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

鵜飼孝典, 三末和男
富士通研究所ドキュメント処理研究部

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

インターネットユーザの多くが, 良くアクセスする URL を自分のブラウザにブックマークとして保存している．しかしながら，URL は変化が激しく，すぐに陳腐化するという問題がある．本稿では, ブックマーク内の URL を新鮮で有効なものに保つことを目的とするブックマーク管理システムについて報告する．本システムのアイデアは協調フィルタリングをベースとし，Yahoo! や OpenDirectory Project のような大規模ディレクトリを一ユーザとして扱うというものである．評価実験により，ユーザ数が少なくても有効な URL の推薦が可能だとわかった.

キーワード： ブックマーク， 知識共有

A bookmark maintenance system with collaboration among users and large-scale 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

Most Internet users keep URLs that they are going to access often in their browsers as the bookmarks. Bookmarks in the browser are easy to be obsolete. In this article we describe a bookmark maintenance system that supports to keep users' bookmarks fresh and useful. The system recommends the useful URL to users with social filtering the users and large directory like Yahoo and the Open Directory. The collaboration of users and such large-scale directory enable useful recommendation to the small number users.

Keywords: Bookmark, Directory, Social filtering, Knowledge sharing

1. はじめに

インターネットユーザの多くが，良くアクセスする URL を自分のブラウザにブックマークとして登録している．

しかしながら，URL は変化が激しく，すぐに陳腐化するという問題がある．たとえば，URL の指す Web ページが無くなる／移動する，いわゆるリンク切れが代表的である．Web ページの内容が古くなり役立たなくなるということもある．他の有益な Web ページが増えているのに登録されないというのもブックマークにとっては陳腐化と言えよう．また，登録 URL の増加に連れブックマークの分類体系が合わなくなり破綻するという問題もある．

このような問題に対して，ブックマークを新鮮で有効なものに維持することは，規模の違いはあるものの，Yahoo!や Open Directory Project のような Web 文書のディレクトリの維持と同じような手間が掛る．実際，インターネットやイントラネットの文書ディレクトリの維持管理は，多大なコストをかけて行なわれている．そこでは維持管理コストの低減が重要な課題である．

我々はブックマークの協調フィルタリングをベースとし，文書ディレクトリを一ユーザとして参加させるという方式を試みている．文書ディレクトリをユーザとして参加させることで，参加ユーザが比較的少数でも有効な推薦が受けられ，逆に参加人数が増えてくれば，ディレクトリの維持管理側はユーザからのフィードバックを受けることで，維持管理に役立つ情報を自然に獲得できるというメリットが期待できる．

以下第２章では，既存の技術について述べ，3 章では我々が開発しているシステムについて述べ，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

ブックマークの陳腐化を防ぐためには自分が保存している URL よりももっと新しい情報を含む URL や同じような情報でももっと良い URL を捜し出して，適切なカテゴリに分類を行なうことが重要である．本章では,有用な URL を推薦する既存の手法について述べる．

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を目的として,URL の自動推薦システム[1]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[4] の Page Rank を改良したアルゴリズムを用いて優良 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インターネット上のサービス

ブリンク[6]では，ブックマークをサーバに保存し，他人のブックマークと比較して，同じ URL を持つ人を捜し出して，その人のブックマークとの差分を示すことで個人のブックマークを新鮮に保つサービスを提供している．

この方法では，インターネットで十分に多くのユーザ(数万人以上)のブックマークを管理している場合には有効であるが,イントラネットのユーザを想定した場合(数十人から数百人)ではうまく機能しない．また,ユーザグループがわかれているときに，両方のグループにまたがる URL を集めているユーザが数多くいないと URL が共有されないという問題がある．

3. システムの概要

本章では，ブックマークの維持管理コスト軽減化のために，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. ユーザと大規模ディレクトリの協調フィルタリング

本システムではブリンク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 を持つカテゴリを捜し出して, 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[7], 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ることには次の３つの利点が期待される.

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に比較して, 部門情報など内容ではなく文書の所属組織などで分類しているカテゴリでもふさわしい URL がえられる.
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

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:カテゴリの集合を L, URL の集合を U とするとき, カテゴリ $C \in L$ は $C \subseteq U$ とする.

定義 2:カテゴリ A と B の類似度 F を $F(A,B)=|A \cap B|/|A \cup B|$ と定義する.

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:$TOP(U,m)$ 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:$TOP(R(A,r),m)$ を $Tm,r(A)$ 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, Tm,r（C）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３URL 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. $T(A)$ 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で登録したまま使われない URL や以前は利用したが最近は使われなくなった URL を排除することが出来る.

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 $P1(u)= \sum Count(C,u),\ u \in C$ 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 $P1(u)= \sum \alpha \times t$ 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 をまとめて一つのカテゴリとして取り扱い，それぞれに URL1 と URL6 を推薦する．これによって B1 に対して URL4,URL5 を推薦するような無駄を減らすことができる.

3.5. その他の機能

本システムはブックマーク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. どこに登録したか忘れた場合，自分が登録した内容をキーワードで検索できる．

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, ユーザと同じ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	25
全カテゴリ数	67
全URL数	258

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 月からサービスを開始している.

推薦の対象にしたカテゴリ(以下 U1,U2...,U5 と記す)は，ある一人のユーザのカテゴリ５つとする．XML は社内の URL が３つ社外の URL が３つ, JAVA,PERL にはすべて社外の URL が登録されているが，JAVA は３つが日本語のコンテンツを含む URL で PERL はすべて英語のコンテンツを指す URL である.社内手続き，製品担当連絡先はすべて社内の URL が登録されている．社内手続きには社内で良く使われる URL が登録されており，同じ URL を数多くのユーザが登録している.

表 4: 推薦対象のカテゴリ数

カテゴリ名	登録URL数
JAVA	5
XML	6
PERL	4
社内手続き	7
製品担当連絡先	12

表 5:実験に用いた大規模ディレクトリ

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

表 5は協調フィルタリングに用いた大規模ディレクトリである. 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(U,0.5,2)* を算出した．正解数は，ユーザ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

ユーザだけによる協調フィルタリング

カテゴリ名	推薦数	正解数
JAVA	3	2
XML	5	1
PERL	4	1
社内手続き	2	0
製品担当連絡先	1	0

JAVA,XML,PERL の技術情報は，ユーザの興味がばらけていてみなが同じ URL を持っていないため正解率が低い.一方社内手続きはみなが同じ URL を集めているために推薦されるものが少なかった.また推薦されたのが部署毎に異なる手続きであったために正解とならなかった. 製品担当連絡先は,他におなじ URL を持った人がいなかった.

ユーザとディレクトリの協調フィルタリング

25 人のユーザにそれぞれのディレクトリを加えて協調フィルタリングを行なった結果である.

ディレクトリ：ODPJ

カテゴリ名	推薦数	正解数
JAVA	5	3
XML	5	1
PERL	4	1
社内手続き	2	0
製品担当連絡先	1	0

ODPJ は日本語コンテンツのみが含まれているので PERL には ODPJ からの推薦は無かった.また XML のコンテンツがあまり含まれていなかったため ODPJ からの推薦が無かった.

ディレクトリ：ODP

カテゴリ名	推薦数	正解数
JAVA	5	2
XML	5	1
PERL	5	3
社内手続き	2	0
製品担当連絡先	1	0

ODP は英語コンテンツのみが含まれているので PERL には PERL への正解数が増えた.また XML のコンテンツはユーザからの推薦だけで上位 5 位を占めたため正解数は増えなかった.

ディレクトリ：INTRA

カテゴリ名	推薦数	正解数
JAVA	5	2
XML	5	1
PERL	5	1
社内手続き	5	3
製品担当連絡先	5	3

INTRA は社内コンテンツに関して有効に働いた.

ディレクトリに重みをつけた場合,

ディレクトリの重みを 3 にした場合, URL の人気度計算の際にディレクトリに含まれる URL は 3 つのカテゴリが含んでいるとして計算する. 2 人のカテゴリに含まれる URL よりもディレクトリに含まれる URL を優先的に推薦する.

ディレクトリ：ODP

カテゴリ名	推薦数	正解数
JAVA	5	2
XML	5	2
PERL	5	3
社内手続き	2	0
製品担当連絡先	1	0

ディレクトリ：INTRA

カテゴリ名	推薦数	正解数
JAVA	5	2
XML	5	2
PERL	5	1
社内手続き	5	3
製品担当連絡先	5	3

ODPJ を用いた場合は，重みが１の場合と結果に変化は無かった.ODP と INTRA を用いた場合に XML において正解数が増えた.これはユーザが持っていた XML に関する URL にあまり有効なものが無かったことを示す.

次にディレクトリの重みを１０にした場合を示す.この場合ディレクトリに含まれる URL は必ず採用され，ディレクトリから推薦候補が得られない場合にのみ他のユーザから推薦する.

ディレクトリ：ODPJ

カテゴリ名	推薦数	正解数
JAVA	5	1
XML	5	2
PERL	4	1
社内手続き	2	0
製品担当連絡先	1	0

ディレクトリ：ODP

カテゴリ名	推薦数	正解数
JAVA	5	2
XML	5	1
PERL	5	2
社内手続き	2	0
製品担当連絡先	1	0

INTRA を用いた場合は, 重みが 3 のときと変化が無かったが，ODPJ を用いた場合，JAVA において，ODP を用いた場合 PERL において正解数が減った.これは他のユーザから得た有効な URL が推薦されなくなったことによる.

参考実験

すべてのディレクトリを一緒に用いた場合,

カテゴリ名	推薦数	正解数
JAVA	5	2
XML	5	2
PERL	5	3
社内手続き	5	3
製品担当連絡先	5	3

本実験では各ディレクトリの重みはそれぞれ1とした.

内容分類による推薦システム

カテゴリ名	推薦数	正解数
JAVA	5	2
XML	5	2
PERL	5	1
社内手続き	5	1
製品担当連絡先	5	0

分類対象とした URL は社内外から文献[6]の Page Ranking を改良したアルゴリズムで抽出した 2000 の URL を用いた.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，正解が得られない.

Blink:

カテゴリ名	推薦数	正解数
JAVA	5	2
XML	5	1
PERL	5	1

各カテゴリに登録されている URL を Blink に登録して, 類似 URL の検索を行ない, 上位５つについて評価した.セキュリティ上の理由から社内 URL はのぞいた. Blink から ODP の重み 10 のときと同じ URL の推薦が 15 のうち 10 あった.

4.4. 考察

実験の結果から，大規模ディレクトリを協調フィルタリングに用いることで，有効な URL を得られることがわかった.ディレクトリによって得手，不得手があるため社内のポータルサイトなどと ODP のようなインターネットの大規模ディレクトリを組み合わせることでより良い結果が得られることがわかる.大規模ディレクトリを適当な重みをつけて利用することで少人数のユーザの場合でも有効な URL を推薦できることがわかった.

5. 他の応用

ディレクトリへの推薦

本稿では，ユーザのブックマークの維持管理に協調フィルタリングを用いた結果を報告したが，同じシステムで，ディレクトリに対して，URL を推薦することができる．多くのユーザが，登録している URL をディレクトリにすすめることが出来る.

パーソナライズ版ディレクトリの提供

協調フィルタリングにおいてディレクトリの重みを大きくすると，ディレクトリに登録されている URL が多く薦められるようになる．これによってユーザは，ディレクトリのパーソナライズ版として利用することが出来る.本手法では, カテゴリの構造，名前にかかわらず有効であるので，ユーザは，ディレクトリの一部を，自分の好きな階層構造で，自分の好みで取捨選択，追加したディレクトリを手に入れることが出来る.

6. まとめ

本稿では，ブックマーク内の URL を新鮮で有効なものに保つことができるブックマーク管理システムについて報告した．本システムではブックマークの管理に，Yahoo! や ODP のような大規模ディレクトリを含めた協調フィルタリングを行なうことで，少人数のユーザでも有効な URL の推薦を行なうことが出来る.実験からイントラネットで利用する場合にはイントラネット上のポータルサイトと，インターネットのサービスを組み合わせて利用すると良いことがわかった. 本稿では，3.4節で示した，カテゴリの詳細度に差がある場合の実験結果を示していない．今後これについての実験，分析を行ないたいと考えている．また大人数，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．また，本手法をディレクトリ管理に用いるシステムの開発を行ないたいと考えている.

参考文献

1. 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).
2. 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)
3. 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).
4. Google: http://www.google.com/
5. Yahoo: http://www. yahoo. com/
6. Blink: http://blink.co.jp/
7. Open Directory Project: http://dmoz.org/