Vol.2010-GN-74 No.2
2010/1/21

# ブックマーク自動分類システムにおける分類成功率の評価

渡辺伸一[†]　服部哲[††]　速水治夫[†,††]

一度見つけた Web ページを再度見つけるためにブックマークが広く利用されている．しかしブックマークは分類の手間が大きい．また，分類を行わないと Web ページを探しづらくなる．そこで本論文では，ユーザの持っているカテゴリに自動分類を行うブックマークシステムを提案する．本システムはカテゴリ毎の Web ページの本文に含まれる単語を学習する．これを利用し，登録の際に最も近いカテゴリをナイーブベイズで決定し，分類する．実際のブックマークを登録して評価を行い，分類精度を調査した．

# An Experiment of Automatic Bookmark Classification System

SHINICHI WATANABE[†]　AKIRA HATTORI[††]
and　HARUO HAYAMI[†,††]

Many people use bookmark to re-find web pages found before. However, there are some problems of the classification of the bookmark. In addition, we cannot look for a Web page unless we classify it. In this paper, we propose the system that automatic bookmark classification to a category. Our system learns a word in the text of the Web page. At the time of registration, our system classify the nearest category using Naive Bayes method. We registered bookmark with a system and tested it and investigated form of the bookmark which system could classify definitely.

## 1. はじめに

Google や Yahoo などの検索エンジンの進化に伴い，目的の Web ページを素早く見つけることができるようになった．しかし，ニュース記事や参考資料は Web ページを移動する中で見つけるものが多い．また，検索エンジンに多数の条件を与えて見つける Web ページも存在する．これらを再度検索エンジンで見つけることは困難である．

Web ページに再度アクセスするための URL 管理ツールとして，Web ブラウザに付属されているブックマークがある．利用の手軽さから多くのユーザはこれを利用している．しかしブックマークには，分類を行わないと目的の Web ページを検索するのに時間がかかる点と，迅速な検索が可能なように分類するには手間が大きいという二つの問題点がある．

そこで本論文では，Web ページの URL をブックマークに登録する際に，本文の内容が近い Web ページが多く入っていると考えられるカテゴリに自動的に分類するシステムを提案する．自動分類には，迷惑メールの自動分類で実績を持つナイーブベイズを利用し，学習データにはユーザのブックマークに登録されている Web ページの本文中の単語とその出現頻度を利用した．

以下，２章で研究対象の現状と問題点，３章で提案手法，４章で試作システムの詳細，５章で評価実験，６章で結びを述べる．

## 2. 研究対象の現状と問題点

### 2.1 ブックマーク

ブラウジングの中で再度訪れたいと感じた Web ページの URL を保存する際，ユーザは Web ブラウザに付属しているブックマークを利用する．多くのブックマークでは，Web ページの URL を階層構造のカテゴリに分類し，Web ページ名を参照しながら検索する．

ブックマークには二つの問題点がある．

一つはカテゴリによる階層構造の管理を適切に行わないと検索が困難となる点である．ブックマークから Web ページを検索する場合，Web ページ名とカテゴリ名を目視で検索するしか方法がない．しかし，対象の Web ページ名がその内容自体を指し示しているとは限らない．そのため，階層構造の完成度を上げないと Web ページの検索に時間がかかってしまう．

もう一つは分類の手間が大きい点である．常に登録ページが増えていくブックマー

[†] 神奈川工科大学　大学院　工学研究科
Graduate School of Engineering , Kanagawa Institute of Technology
[††] 神奈川工科大学　情報学部
Infomation Faculty , Kanagawa Institute of Technology

クでは，毎回の分類や定期的な整理が必要である．URL の登録数が多いほどこの手間は大きくなり，ユーザは分類を行わなくなる．David らの研究では，ブックマークを利用しているユーザのうち，30%程度は単一階層でしか分類を行わず，35%程度は分類を行わないことが示されている[1].

この二つの点から，ユーザが有用な Web ページを見つけてブックマークに保存しても，後々検索が困難になることは明らかである．

### 2.2 関連研究・システム

#### 2.2.1 Web ページが持つ情報を利用したブックマークの分類，検索システム

従来のブックマークでは，URL と Web ページ名のみで検索を行う．Web ページはこれ以外にも多くの情報を持っているため，この情報を利用したブックマークシステムの研究がおこなわれている．

John らは，保存日時やアクセス回数，html メタタグなどをブックマークに保存することで高度な検索を利用できるとともに，情報の編集が行いやすいシステムを開発した[2]．メタタグ情報は自動分類に利用され，これをカテゴリとした管理が行われる．

Mac OS X では Spotlight 機能により，閲覧した Web ページから抽出したメタデータが保存されている．これによりブックマークや履歴の高速な全文検索が可能である．

検索に利用できる情報を新しく定義する研究も行われている．中島らは，ブックマークまでの閲覧履歴にユーザの意図が隠されていると考え，ユーザの閲覧履歴を反映した Web ブックマークの概念を提案した[3]．この研究では検索エンジンで見つけた Web ページをブックマークに加えるというプロセスにより，そのページの前後に閲覧した Web ページをもとにして代表キーワードやランキング値をメタデータとして保存するものである．

#### 2.2.2 ブックマーク機能の拡張

ブックマークの機能を様々な面から拡張し，管理の効率化を目指す研究が行われている．

Li らは，キーワードによるブックマーク自動分類等の機能を実装した拡張ブックマークシステムである PowerBookmarks を開発した[4] ．このシステムでは，ブックマークの自動追加やリンク切れの検知，他者とのブックマーク共有，スライドショーでの表示など，高機能なブックマーク機能を提供している．

#### 2.2.3 ソーシャルブックマーク関連

はてなブックマークや livedoor クリップなど，インターネット上にブックマークを公開し，不特定多数のユーザと共有できるサービスとしてソーシャルブックマークがある．多くのサービスは従来のブックマークと違い，カテゴリによる階層構造ではなくタグという名前のキーワードを付加して分類を行う．ユーザが多く，集合知を利用できるソーシャルブックマークでは，タグを利用した協調フィルタリングの研究が多数行われている．

#### 2.2.4 カテゴリの生成と自動分類システム

著者らは，日本語語彙体系を利用するカテゴリの生成と，それに対する自動分類システムを提案した[5]．このシステムは，日本語のシソーラスである日本語語彙大系をもとに対象ページの本文のジャンルを推定し，そのジャンルを階層構造として自動的に構築しながらカテゴリ分類を行う．更に登録日や文章量，アクセス回数などの検索補助を利用することにより，様々な記憶から目的の URL を見つけることができる．学習の必要がなく，ブックマークを持っていないユーザでもすぐに利用できる利点があるが，それぞれのユーザが自分好みの階層構造を利用できないという欠点があった．

### 2.3 解決手法

本論文ではブックマークへの登録時の分類の手間をなくすために，システムによる自動分類手法を検討する．html メタタグや共有タグなど，Web ページによって利用できる数に差があるものを利用して自動分類を行う手法は，その数によって提示される結果が左右されてしまう．Web ページから取得できる情報が多い場合には有効に動作するが，情報が少ない Web ページは埋もれてしまう可能性があるため，URL の管理には向かない．そこで本論文では，従来のブックマークと同様に，1 つの URL は 1 つのフォルダのみに存在する形をとり，そこに分類をおこなう．また，Web ページの情報の中で内容を直接的に示す「本文」を利用することでより正確な分類を行う．

また，多くの関連研究では，複数の画面を横断的に利用しなければならない問題点がある．ブックマークの利用時には，Web ページをブラウザで参照しながら管理や検索を行うことが多い．画面スペースを広く占有する高機能なシステムがあっても，ブックマークを利用しようとするたびにその画面に移動するのでは逆に効率が悪くなってしまう． そこで，ブラウジングを邪魔しないサイズでのインターフェイスに関しても検討を行う．

## 3. 提案手法

### 3.1 着眼点と提案手法

迷惑メールの分類手法として広く知られているものに，ナイーブベイズがある．この手法は古くから使われている手法ではあり，多くのシステムで実装されている実績を持つ．そこで，この手法をブックマークに適用することで自動分類を行うシステムを提案する．

この手法に限らず事前学習を必要とする手法では，適切なカテゴリに分類できるまでの初期学習がユーザのハードルとなる．これに対して多くのユーザがすでにブックマークを持っていると考えられるため，これをインポートすることで学習を行う利用形態を想定する．

### 3.2 ナイーブベイズによる自動分類

#### 3.2.1 ナイーブベイズ

ナイーブベイズは過去の事例をもとに，文書があらかじめ与えられているどのカテゴリに属すかを決定する分類手法である．ナイーブベイズを用いた Web ページの自動分類では，Web ページの本文に含まれる単語によって特定のカテゴリに分類される確率を計算する．カテゴリは単語とそれぞれの単語の頻度のリストで構成され，本文に含まれる個々の単語が特定のカテゴリに含まれる確率，URL がカテゴリに含まれる確率などから URL を分類するカテゴリを決定する．ナイーブベイズでは，単語が出現する確率は他の単語とは関係なく独立であると仮定している．

特定の Web ページ H がカテゴリ Ci に含まれる確率である P(Ci|H)は以下のように計算される．

$$P(C_i|H) = \frac{P(H|C_i) \times P(C_i)}{P(H)}$$

P(H|Ci)は H に含まれる単語がカテゴリ Ci の中に現れる確率であり，P(Ci)は与えられたカテゴリが選ばれる確率である．特定の URL が現れる確率である P(H)は，どの計算においても等しいのでこれを無視して以下の計算を行う．

$$P(C_i|H) = P(H|C_i) \times P(C_i)$$

H をどのカテゴリに分類するか計算するためには，この P(Ci|H)をそれぞれのカテゴリについて計算し，最大のものを探す．

まず，Web ページ H の本文を単語リストに分割する．そして P(H|Ci)としてそれぞれの単語 W の出現確率の積を計算する．これは単語 W が含まれる Web ページ数をカテゴリ中すべての Web ページ数で割ったものである．次に P(Ci)としてカテゴリ Ci に含まれる URL 数を分類されている全 URL 数で割る．これらの積から最大のカテゴリを選び分類を行う．

#### 3.2.2 ゼロ頻度問題

単語 Wi が Web ページ H に一度も現れない場合，Wi の出現率は 0 と推測される．本文の確率は単語列の積で計算するため，この場合，本文中のどれか 1 つの単語の発生確率が 0 だと，本文全体の確率も 0 になってしまう問題が発生する．この問題はゼロ頻度問題と呼ばれている．

この問題を避けるため，Wi の出現確率を求める際に単語の出現回数を補正した値を用いるスムージングという処理を行う必要がある．

本システムでは，仮の確率の 0.5 を出現確率として，実際の確率との平均を計算する．

#### 3.2.3 日本語形態素解析

対象ページの URL をブックマークに登録する際，本文を単語に区切り，頻度とともにデータベースに登録する．本文を単語に区切る形態素解析処理には Yahoo デベロッパーネットワークの日本語形態素解析処理を利用した．登録する品詞は形容詞，形容動詞，副詞，名詞，動詞，助詞，助動詞である．

### 3.3 システム環境

実際の利用シーンを想定するため，常駐しても邪魔にならず，検索を行えるスペースを検討する．ブラウザのブックマークで利用されるスペースは，アドレス部分の直下に存在するアドレスバーと，画面左側に縦長で表示されるものがある．アドレスバーはアクセスが容易なため頻繁に訪れる Web ページには便利であるが，スペースが限られるため多くの URL を置くことができない．そこで本論文では画面左側に常駐するシステムを提案する．また，ブラウザ依存を避けるため，アドオンでなく単体のソフトとして開発する．

## 4. 試作システム

### 4.1 システム概要

本システムは，Web ページを読込み単語に分割するデータベース登録部と，ベイズ推定を行ってカテゴリを決定するカテゴリ推定部に分かれる．

システムの構成図を図 1 に示す．

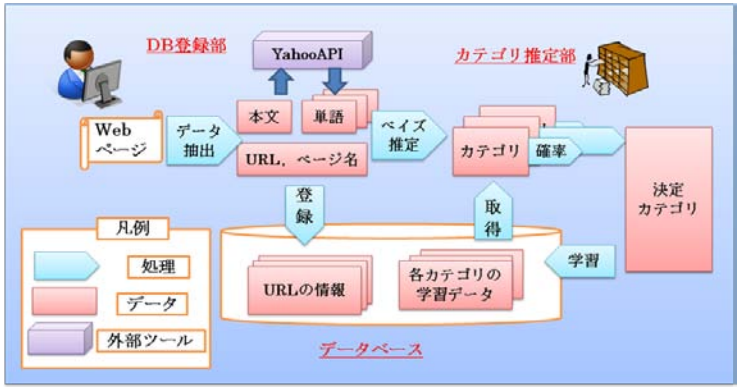


図 1　システム構成図

### 4.2 システム機能の詳細

#### 4.2.1 データベース登録部

登録部では，通常のブックマークと同様に URL と Web ページ名を登録する．その後，対象ページの HTML ソースを取得し，HTML タグを取り除く．そしてその本文を Yahoo 形態素解析で分割する．

#### 4.2.2 カテゴリ推定部

カテゴリ推定部では，分割した単語と各カテゴリの学習データを利用してベイズ推定を行い，もっとも確率の高かったカテゴリをその Web ページの入るカテゴリとして決定する．カテゴリ決定後にそのカテゴリと単語をデータベースに学習させる．

#### 4.2.3 その他機能

本システムでは，Internet Explorer 等のブックマークからインポートを行うことが可能である．操作を行うことで，登録されているすべての Web ページを本システムに登録することができる．

### 4.3 表示画面

#### 4.3.1 登録時表示画面

登録時の表示は，ブラウジングの邪魔を極力しないよう，アイコンのみの表示とした．ここにブラウザの URL 横のアイコンをドラッグアンドドロップすることでシステムにブックマークが登録される．通常時の表示画面と登録イメージを図 2 に示す．


図 2　登録時表示画面

#### 4.3.2 URL 検索時表示画面

検索時の表示は，ブラウジングしながらソフトを操作できるよう，縦長でブラウザの横に常駐きる形である．検索時の表示画面を図 3 に示す．

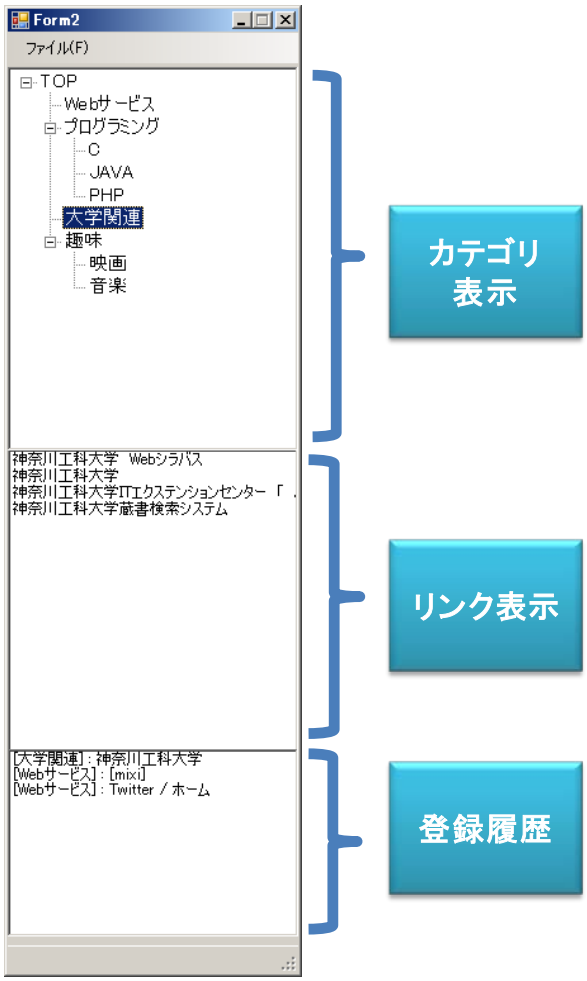


図 3　検索時表示画面

上部にあるのが階層構造であり，カテゴリを選択することで Web ページ名が中心に表示される．Web ページ名をクリックすることでその URL にアクセスできる．また，下部には，登録された URL が入っているカテゴリが表示される．

## 5. 評価実験

### 5.1 分類精度評価

#### 5.1.1 概要

試作システムの分類精度を評価する．実験データとして，実験協力者 3 名が実際に利用しているブックマークを利用した．これらのブックマークの最後に登録した URL が正しく分類できるかを調査するため，以下の実験を行う．

① 試作システムにブックマークをインポート
② URL1 つを一時的に削除し，システムに再登録
③ 分類の結果を記録
④ URL がもとのカテゴリに入らなかった場合，正しいカテゴリに移動
⑤ 2,3,4 をすべての URL で繰り返す

実験協力者 3 名のブックマーク情報を表 1 に示す.

表 1　ブックマーク詳細

| | URL数 | カテゴリ数 | URL / カテゴリ | 最大カテゴリ深度 |
|---|---|---|---|---|
| 協力者A | 136 | 26 | 5.2 | 2 |
| 協力者B | 62 | 10 | 6.2 | 1 |
| 協力者C | 149 | 16 | 9.3 | 1 |

#### 5.1.2 実験結果

分類精度を表 2 に示す. 有効 URL とは試作システムに登録を行うことが出来た URL であり，存在しなかった Web ページ，実験環境からはアクセスできないローカルのファイルなどが含まれる.

表 2　分類精度

| | 有効URL数 | 分類成功 | 成功確率 |
|---|---|---|---|
| 協力者A | 98 | 14 | 14.3% |
| 協力者B | 55 | 32 | 58.2% |
| 協力者C | 135 | 35 | 25.9% |

### 5.2 カテゴリ内 Web ページ本文の相互類似度との相関

#### 5.2.1 概要

どのようなカテゴリに対しての分類が成功しやすいかを調査する.

#### 5.2.2 コサイン類似度

文書間の類似度を求める方法の一つとして、コサイン類似度がある．これはベクトル化した文章の類似度を判断するものであり，以下の計算式で計算を行う.

$$\cos(V^c, V^g) = \frac{V^c \cdot V^g}{|V^c||V^g|}$$

コサイン尺度は 0 から 1 の値を取り、1 に近いほど類似性が高いことを示す．なお，この処理にはライブラリである Slothlib を使用した[7]．文章の特徴ベクトルの計算には tf-idf 法を利用した．これは，特定の文書にしか出現しない単語に重みを加えた方法である.

#### 5.2.3 手法

ブックマーク内の Web ページの本文すべての組み合わせでコサイン類似度の計算を行う．その後，カテゴリごとに平均をとり，カテゴリの分類成功率との相関をとった.

#### 5.2.4 結果

実験結果を表 3 と図 4 に示す.

表 3　相関

| | 相関 |
|---|---|
| 協力者A | 0.511647 |
| 協力者B | −0.07601 |
| 協力者C | 0.582752 |
| Webページ登録数が5以上のカテゴリ | 0.596471 |
| 全体 | 0.562518 |

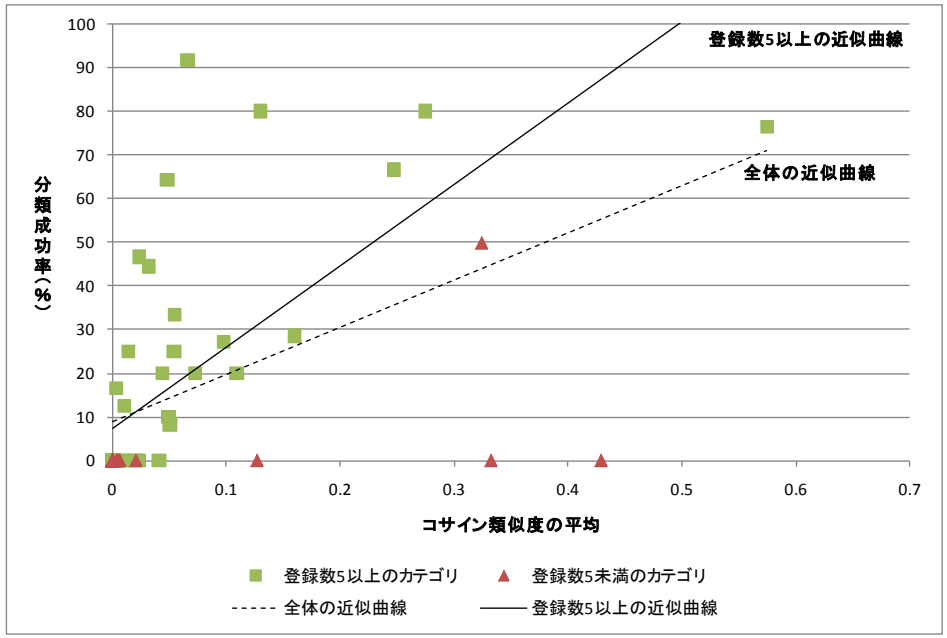


図 4 コサイン類似度と分類成功率の相関

3 人のデータを合わせたものでは，中程度の相関がみられた．また，Web ページ登録数が 5 未満のカテゴリではコサイン類似度の平均が高くても分類成功率が低かった．

### 5.3 考察

分類精度の結果から，利用するブックマークによって分類の成功率に大きな差が出ることが分かった．また，類似度の結果から，5 つ以上の Web ページの本文を学習させたカテゴリでは，より内容の近い Web ページが集まっているほど分類が成功しやすいことが分かった．

### 5.4 今後の課題

本手法では，分類手法としてナイーブベイズを利用した．分類手法は他にも多数提案されており，より精度が高いといわれている分類手法との比較が必要だと考える．また，実験件数を増やし，どのようなブックマークでより分類精度が上がるかを調査していく必要がある．更に，本論文では動作速度やインターフェイスの評価を行っていないため，この点の実験も今後行っていく．

## 6. おわりに

本論文では，ナイーブベイズにより自動的にカテゴリ分類を行うブックマークシステムを提案した．実験の結果，ブックマークによって分類精度に差が見られた．また，コサイン類似度との平均と分類成功率に相関がみられたことから，より本文が近い Web ページが入っているカテゴリの成功率が高いことが分かった．今後は実験を進め，精度を上げていく．

## 参考文献

1) David Abrams, Ron Baecker, and Mark H. Chignell. Information Archiving with Bookmarks: Personal Web Space Construction and Organization. In CHI, pp. 41–48,(1998).
2) Robert, J., Ruiz, J. and Lank, E. Making Favorites Useful( 2005).
3) 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．情報処理学会論文誌：データベース, Vol.43, No.SIG5(TOD14), pp.23-36(2002)
4) W.S. Li, Q. Vu, D. Agrawal, Y. Hara, and H. Takano:”PowerBookmarks: A System for Personalizable Web Information Organization, Sharing, and Management.” 8th International World Wide Web
5) 渡辺伸一，服部哲，速水治夫 ：ブックマーク自動分類システムの提案　情報処理学会　マルチメディア，分散，協調とモバイル (DICOMO2009)シンポジウム 論文集 pp. 235 - 240
6) Yahoo!デベロッパーネットワーク － テキスト解析 － 日本語形態素解析
http://developer.yahoo.co.jp/webapi/jlp/ma/v1/parse.html
7) 大島裕明，中村聡史，田中克己 ：SlothLib: Webサーチ研究のためのプログラミングライブラリ 日本データベース学会Letters Vol.6, No.1, pp.113-116