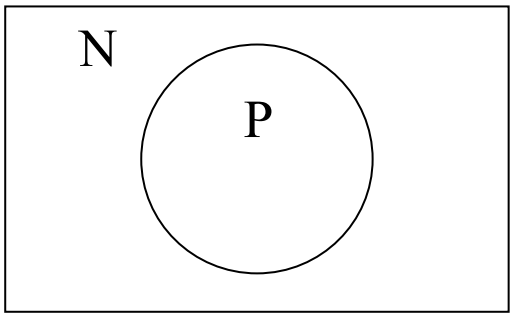

図 1 ブックマークの構造

全体のブックマーク集合を N, ユーザの所有する集合を P とする．ここで P と N における，ある語 t の DF の比を考える．この値は P における t の出現率の偏りであり，これを語の偏り：TB（Term Bias）と定義する．

$$TB(t) = \frac{DF(t,P)}{DF(t,N)}$$

例えば，ブックマークの状態が図 2 のようになっている場合，DF(t, P)=5/10=0.5，DF(t, N)=10/30=0.333 となり，TB(t)=0.5/0.333=1.5 となる。

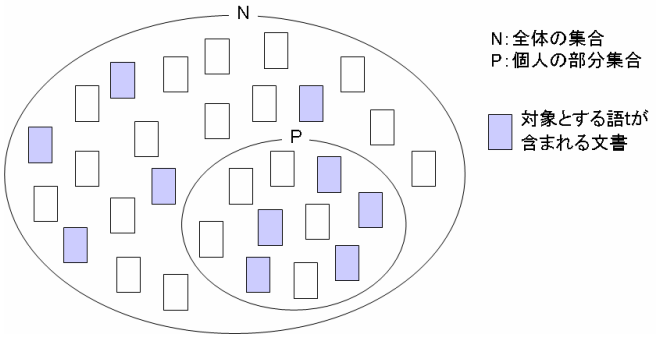

図 2　TB の定義

TB が高い語はユーザの所有する文書に偏って存在する語であり，ユーザの嗜好を反映し特徴づける語である.P におけるすべての語について TB を求めると，TB 群はユーザごとの特徴ベクトルになる.文書比較の場合と同様に，ユーザの特徴ベクトルのなす角が小さいほどユーザの類似度が高いといえる.

3.3. 類似度スコア

文書およびユーザの特徴ベクトルの類似度をあらわすスコアについて，A の特徴ベクトルを F(A)，B の特徴ベクトルを F(B)としたとき，類似度スコア S を以下のように定義する.

$$S = \frac{F(A) \cdot F(B)}{|F(A)||F(B)|}$$

ここで，文書の類似度スコア S を求める場合は特徴ベクトル F として TF・IDF を用い，ユーザの類似度スコア S を求める場合は TB を用いることとする.

類似度スコア S は 2 つのベクトルのなす角 θ に対する $\cos\theta$ の値であり，$0 \leqq S \leqq 1$ である．S の値が大きいほど類似度が高い.

4. プロトタイプの実装

これらの手法を適用したプロトタイプを実装した.プロトタイプの動作環境を表 2 に示す.

表 2 プロトタイプの動作環境

OS	FedoraCore5 Linux
Web サーバ	Apache2.2.2
DBMS	PostgreSQL8.1.4
使用言語	PHP5.1.6
文書解析エンジン	MeCab0.9.3

文書から名詞を抽出する手段として形態素解析を用い，形態素解析エンジン MeCab[2]を利用した.

プロトタイプの構成を図 3 に示す.

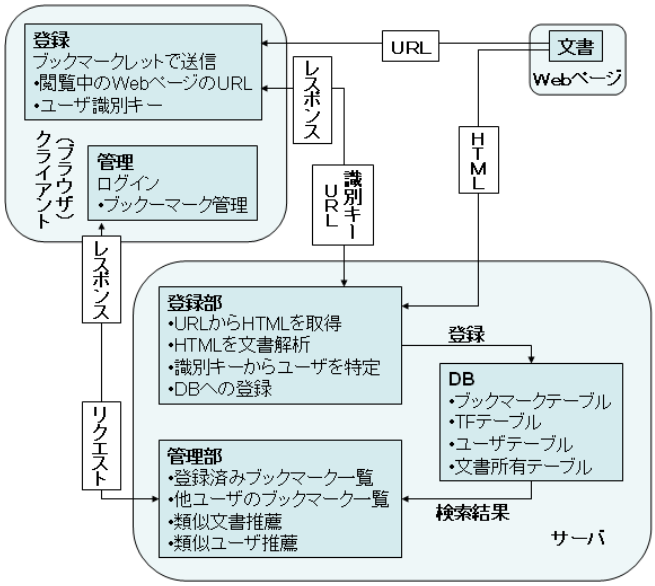

図 3　プロトタイプの構成

サーバ側のシステムはブックマークをデータベースに登録する登録部と，登録されたブックマークから類似エントリおよび類似ユーザの推薦を行う管理部に分けられる.

4.1. 登録部

ブックマークの登録は任意の Web ページに対してブックマークレットを実行することで行う．ブックマークレットとはブラウザのブックマークに記述する短い JavaScript コードで，ブラウザの機能を拡張して，Web ページの表示方法の変更やデータの抽出，別サイトへの転送などを行うことができる．これを利用してサーバ側のブックマーク登録用スクリプトへ URL とユーザ識別用のキーを送信する.

ブックマーク登録用スクリプトでは，受信した URL から PEAR::HTTP/Request により本文を取得して MeCab に渡し，名詞を抽出，名詞ごとにカウントし TF を求める．URL，タイトル，TF を DB に格納する.

4.2. 管理部

ユーザ認証を済ませると，いままでに登録したブックマークの一覧が表示される（図 4）.

図 4　登録したブックマークの一覧表示の一部

各ブックマークについて，「類似ブックマークを表示する」というラベルのリンクがある．これをクリックすることで，類似したブックマークを一覧表示する（図 5）.

図 5　類似ブックマークの一覧表示の一部

また，類似ユーザ発見スクリプトにアクセスすると，類似ユーザの一覧が表示される（図 6）.

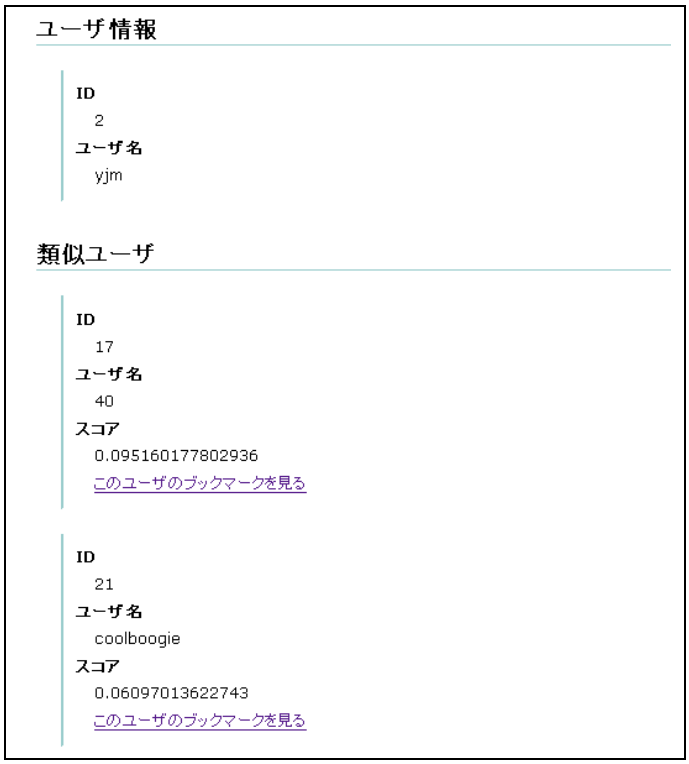

図 6　類似ユーザの一覧表示の一部

リンクより類似ユーザのブックマーク一覧を見ることができる．各ユーザはこれを見ることにより有用な文書を探すことができる（図 7）.

図 7　類似ユーザのブックマークの一覧表示の一部

5. 評価

3 章で述べた類似文書および類似ユーザの検出手法について，4 章で述べたプロトタイプを用いて評価を行い，有効性を検証した．

なお，評価時点での登録ブックマーク数は 153 件，登録ユーザ数は 9 名である．

5.1. 類似文書

提案した類似文書の検出手法が有効であるということは，類似度スコアとその文書が有効である割合の間に正の相関があるということである．つまり，類似度スコアが高いものほど有効である可能性が高いということが示せればよい．

これを示すため，以下のような手順で評価を行った．

(1) ブックマークされた文書から任意にＡ件の文書を選ぶ．
(2) それぞれについて，他の文書との類似度スコアを算出し，上位Ｂ件を求める．
(3) Ａ×Ｂ件の文書対について，実際に類似しているかを人間が判断する．類似している文書を有効文書とする．
(4) 類似度スコアを一定間隔で区切り，それぞれの区間での全文書数に占める有効文書数を有効率とする．スコア区間と有効率の関係を評価する．

今回の評価では，Ａ・Ｂをともに 10 件とした延べ 100 件について評価した．また，類似度スコアの区切りは 0.1 刻みとした．

評価結果を表 3 に示す．

表 3 類似文書の類似度スコアと有効率

スコア区間	有効文書数	全文書数	有効率(%)
0～0.1	1	14	7.14
～0.2	1	31	3.23
～0.3	7	21	33.3
～0.4	4	15	26.7
～0.5	1	6	16.7
～0.6	2	2	100
～0.7	3	5	60
～0.8	1	1	100
～0.9	0	0	－
～1.0	0	0	－

100 件について判定を行ったが，重複が存在したため 5 件を除外した．また，0.8 を超えるスコアは存在しなかった．

類似度スコア区間とその区間における文書の有効率をプロットしたものを図 8 に示す．

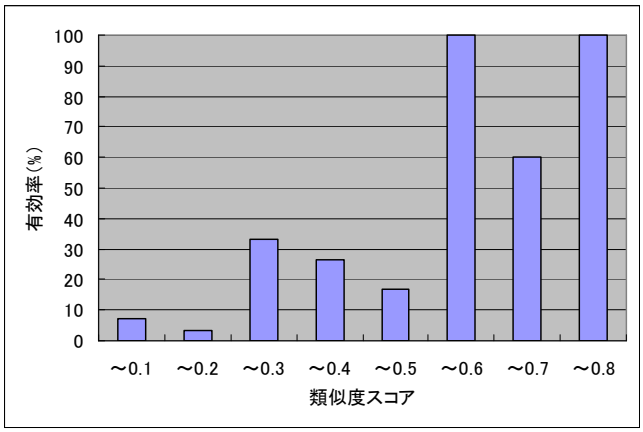

図 8 類似文書の類似度スコアと有効率

この図からスコア 0.5 以上では 75%の文書が有効であり，一方 0.5 未満では有効な文書は 16%ほどしかない．従って，スコアと有効率との間に正の相関あるといえ，類似度スコアによる類似文書発見の手法は有効であるといえる．

5.2. 類似ユーザ

推薦された類似ユーザが妥当であるということは，類似度スコアの高いユーザのブックマークにおいて，自身にとって有用な文書の割合が高いということである．

これを検証するため，著者自身の類似ユーザの類似度スコアと有効ブックマーク割合との相関について調べた．登録ユーザ数が 9 名であるため，筆者自身を除く 8 名との関係を検証した．

評価結果を表 4 に示す．表 4 の各行は一人のユーザに対応し，有効文書数はそのユーザが所有するブックマークのうち著者のブックマークと類似していると判断された文書数，全文書数はそのユーザが所有する全ブックマーク数であり，有効率は有効文書数/全文書数である．

表 4 類似ユーザの類似度スコアと有効率

スコア	有効文書数	全文書数	有効率(%)
0.0048	0	1	0
0.0137	1	3	33.3
0.0277	2	14	14.3
0.0350	2	7	28.6
0.0388	3	10	30
0.0408	3	24	12.5
0.0610	6	40	15
0.0952	10	16	62.5

表 4 におけるスコアと有効率の関係をプロットしたものを図 9 に示す．

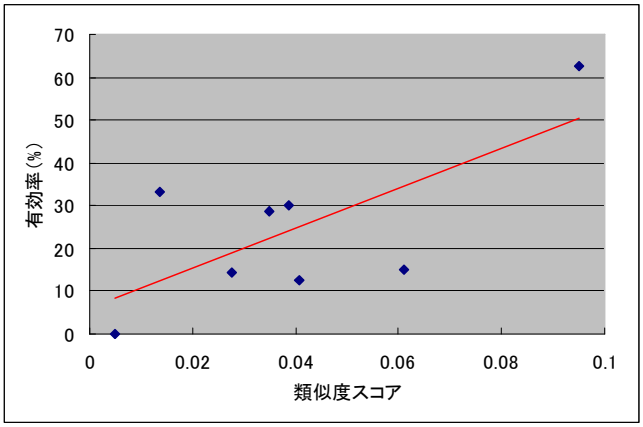

図 9 類似ユーザの類似度スコアと有効率

赤直線は最小二乗法による近似線である．サンプルとなるユーザの件数が少ないため正確性は十分ではないが，ユーザの類似度スコアと文書の有効率には正の相関がある．つまりスコアが高いユーザほど有効な文書の保有割合が高く，類似ユーザの推薦手法として有効であるといえる．

6. まとめ

本稿では文書解析情報を利用して特徴ベクトルを算出し，類似文書および類似ユーザを推薦する手法を提案した．精度の向上などが課題であるが，検証によってその有効性が示された．

提案した手法のうち，類似ユーザの推薦方法で定義した TB はある全体集合の部分集合における語の偏りであり，部分集合においての語の重要度を表す．ブックマークと同様に部分集合がユーザの選択によるものである場合，ユーザにとっての語の重要度になる．この手法を一般化することにより，SBM 以外への応用，たとえば文書検索の際の語の重み付けなどにも利用できると考えられる．

7. 今後の課題

(1) 精度の向上

文書比較において，同じサイトの文書であれば内容的な類似度が高くないものでも類似度スコアが高くなる現象が見受けられた．これはコンテンツ以外のサイト共通の記述部分によって類似度スコアが高くなったためと考えられる．抽出や比較のアルゴリズムの改良によってそれらの要因を除外し，精度の向上を図る必要がある．

(2) 計算量の減少

実装したプロトタイプでは，TF・IDF や TB，類似度スコアの算出を行う際，単にすべての語について計算を行っているため計算量が多く，動作速度に難がある．登録データの増加に伴いさらに遅くなると考えられるので，実用のために結果の妥当性を確保しつつ計算量を減らす必要がある．

(3) 文書解析情報を利用した管理機能

類似文書および類似ユーザの推薦に用いた，文書解析で得られた文書の特徴ベクトルを，基本的な管理機能である検索や分類等にも反映させることで，より有効な管理機能を提供することができると考えられる．

(4) 文書の変更・削除の問題

Web ページは変更されうるため，同じ URL のバージョンの違いが存在しうる．それによって DB に格納されている解析情報との間に齟齬が生じる．この問題を解決するために，URL のみではなく文書のタイムスタンプなどによってバージョンの違いを検出する必要がある．

また，ブックマークされた文書が削除され閲覧できない，変更された文書の以前の状態を閲覧したいといった状況が考えられる．ブックマーク時点の文書をアーカイブしておくなど，変更・削除されても閲覧できるような仕組みが必要である．

(5) 客観的な検証手法

今回行った検証においては，文書が有効であるかの判断が個人に依存し，客観性に欠ける．提案手法の有効性を客観的に評価するような検証手法の開発が必要である．

また，プロトタイプの運用によりデータを蓄積し，より多くのサンプルを用いた検証を行う必要がある．

文　献

[1] Social Bookmarking Tool Comparison
http://www.consultantcommons.org/node/239

[2] MeCab
http://mecab.sourceforge.jp/

[3] はてなブックマーク
http://b.hatena.ne.jp/

[4] del.icio.us
http://del.icio.us/

[5] 賀家智代，角谷和俊，“質問キーワードの順序依存性に基づく Web アーカイブ検索方式”，DEWS2006

[6] 中島伸介，黒田慎介，田中克己，“閲覧履歴を反映したコンテクスト依存型 Web ブックマーク”，TOD14

[7] 白井慧，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”，研究報告 情報処理学会, July 2006

[8] 松岡有希，坂本竜基，伊藤禎宣，武田英明，小暮潔，“Web 文書に対するマーキングからの個人知識の獲得”，The 20th Annual Conference of the Japanese Society Artificial Intelligence,2006

[9] 寺野隆雄，“Web 上の情報推薦システム”，情報処理，44 巻，7 号，July 2003

[10] 宇田隆幸，藤井敦，石川徹也，ユーザ投票と情報アイテム間類似度を併用した情報推薦システムー擬似投票方式の提案と評価ー“，研究報告 データベースシステム，January 2004