グループウェアと
ネットワークサービス 44－9
（２００２．５．２４）

ブックマークの協調フィルタリングを利用したディレクトリ管理

鵜飼孝典，三末和男
富士通研究所ドキュメント研究部

住所：川崎市中原区上小田中 4-1-1
TEL：044-754-2671
EMAIL：ugai@jp. fujitsu. com, misue. kazuo@jp. fujitsu. com

概要

有用な情報(URL)を階層化されたカテゴリに分類して整理し，提供するYahoo!のような文書ディレクトリは，情報を探すためのツールとしても情報共有システムとしても非常に有効である．しかし，こういった文書ディレクトリの維持管理では，優良なURLを探し出して，適切なカテゴリに分類を行なうことに非常にコストがかかる．本稿では，サーバに保存されているユーザのブックマークと文書ディレクトリに対して協調フィルタリングを行なうことで，ディレクトリに対してURLのリコメンデーションを行なうシステムについて述べる．内容分類を用いた手法との比較実験で，相互に補完し合うことでより良い結果がえられることがわかった．

キーワード：　ブックマーク，知識共有

A directory maintenance system with social filtering of users' bookmarks and the directory

Takanori Ugai, Kazuo Misue
Fujitsu Laboratories Limited
Address:4-1-1 Kamikodanaka, Nakaharaku, Kawasaki
TEL: 044-754-2671
EMAIL: ugai@jp. fujitsu. com, misue. kazuo@jp. fujitsu. com

Abstract

Yahoo! like document directory which classifies useful information (URL) into the category and hierarchies them is a very effective system as the tool to look for useful information or one of informational common systems. However, it costs for searching out excellent URL, and the classification into an appropriate category in the maintenance management of such a document directory. We describe a system that recommends URL to the directory in the social filtering the user preserved bookmarks in the server and the directory. We got the better result to supplement with the comparison experiment with the technique that uses the content classification.

Keywords: Bookmark, Directory, Social filtering, Knowledge sharing

1. はじめに

インターネットやイントラネットにおいて近年一般のユーザが自由にホームページを開設することが容易になった．その一方，本当に有用な情報が何処にあるのか探すことが難しくなった．このような状況において有用な情報(URL)を階層化されたカテゴリに分類して整理し，提供する Yahoo!のような文書ディレクトリは，情報を探すためのツールとしても情報共有システムとしても非常に有効である．しかし，こういった文書ディレクトリの維持管理では，優良な URL を探し出して，適切なカテゴリに分類を行なうことに非常にコストがかかる．例えば Yahoo![7] では数百人もの「サーファ」と呼ばれる専門化がこれを行なっている．

これに対して我々は，Google の PageRank と同じようなアルゴリズムを選別に用いて優良な URL を自動的に収集し，内容の類似性による自動分類を用いたカテゴリへの自動割当を行なうシステムを開発し，実験，運用を行なってきた．この手法では，新しく作成された URL がたとえ優良なものであっても収集されない，部門情報など内容ではなく文書の所属組織などで分類しているカテゴリでは，分類誤りが増えるという問題がある．

本稿では，サーバに保存されているユーザのブックマークと文書ディレクトリに対して協調フィルタリングを行なうことで，ディレクトリに対して URL のリコメンデーションを行なうシステムについて述べる．本システムでは，他からリンクされていないページで口コミによって知られているような URL でもユーザのブックマークに登録されていればリコメンドでき，協調フィルタリングを用いているので，文書の内容とは異なった観点で分類されているカテゴリに対してもリコメンデーションを行なうことができる．またユーザはブックマーク管理システムを利用しているだけで自動的にディレクトリに情報提供を行なうことができ，自然と情報共有が行なわれる．

以下第2章では既存の技術について述べ，第3 章では我々が開発しているシステムとシステムで用いている協調フィルタリングのアルゴリズムについて述べ，4 章で我々が開発した手法の実験結果を示す．最後に 5 章で，本稿で報告する手法について考察を加える．

2. 既存技術

本章では，我々がこれまでに開発した，内容の類似性による自動分類を用いたカテゴリへの自動割当を行なうシステムと，本稿で述べるシステムと同様の手法でブックマークの維持管理コスト軽減化のために，有用な URL を推薦するシステムについて述べる．

2.1. 自動分類による URL 自動推薦システム

我々は文書ディレクトリの維持管理作業の軽減化を目的として，URL の自動推薦システム[2]を開発した．このシステムはインターネットやイントラネットからディレクトリに載せるべき優良な URL を収集する部分と，URL を適当なカテゴリに分類する部分から構成されている．このシステムは Google[6] の Page Rank を改良したアルゴリズム[5]を用いて優良 URL を選び出す．得られた URL を各カテゴリに分類する部分では，URL が指す文書をサンプル文書として学習し，文書の特徴を示すキーワードベクトルの余弦や距離を利用するベクトル空間法，カテゴリとの関連度を数値化した分類規則をあらかじめ用意するルールベース法，単純な分類方法を繰り返し実行して分類結果の多数決を利用するブースティング法の３種類の方法で分類する．

このシステムの優良 URL を選び出す部分では，新しく作成されて他からはまだ余り多くリンクされていない URL はたとえ優良であっても収集されないという問題がある．また URL を分類する部分では，文書の内容が似ているカテゴリについては高い率でそのカテゴリにふさわしい URL が得られる．しかし部門情報など内容ではなく文書の所属組織などで分類しているカテゴリでは，分類誤りが増え，自動推薦の効果が低くなるという問題がある．

2.2. 協調フィルタリングを用いたブックマーク管理システム

われわれはこれまでにブックマークの維持管理軽減化のために，優良 URL を推薦するブ

ックマーク管理システム[1]を開発してきた．このシステムではブリンク[8]と同様にサーバ上にカテゴリ毎に分類されているブックマークを他人のものと比較して，同じURLを持つカテゴリを探し出して，そのカテゴリとの差分を推薦URLとしてユーザに示す．システムは他のユーザの各カテゴリと同様に，Yahoo!やOpen Directory Project[9]，社内ポータルなどの大規模ディレクトリの各カテゴリとも比較する．

大規模ディレクトリの各カテゴリを他のユーザと同様に協調フィルタリングに用いてブックマークにURLを推薦することには次の2つの利点がある．

1. 協調フィルタリングを用いることで文書の内容の類似性に基づいて分類する方法と比較して，部門情報など内容ではなく文書の所属組織などのメタ情報で分類しているカテゴリでも正解率が高くなる．
2. カテゴリ単位での協調フィルタリングを行なうので，ディレクトリのカテゴリ構造とブックマークのカテゴリの構造が全く異なっていてもかまわない．
3. 数十人から数百人の小人数でも網羅的な大規模ディレクトリとの比較によって有効なURLの推薦を得ることができる

ユーザはシステムから推薦されるURLをチェックして，有用であれば自分のブックマークに登録する．

このシステムは，有用なURLを推薦する以外に次の機能を提供する．

1 登録されているURLに定期的にアクセスしてリンク切れを起こしていないかチェックすること．URLに変更があり，リダイレクトされているときは，自動的にURLを変更する．

2 URLを登録するときに，コンテンツから自動的にタイトル，キーワードを取り出し，入力を軽減化する．

3. システムの概要

本章では，我々が開発している文書ディレクトリの維持管理軽減化のためにシステムが提供する機能について述べる．まず有用なURLを推薦する手法について述べ，その後，それ以外にシステムが提供する機能について述べる．

3.1. ブックマークとディレクトリの協調フィルタリング

本システムではサーバ上にカテゴリ毎に分類されているユーザのブックマークとディレクトリの各カテゴリを比較して，同じURLを持つカテゴリを探し出して，そのカテゴリとの差分を推薦URLとしてディレクトリに推薦する．

ユーザのブックマークと協調フィルタリングを行なうことには次の2つの利点が期待される．

1 2.2節で述べたブックマークへの推薦と同様に内容ではなく文書のメタ情報で分類しているカテゴリでも正解率が高くなる．

2 ユーザはブックマーク管理システムを利用しているだけで自動的にディレクトリに情報提供を行なうことができ，自然と情報共有が行なわれる．

ディレクトリ管理者はシステムから推薦されるURLをチェックして，有用であればディレクトリに登録する．

3.2. 本システムで用いたアルゴリズム

本節では，本システムで用いた協調フィルタリングのアルゴリズムを定義し，例を用いて説明する．

定義 1:URLの集合をURLとするとき，カテゴリCは$C \subseteq URL$と定義する．

定義 2:カテゴリAとBの類似度Fを$F(A, B)=|A \circledcirc B|/|A \cup B|$と定義する．ただし$\circledcirc$は排他的論理和演算子を示す．

定義 3:uというURLの人気度$P(u)$を$P(u)=|\{C|u \in C\}|/|AC|$と定義する．ただし$AC$はすべてのカテゴリとする．

定義 4:閾値rとしたとき，カテゴリCに推薦するURLの集合$R(C, r)$を$R(C, r)=\{u|u \in A \wedge F(A, C)<r \wedge \neg u \in C\}$と定義する．

定義 5:$MAX(U, m)$をUの要素から人気度が大きい順にm個集めた集合と定義する．

記法:$MAX(R(A, r), m)$を$Max(A, r, m)$とする．

本システムは，他のユーザのカテゴリ，他のユーザと一緒に用いる大規模ディレクトリの各カテゴリを集合 *URL* とする．そして類似度の閾値 r と推薦する数 m を定数としてユーザの各カテゴリ C に対して，*MAX (C, r, m)* を算出して，人気度の大きい順に並べてユーザに提供する．

つぎのような3つのカテゴリのそれぞれに○がついている URL が登録されているとする．閾値 r を 0. 5，最大推薦数 m を 2 とする．

表 1:カテゴリの例

カテゴリ	URL1	URL2	URL3	URL4	URL5	URL6
A	○	×	○	○	○	×
B	×	○	○	○	○	○
C	○	○	○	○	○	×

A と B, A と C はどちらも 3 URL が一致し，類似度は 0. 75 になる．A に推薦すべき URL を示す R(A, 0. 5)は B と C の URL のうち A に含まれない URL{URL2, URL6}となる．URL2 と URL6 の人気度はそれぞれ 0. 67 と 0. 33 となる．MAX(A, 0. 5, 2)は{URL2, URL6}となり，システムは A に対して URL2, URL6 の順に推薦する．

3.3. ユーザの利用頻度の利用

ブックマークにおいては，一度登録したがアクセスしなくなる URL が少なくない．しかもそのような URL も削除されずに残る．そこで推薦する URL の決定要素として，アクセス回数や，アクセス履歴を考慮することが必要であると考えられる．

定義 6:[利用頻度の利用]url u の利用回数が *Count(C, u)* で与えられているとしたとき，URL u の人気度 $P1(u)= \Sigma Count(C, u), u \in C$ と定義する．

定義 7:[利用履歴の利用]カテゴリ C における url u の利用履歴が *History(C, u)* が現在からの時間の列(t1, t2, t3, t4, . .)で与えられているとしたとき，URL u の人気度 $P1(u)= \Sigma \alpha \times t$ と定義する．ただし α は減衰定数とする．

3.4. カテゴリの粒度の大きさの吸収

大規模ディレクトリは網羅性が高く，広い範囲のカテゴリを持っているが，そのため専門性が低くあまり詳細に分割していないことがある．このような場合，ユーザのカテゴリがより詳しく細かく分類していて，本システムのアルゴリズムがうまく働かないことがある．

表 2:カテゴリの大きさに差がある例

カテゴリ	URL1	URL2	URL3	URL4	URL5	URL6
A	○	○	○	○	○	○
B1	×	○	○	×	×	×
B2	×	×	×	○	○	×

例えば，表 2のようにディレクトリが A というカテゴリ，あるユーザが B1 と B2 というカテゴリを持っているとする．B1 と B2 に含まれる URL がすべて A に含まれる．この場合は B1 と B2 に URL1 と URL6 を推薦する．これによって無駄な推薦を減らすことができる．

3.5. その他の機能

本システムはディレクトリの維持管理軽減化のために，有用な URL を推薦する以外に次の機能を提供する．

1. 登録されている URL に定期的にアクセスしてリンク切れを起こしていないかチェックすること．URL に変更があり，リダイレクトされているときは，自動的に URL を変更する．
2. URL を登録するときに，コンテンツから自動的にタイトル，キーワードを取り出し，入力を軽減化する．
3. 情報の編集は，フォームを通して行ない，各カテゴリの画面はテンプレートを利用して作成されるために，統一性を保つことができる．
4. 更新されたカテゴリや URL の一覧を“新着情報”として自動的に作成する．

3.6. 動作画面例

図 1は文書ディレクトリの管理者向け編集画面の例である．図 2は編集者向けの推薦された URL の表示，登録画面である．カテゴリごとにまとめて推薦された URL が表示される．管理者が推薦された URL を有用だと判断したら登録する．

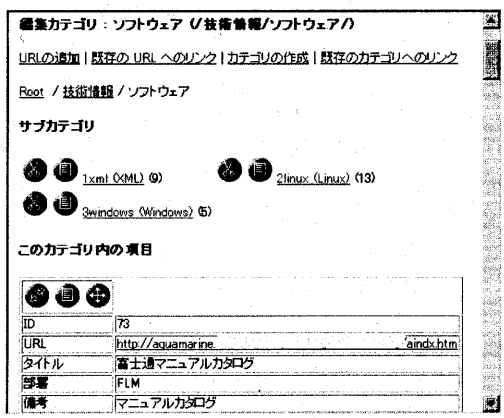

図 1:管理者向け編集画面例

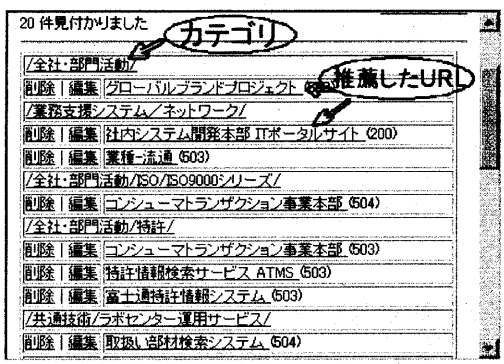

図 2:URL 推薦画面例

4. 評価実験

本章では，システムが提供する協調フィルタリングを用いて有用なURLを推薦する機能について行なった評価実験の結果について述べる．

4.1. 実験に用いたデータ

本節では，実験に用いたブックマークのデータ，推薦対象に用いた大規模ディレクトリの諸データについて述べる．

表 3:全ブックマークとカテゴリ

ユーザ数	32
全カテゴリ数	71
全URL数	281

表 3は現在システムが管理しているすべてのブックマークの大きさである．2002 年 1 月から社内向けにサービスを開始している．

表 4:推薦対象に用いたディレクトリ

対象ディレクトリ	URL数	カテゴリ数	データ名
Open Directory Japanese/コンピュータ/プログラミング言語	173	12	ODPJ
Open Directory Computer/Programming/Languages	9746	543	ODP
社内独自ディレクト	1560	210	INTRA

Open Directory はボランティアによって維持管理されている大規模ディレクトリで，そのデータは商用，非商用を問わず自由に利用することができる．今回用いたのは，その一部である．社内独自ディレクトリは，社内全体を広く浅く網羅するポータルで 1560 の内 1450 の URL が社内の URL である．

4.2. 実験方法

実験は各カテゴリにおのおの最大 5 つの URL を推薦することとし，3.2節で定義した *MAX* を，閾値を変えて算出した．正解数は，実験者が採用しても良いと判断したものを正解とし数えた．

4.3. 実験結果

協調フィルタリングによる推薦

表 5は協調フィルタリングを用いて推薦を行なった実験の結果である．

表 5:協調フィルタリングの結果

ディレクトリ	推薦カテゴリ	推薦数	正解数
ODP	5	7	3
ODPJ	4	6	3
INTRA	17	21	18

ODPについては543のカテゴリに対して5のカテゴリに何らかの推薦が得られた．ODPのこの部分は詳細化が進んでいて重要なものはかなり網羅されているため多くの推薦は得られなかった．推薦数7に対して正解数は3であるが，英語のコンテンツを含むURLだけを正解としたため3にとどまったが内容的には

もう一つ正解と出来るものがあった．ODPJいついては12のカテゴリに対して4つのカテゴリに何らかの推薦が得られた．こちらは日本語主体のコンテンツを正解としたが，内容的には ODP と同様に4つの正解が得られた．INTRA については 210 のカテゴリに対して，17 のカテゴリに推薦が得られた．21 の推薦数に対して 18 の正解で 85%の正解率となった．

内容分類による推薦

はユーザがブックマークに登録している 281 の URL を内容分類によって各カテゴリに推薦した結果である

表 6: 内容分類による推薦結果

ディレクトリ	推薦カテゴリ	推薦数	正解数
ODP	14	20	3
ODPJ	6	8	3
INTRA	30	43	13

5. 考察

実験の結果から閾値を大きくすると，適合率(正解数／推薦数)が大きくなり，閾値を小さくすると適合率が小さくなることがわかる．これはディレクトリのカテゴリとの共通 URL が多いブックマークから得られた推薦は正解となりやすいことを意味する．このことは類似性の高いカテゴリ同士からえた URL が推薦されるべき URL となるべきという直感と一致するものである．

内容分類との比較実験において，内容分類を用いたシステムでは有効な推薦が得られなかった，“部門別座席表”というカテゴリに対して協調フィルタリングを用いたシステムでは2つの URL を推薦し，いずれも正解であった．この結果から，今回開発したシステムが，部門情報など内容ではなく文書の所属組織などで分類しているカテゴリでも有効な URL を推薦できることが示されたといえる．

今後大人数，多くのカテゴリのデータを収集して本手法の有効性を確認したいと考えている．

推薦アルゴリズムのパラメータとなる類似度の閾値と推薦数は，次のような方針で運用フェーズによって使い分けることが有効であると考えている．1) ディレクトリの立ち上げの段階では掲載 URL を増やすために多少適合率が低くても編集者によって適当なカテゴリに振り分けなおすことができるように，類似度の閾値を小さく，推薦数を大きくする．2) 運用段階で十分 URL が集まっている場合は，不要な URL が掲載されることが無いように適合率が高くなるように類似度の閾値を大きくする．

6. まとめ

本稿では，文書ディレクトリの運用支援を目的として，URL の収集，分類を半自動化する URL 推薦システムについて報告した．本システムはサーバに保存されているユーザのブックマークと文書ディレクトリに対して協調フィルタリングを行なうことで，ディレクトリに対して URL の推薦を行なう．内容分類を用いた手法との比較実験で，相互に補完し合うことでより良い結果がえられることがわかった．推薦アルゴリズムのパラメータとなる類似度の閾値と推薦数は，運用フェーズによって使い分けることが有効であると考えられる．

参考文献

1. 鵜飼，三末:人と大規模ディレクトリの協調によるブックマーク管理，情報処理学会グループウェア研究会(2002)
2. 鵜飼，片山，津田: 文書ディレクトリ管理のための自動収集，自動分類の利用，人工知能学会全国大会 (2001).
3. 鵜飼，イントラネット向けディレクトリ管理システム，情報処理学会グループウェア研究会(1999)
4. 片山: 多様な要求に対応するテキストの自動分類システム，情報処理学会 62 回全国大会(2001).
5. Tsuda, Ugai, Misue: Link-based Acquisition of Web Metadata for Domain-specific Directories, PKAW2000, pp. 317-324 (2000).
6. Google: http://www. google. com/
7. Yahoo: http://www. yahoo. com/
8. Blink: http://blink. co. jp/
9. Open Directory Project: http://dmoz. org/