DEWS2004 6-B-04

ブックマークの階層構造情報を組み込んだ協調フィルタリングによるWebページの推薦手法

佐保田圭介† 波多野賢治† 宮崎 純† 吉川 正俊†† 植村 俊亮†

† 奈良先端科学技術大学院大学 情報科学研究科 〒 630–0192 奈良県生駒市高山町 8916–5
†† 名古屋大学 情報連携基盤センター 〒 464–8601 愛知県名古屋市千種区不老町
E-mail: †{keisuk-s,hatano,miyazaki,uemura}@is.aist-nara.ac.jp, ††yosikawa@itc.nagoya-u.ac.jp

あらまし ブックマークは，利用者の嗜好が顕著に反映された情報である．そのため，協調フィルタリングによるWeb ページ推薦において，ブックマークは共有情報として用いられている．これまでの研究では，ブックマークから得られる特徴として，単語の出現頻度のみが用いられてきた．しかしながら，ブックマークは，その構造にも利用者の嗜好が反映されていると考えられる．本稿では，ブックマークの構造も共有情報として組み込んだ協調フィルタリングに基づく Web ページの推薦手法を提案する．我々は，利用者の類似度判定の基準として，単語の出現頻度に基づいた Web ページ間の類似度に加えて，ブックマークの階層構造から得られるフォルダの類似度，構造の類似度の二つを利用する．本手法によって，単語の出現頻度に基づいた Web ページ間の類似度のみを用いる場合に比べ，利用者の嗜好をより反映した他の提案システム利用者の Web ページを推薦することが可能となった．
キーワード WWW，協調フィルタリング，推薦，ブックマーク

A Web Page Recommendation Method with Collaborative Filtering Using User's Bookmark Hierarchy

Keisuke SAHODA†, Kenji HATANO†, Jun MIYAZAKI†, Masatoshi YOSHIKAWA††, and Shunsuke UEMURA†

† Graduate School of Information Science, Nara Institute of Science and Technology
8916–5, Takayama, Ikoma, Nara 630-0192 Japan
†† Information Technology Center, Nagoya University
Furo, Chikusa, Nagoya, Aichi 464–8601 Japan
E-mail: †{keisuk-s,hatano,miyazaki,uemura}@is.aist-nara.ac.jp, ††yosikawa@itc.nagoya-u.ac.jp

Abstract Users' bookmarks reflects in the users' favorites. That is, these users' bookmarks can be considered as the applied shearing information for Web page recommendation systems with collaborative filtering techniques. In the past researches, only term frequency is extracted from users' bookmark, and used as the users' bookmark features. On the other hand, we assert that the bookmark structures are also reflected in the users' favorites. In this paper, we propose the novel features that are extracted from the bookmark hierarchies for Web page recommendation systems with collaborative filtering techniques. We use three kinds of features, such as the feature of the term frequency, the feature of similarities between the users' bookmark folders, and the feeature of the similarities of structure obtained from users' bookmark hierarchies. From our experimantal results, we found out that our method can recommend more relevant Web pages to the user's favorites than other recommendation system .
Key words WWW，Collaborative Filtering，Recommendation，Bookmark

1. はじめに

インターネット上では，Web ページを用いて誰もが情報を公開し発信できる．そのため，日々Web ページの数は増加の一

途をたどり，2004 年 2 月現在，Google(注1)に登録された Web ページ数は 42 億ページ以上存在する．そこで，膨大な量の Web ページから，利用者が必要な Web ページを効率よく見つけるための手段として，様々な Web 検索エンジンが公開され，広く一般に利用されている．しかし，この Web 検索エンジンには，検索結果が利用者の嗜好に依存しないという問題点がある．

例えば，Java の初心者がキーワード“ Java ”を Web 検索エンジンに入力し，Java に関する Web ページを検索したとする．この場合の検索結果には Java の初心者向けの Web ページも含まれるが，Java の上級者向けの Web ページも多く含まれる．ここで，初心者にとって上級者向けの Web ページは難解であり，不要な Web ページといえる．なぜなら，このような場合，利用者は Web 検索エンジンに，利用者の嗜好を判断し，Java 初心者であれば Java 初心者向けの Web ページ検索して欲しいからである．

このような問題に対し，利用者の嗜好を検索結果に反映させる方法として協調フィルタリングが利用され始めている．協調フィルタリングとは，利用者間で嗜好を共有し，嗜好の似た利用者を特定し，その利用者がよいと評価した Web ページを推薦する手法である．協調フィルタリングでは，類似した嗜好の利用者の特定のために，共有情報が重要になる．近年，ブックマークは利用者の嗜好が顕著に反映された情報であるという基本的概念の基づいて，ブックマークが共有情報として用いられた研究が多く存在する．

そこで本稿では，利用者独自に作成しているブックマークを利用し，ブックマークの階層構造を共有情報として組み込んだ協調フィルタリングに基づく Web ページの推薦手法を提案する．類似した嗜好を持つ利用者を特定するために，本研究では，ブックマークの階層構造から得られるフォルダの類似度，構造の類似度を定義した．二つの類似度は，ブックマーク全体での利用者の嗜好ではなく，ブックマークの一部を利用し利用者の部分的な嗜好の類似性を判定するものであるという特徴を持つ．

本稿では，ブックマーク中の Web ページをブックマークページ，利用者がブックマークに加えた Web ページを追加ブックマークページと定義する．また，提案システムを利用し，推薦を受ける利用者を推薦情報享受者，システムにブックマーク情報を提供し，推薦を提供する利用者をブックマーク提供者と定義する．

2. ブックマークとその関連研究

2.1 ブックマークの優位性

協調フィルタリングでは嗜好の似た利用者を特定するために，共有情報が重要となるが，本稿ではブックマークに着目する．ブックマークを用いた理由は三つ存在する．

一つ目はブックマークは利用者の嗜好が顕著に反映されているからである．ブックマークは数ある Web ページの中から利用者が興味を持った Web ページだけで構成されている．さらに，ブックマークページ数が多くなると，管理が困難になるため，利用者はブックマークを自分がアクセスしやすいように構造化する．このように，利用者独自の嗜好で収集され，階層構造に構造化されたブックマークは，利用者の嗜好を顕著に反映されたものであると言える．

二つ目はブックマークは多くの利用者に利用されているからである．1996 年の調査 [1] では，Web 利用者の 92% 以上がブックマークを利用し，37% 以上の利用者がブックマークページを 50 以上作成していると報告されている．この調査から，多くの利用者がブックマークを利用していることがわかる．

三つ目はブックマークの作成を補助する研究が行われ，ブックマークから利用者の嗜好を抽出しやすくなっているからである．利用者がブックマークを作成する際，効率的に記憶した Web ページにアクセスするために，ブックマークにカテゴリを作り，階層構造に分類・管理することができる．これは利用者にとって労力を要する作業であるが，WebTagger [2] では，それぞれのブックマークに複数のカテゴリ名などの属性情報をメタデータとして付加した上でブックマークのデータベース化・共有を行い，ブックマークの統合的な管理を実現することができ，利用者の労力を軽減できる．このような研究により，利用者の興味のある Web ページがブックマークに自動的に登録され，階層的に管理されることが期待できる．また，利用者は，ある Web ページが別の Web ページからリンクされていることや，あるキーワードを検索エンジンに与えれば，所望の Web ページが得られることを記憶しているため，利用者が興味を持った Web ページのうち，半数以下の Web ページしかブックマークしないと言われている [3]．これに対して，PowerBookmark [4] は利用者の Web アクセス行動の解析により，自動ブックマーク，ブックマーク更新などのさまざまな個別化されたサービスを提供する．PowerBookmark では，追加ブックマークページの内容を解析し，自動的に分類，構造化を行う機能も持つ．さらに，中島ら [5] はブックマークに Web ページを加えるというプロセス (コンテクスト) に着目し，利用者が持つブックマークの価値や追加ブックマークページとブックマークページのと違いについて定義している．これらの情報を用い，様々なアプリケーションに応用することにより，ブックマークを高度利用促す環境を提供している．以上の研究により，利用者にとってブックマークがより利便性の高いものになり，より多くの利用者がブックマークを利用することが期待できる．

2.2 関 連 研 究

本節では，2.1 節で説明したブックマークは利用者が嗜好を顕著に反映しているという基本概念に基づき，ブックマークを嗜好として利用者間で共有し，ブックマーク提供者のブックマークページを推薦する手法について述べる．

2.2.1 ブックマークを利用した協調フィルタリングに基づく Web ページ推薦手法

森ら [6] は現在推薦情報享受者が閲覧している Web ページに対して，ブックマーク提供者のブックマークページを推薦するシステムであるブックマークエージェントを開発した．このシステムは，現在閲覧中の Web ページから推薦情報を生成するため，推薦情報享受者の過去の嗜好に依らない推薦ができると

(注1)：http://www.google.com/

いう特徴を持つ．

また，Ruckerand らは，ある推薦情報享受者のブックマークフォルダ内におけるブックマークページが，ブックマーク提供者のブックマークフォルダ内におけるブックマークページといくつ一致しているかによってフォルダの類似度を定義し，類似度の最も高いフォルダ内のブックマークページを推薦するシステムである Siteseer [3] を開発した．しかしこのシステムでは，ブックマークのフォルダに含まれる Web ページが多いほど推薦フォルダになり得え，有益なブックマークページが存在しても，ブックマークフォルダに含まれる Web ページが少ないフォルダに入っていると推薦されにくいという問題点が存在する．

さらに，濱崎ら [7] はブックマークフォルダに含まれるブックマークページから利用者間のつながりを発見し，関連付けられたブックマーク提供者のフォルダ間に存在するブックマークページを推薦するシステムを提案した．このシステムは推薦情報がどのように推薦情報享受者に推薦されるか理解することができるが，推薦情報がどの程度有用であるかの判断はなされていないため，推薦情報が受け入れられにくいという問題点を持つ．

以上のシステムでは，ブックマークの階層構造を利用せず，ブックマークを深さ 1 の幅が広い木として扱っているものや，フォルダ構造を 1 階層しか利用していないため，ブックマークを十分に活用しているとは言えず，推薦精度が問題となる．

2.2.2 ブックマークの階層構造を利用した協調フィルタリングに基づく Web ページ推薦手法

2.2.1 節での問題点に対し，ブックマークの構造を Web ページ推薦に利用する研究もいくつか存在する．Jung ら [8] はベイジアンネットワーク [9] を利用して，ブックマークの階層構造からユーザの興味を表現する手法を提案した．まず推薦情報享受者は，Yahoo などの組織的に分類されたカテゴリに，興味のある Web ページを加える．そこから，ベイジアンネットワークを利用し，利用者の興味をカテゴリごとに数値化し，利用者が興味を持つカテゴリを割り出す．そして最後に，利用者同士のブックマークの構造を評価し，嗜好の似たブックマーク提供者を発見し，ブックマーク提供者がブックマークした Web ページを推薦する手法である．しかし，定められたカテゴリ構造以外ではブックマーク構造を表現できないため，多くの利用者が独自の基準で構造化しているブックマークをそのまま利用することはできない．

3. 提 案 手 法

2 章で述べたように，既存の研究には，1) ブックマークから得られる知識として，Web ページの単語の出現頻度のみを用いて，ブックマークの階層構造利用していないためブックマークから得られる嗜好を十分に表現できていない，2) 定められたカテゴリ構造でのみでしかブックマーク構造を表現できないため，多くの利用者が独自の基準で構造化しているブックマークをそのまま利用することはできない，という二つの問題が存在する．そこで本章ではこれらの問題点を解決するためのブックマークの階層構造を利用した協調フィルタリングに基づく Web ページ推薦手法を提案する．

3.1 システム概要

提案システムは，推薦情報享受者が Web ブラウズ中に興味を持った Web ページをブックマークに加えたとき，追加ブックマークページと推薦情報享受者のブックマークの階層構造の情報を用いて，ランキング形式で Web ページを推薦するシステムである．具体的に説明すると，推薦情報享受者の追加ブックマークページ α に対して，推薦情報享受者自身のブックマークの階層構造情報を加味し，嗜好の類似したブックマーク提供者を特定し，ブックマーク提供者がよいと判断したブックマークページを推薦するシステムである．図 1 は提案システムが，どのような手順で推薦情報享受者に推薦を生成するかを示している．

（1） 推薦情報享受者の追加ブックマークページに対して，推薦情報享受者のブックマークの構造を加味し，Web ページの類似度 R，推薦情報享受者のブックマークフォルダとブックマーク提供者のブックマークフォルダのフォルダの類似度 R'，推薦情報享受者のブックマークの構造とブックマーク提供者のブックマーク構造の構造の類似度 R'' を算出する (3.2 節，3.3 節で説明)．

（2） (1) で計算された類似度を正規化し，評価統合を行う．そして，推薦情報享受者に推薦結果を提示する (3.4 節で説明)．

一方，図 2 は類似度 R，R'，R'' の算出方法の概要について示した図である．

（1） 推薦情報享受者の追加ブックマークページ α とブックマーク提供者のブックマークページ β 間の Web ページの類似度 R を算出 (3.2 節で説明)．

（2） 追加ブックマークページ α を含む全祖先フォルダとブックマークページ β を含む全祖先フォルダ間のフォルダの類似度の内，最も類似度が高いものを示す，フォルダの類似度 R' を算出 (3.3.1 節で説明)．

（3） 利用者間でのブックマークの構造化の仕方の類似度を示す，構造の類似度 R'' を算出 (3.3.2 節で説明)．

提案手法では，1) ブックマークの階層構造から利用者の嗜好を抽出する，2) 利用者が独自に作成しているブックマークからの推薦ができる，点を特徴とする．

3.2 Web ページの類似度の算出

本節では，推薦情報享受者の追加ブックマークページ α とブックマーク提供者のブックマークページ β 間の Web ページの類似度 R の算出方法について述べる．まず，Web 上からブックマーク中の URL の情報をに基づいて，Web ページの HTML ファイルを入手する．次に，本文のみを抜き出すため，得られた HTML からのタグ，JavaScript，コメントを取り除く．そして，得られた本文から茶筌 [10] による形態素解析で名詞だけを抜き出し，単語を切り出す．

次に，本手法では，Web ページの特徴ベクトル作成手法として Managing Gigabytes [11] での tf-idf 法 [12] を近似する手法を用いた．また，提案手法では，検索キーを Web ページとしているため，Web ページを検索キーワードの集合とみなし，以下のような方法で追加ブックマークページ α とブックマー

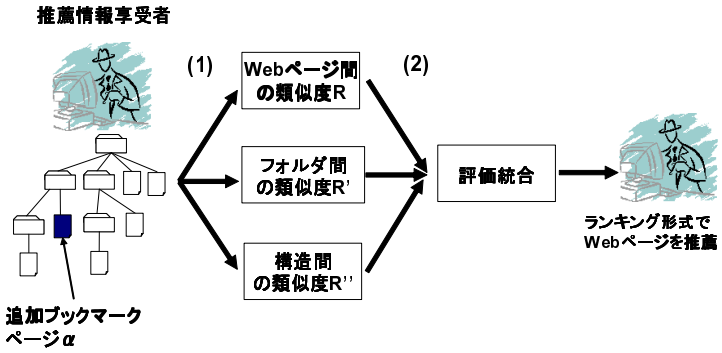

図 1　推薦の流れ

Fig. 1　recommend flow

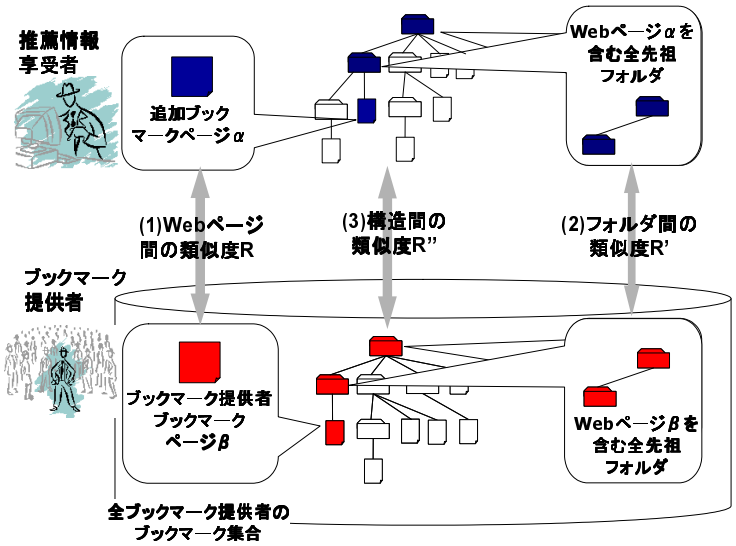

図 2　システム概要

Fig. 2　outline of our proposed system

ク提供者のブックマークページ β の間の類似度を定義した．N を全ブックマーク提供者のブックマークページ数，P を単語 t が含まれる Web ページ数，f_t^α を追加ブックマークページ α の単語 t の出現数，f_t^β をブックマークページ β の単語 t の出現数とした時，ブックマークページ β の単語の重み w_t^β，および，追加ブックマークページ α の単語の重み w_t^α は以下のように表される．

$$w_t^\beta = 1 + \log_e f_t^\beta \tag{1}$$

$$w_t^\alpha = (1 + \log_e f_t^\alpha)\log_e(1 + \frac{N}{P}) \tag{2}$$

また，c を追加ブックマークページ α の単語 t の数，追加ブックマークページの α の単語の重みを $w_{t_1}^\alpha, \cdots, w_{t_c}^\alpha$ とした時，追加ブックマークページ α の特徴ベクトル $\boldsymbol{V}^\alpha$ は以下のように表される．

$$\boldsymbol{V}^\alpha = (w_{t_1}^\alpha, \cdots, w_{t_c}^\alpha) \tag{3}$$

最終的に，ブックマークページ β の特徴ベクトル $\boldsymbol{V}^\beta$ とすれば，追加ブックマークページ α とブックマークページ β 間の類似度 R は次式のように定義される．

$$R = sim(\boldsymbol{V}^\alpha, \boldsymbol{V}^\beta) = \frac{\boldsymbol{V}^\alpha \cdot \boldsymbol{V}^\beta}{|\boldsymbol{V}^\alpha||\boldsymbol{V}^\beta|} \tag{4}$$

3.3　利用者の嗜好を表現する特徴量の算出とブックマーク間の類似度

本節では，類似した嗜好を持つブックマーク提供者を特定するため，ブックマークの階層構造を利用し，追加ブックマークページ α を含む全祖先フォルダの特徴量とブックマークページ β を含む全祖先フォルダの特徴量から得られるフォルダの類似度の内，最も類似度が高いものを示すフォルダの類似度 R'，利用者間でのブックマークの分類の仕方の特徴量から得られる類似度を示す構造の類似度 R''，を定義する．二つの類似度は，ブックマーク全体での利用者の嗜好ではなく，ブックマークの一部を利用し利用者の部分的な嗜好の類似性を判定するものである．

3.3.1　フォルダの類似度による評価

推薦情報享受者の追加ブックマークページ α とブックマーク提供者のブックマークページ β がどのようなフォルダをたどって分類されているかを比較，すなわち利用者間の嗜好の類似性判定に利用する．推薦情報享受者が追加ブックマークページ α の含んだブックマークのルートフォルダから追加ブックマーク

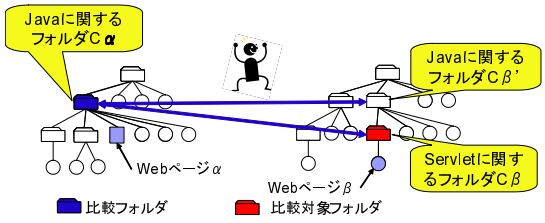

図 3　フォルダの特徴の比較

Fig. 3　comparing particularity of folder

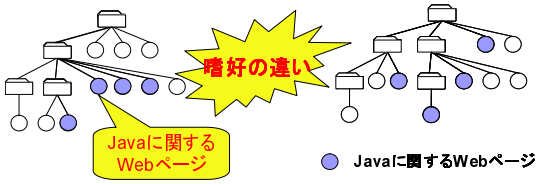

図 4　利用者間での嗜好の違い

Fig. 4　different structure between users

ページ α までのパスに含まれる中間フォルダ C_α と，ブックマークページ β を含んだブックマークのルートフォルダからブックマークページ β までのパスに含まれる中間フォルダ C_β の特徴量を比較する．ここで，C_α と C_β の組合せの数だけ，中間フォルダの特徴ベクトル (次項目で説明) を利用して，フォルダ間の類似度を算出し，その最大値をフォルダの類似度とする．

図 3 を用いて具体的に説明する．追加ブックマークページ α を含むフォルダを C_α, ブックマークページ β を含むフォルダを C_β，ブックマークページ β のひとつ上の階層のフォルダを $C_{\beta'}$ とする．フォルダ C_α とフォルダ C_β の類似度は，ブックマークページ β を含むフォルダ C_β のひとつ上の階層のフォルダ $C_{\beta'}$ とフォルダ C_α の類似度より低い値だと考えられる．そこで，ブックマークのルートフォルダから追加ブックマークページ α，ブックマークページ β が含まれているフォルダまでのすべての中間フォルダ対して，類似度の計算を行っている．ブックマークページ β を含むフォルダ C_β のひとつ上の階層のフォルダ $C_{\beta'}$ の情報を利用することにより，利用者間で異なるブックマークの構造を吸収し，ブックマークから得られる利用者の嗜好を表現できる例を示している．

- フォルダの類似度の算出

フォルダの特徴ベクトルの計算には，tf-idf 法の応用として提案された，tf-icf 法 [13] を利用する．tf-icf 法で単語の重みを計算することで，ブックマークフォルダを反映した単語の重み付けを行い，ブックマークフォルダの特徴ベクトルを生成できる．ブックマーク提供者のブックマーク集合に含まれる総フォルダ数 n，フォルダ c に含まれる単語 t^c と単語 t^c を含むブックマーク内のフォルダ数を f_{t^c}，フォルダ c に含まれる単語 t^c の出現頻度を $f^c_{t^c}$ とすると，フォルダ c における単語 t^c の重み $w^c_{t^c}$ は以下のように定義される．

$$w^c_{t^c} = (1 + \log_e f^c_{t^c}) \log_e (1 + \frac{n}{f_{t^c}}) \tag{5}$$

また，フォルダに含まれる単語の数 M とした時，フォルダの特徴ベクトル $\boldsymbol{V}^c$ は以下のように定義される．

$$\boldsymbol{V}^c = (w^c_{t^c_1}, w^c_{t^c_2}, \cdots, w^c_{t^c_M}) \tag{6}$$

最終的に，比較対象フォルダ g の特徴ベクトル $\boldsymbol{V}^g$ とすれば，フォルダ c と g 間の類似度 R' は次式のように定義される．

$$R' = sim(\boldsymbol{V}^c, \boldsymbol{V}^g) = \frac{\boldsymbol{V}^c \cdot \boldsymbol{V}^g}{|\boldsymbol{V}^c||\boldsymbol{V}^g|} \tag{7}$$

3.3.2　構造の類似度による評価

ブックマークの管理は利用者に任されているため，利用者は独自の構造でブックマークを管理する．図 4 の例では，Java に関する Web ページでも，利用者によって構造化の方法に違いがあるため，異なるブックマークの階層構造をとることがわかる．この階層構造の違いを比較，すなわち，利用者間の嗜好の類似性判定に利用する．本研究では，ブックマークの分類法の類似した利用者からの推薦を受けることが重要であるため，構造の類似度 R'' を算出する．提案手法では，利用者のブックマーク全体の構造ではなく，追加ブックマークページ α に類似した Web ページに関して，類似したブックマーク構造を持つ利用者から推薦を受けることを目的とする．

図 5 は利用者間で異なった嗜好で作成された階層構造を持つブックマークの類似性判定方法について示したものである．図 5 中の推薦情報享受者によって Web ページが加えられたブックマーク 1) とブックマーク提供者のブックマーク 2) から追加ブックマークページ α とある閾値以上の類似度を持った Web ページ r のみを残し，閾値以下の類似度の Web ページを削除するとそれぞれ 3)，4) となる．さらに，推薦情報享受者の追加ブックマークページ α と，それに類似している，あるいは閾値以上の類似度を持った Web ページ r がブックマーク内でどのように構造化されているかを基にブックマーク間の類似性判定を行う．得られたブックマークの構造間の類似度の判定には edit-distance [14] を用いる．しかし，3) と 4) の構造を比較すると利用者間でブックマークの構造が違いすぎるため，利用者の嗜好をうまく反映できない．そこで，ブックマーク同士を統合し，中間的なブックマーク構造 5) を作成し，中間的なブックマーク構造と 4) を比較し，最終的に 4) と 5) のブックマークの構造を比較することで，構造の類似度 R'' を算出した．

edit-distance とは，任意の文字列 p に対して挿入，削除，置換のいずれかを適用して, 任意の文字列 q を求めるための最小コストの算出手法である. ブックマークを深さ優先探索で走査し，ラベル付けを行い，正規化したものを文字列で表す．その文字列間の類似度を指標として edit-distance で算出された値を用いる．値が大きいほど，類似度が低いことを表す．

ブックマークの統合には，データ統合の研究分野の技術であるメディエータと呼ばれるプログラムを利用した．メディエータとは，問合せの再構築や配布を行ったり，結果データをフィルタリングし，統合し，処理したりすることにより，新しい，より密度の高い，より関連性のある情報を生成するように，問合せや結果に対して付加価値の処理を行う手法である [15]．メディエータにはさまざまな手法 [16] [17] があるが，本稿では XML のデータを統合できる XML Data Mediator [18] を利用

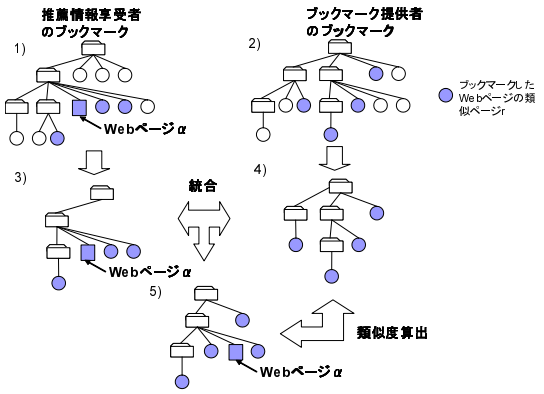

図 5　ブックマークの統合
Fig. 5　integration of tree

する．

そして，類似度 R'' は edit-distance により算出されたコストが γ だったとき，以下のように定義される．

$$R'' = \frac{1}{\gamma + 1} \tag{8}$$

3.4 評価統合

提案する推薦手法では，最終的に推薦されるべき Web ページをランキングするために，それぞれの算出された類似度同士を統合する必要がある．情報検索の分野では二つの値を統合するための関数として様々な関数が用いられており，例えば最大値や最小値，相加平均などが使用されている．ところが，これらの関数の選択によって推薦手法の有効性に差が出る [19] ことがわかっている．Montague らの報告では，テキスト文書検索において，SUM を使った正規化とそれぞれの類似度の和を表す CombSUM を使った評価統合を用いた場合，高い検索精度になることが実験で示されている [20]．そこで，我々はこの二つの関数を使用し正規化と評価統合を行う．

ここで正規化前の類似度 $S_i(O_j)$，ブックマーク中の Web ページ数 N で SUM で正規化を行った後の類似度 $S'_i(O_j)$ を定義すると以下の式で表される．

$$S'_i(O_j) = \frac{S_i(O_j)}{\sum_{j=1}^{N} S_i(O_j)} \tag{9}$$

さら，CombSUM で評価統合を行う前の類似度 $S'_i(O_j)$ 行った後の類似度 $S''(O_j)$ を定義すると以下の式で表される．

$$S''(O_j) = \sum_{i=1}^{N} S'_i(O_j) \tag{10}$$

4. 実　　験

4.1 実験環境

情報検索の分野では検索精度を評価するため，TREC テストコレクション [21] が存在する．しかし本手法の有効性を示す実験では，利用者が独自に作成しているブックマークが必要となるが，我々の想定に合うようなテストコレクションを発見できなかったため，独自にテストコレクションを作成した．

そこで，我々は Web 上ではブックマークを公開している利用者が多いため，それらのファイルを人手で収集して利用した．公開されたブックマークはフォルダ名，URL の情報を持っており，またブックマークの階層構造も表現している．収集したブックマークの詳細を表 1 に示す．

表 1　各利用者のブックマークの詳細
Table 1　collected bookmark

推薦情報享受者数	33 人
総フォルダ数	1724 個
デッドリンクではない Web ページ数	9201 ページ
全ブックマーク提供者のフォルダ数の平均	52.24
全ブックマーク提供者のフォルダ数の標準偏差	62.00
全ブックマーク提供者のブックマークの深さの平均	3.39
全ブックマーク提供者のブックマークの深さの標準偏差	0.555
全ブックマーク提供者のブックマーク中ページの平均	252.24
全ブックマーク提供者のブックマークページの標準偏差	352.13

以上のブックマークデータを用い，以下のような手順で評価実験を行った．

（1） 全ブックマーク提供者のブックマークページの単語の重みから特徴ベクトルを作成し，その情報をファイルに格納する．

（2） Web ページの閲覧中に推薦情報享受者のブックマークに Web ページを加えたとき，加えたページの単語の重みから特徴ベクトルを作成し，(1) で作成された特徴ベクトルとのコサイン相関値から類似度を算出する．

（3） 推薦情報享受者と他の全ブックマーク提供者とのフォルダの階層構造を比較し，評価されたフォルダの類似度と構造の類似度によって，嗜好の似たブックマーク提供者を発見する．

（4） (2)，(3) で得られた類似度を正規化し，統合関数を用いて類似度の統合を行う．

4.2 評価方法

実験の評価は以下のように行った．

（1） 被験者が Web ブラウズ中に Web ページをブックマークに加えることを想定し，著者の友人である 3 人の被験者によって，1 組組づつ合計 3 組の追加ブックマークページについて，適合 Web ページの組を作成する．

（2） Web ページを提案手法に入力し，ランキング形式で推薦された Web ページを得る．

（3） 3 組の追加ブックマークページに対して，ランキング形式で推薦された Web ページ群と (1) で作成された適合 Web ページを比較し，再現率と適合率を求める．その結果から平均適合率を算出し，再現率－適合率グラフに描画し，評価を行う．

4.3 平均再現率適合率

情報検索システムの検索精度を評価する指標として，再現率，適合率が挙げられる [22]．追加ブックマークページに対しての適合 Web ページ集合 R と，推薦結果 A が，図 6 に示すような結果になったとすると，再現率 R_e と適合率 P_r は次のような式で定義される．

$$R_e = \frac{Ra}{R} \tag{11}$$

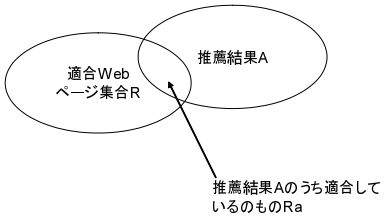

図 6　適合 Web ページ集合と推薦結果
Fig. 6　Recall and Precesion

$$P_r = \frac{Ra}{A} \tag{12}$$

得られた再現率と適合率から推薦 Web ページ数を変化させて，グラフの横軸に再現率を縦軸に適合率をプロットしグラフを描くと，再現率－適合率と呼ばれるグラフが得られる．また，再現率適合率グラフで性能判定が難しい場合，平均適合率と呼ばれる評価指標で表現することがしばしば行われる [23]．B を他の全ブックマーク提供者のブックマーク中の Web ページ数，$O_i(i = 1, 2, \cdots, N)$ を推薦順位第 i 位の文書の適合，不適合を表す変数とした時，平均適合率 v は以下のように定義される．

$$v = \frac{1}{\sum_{i=0}^{B} x(O_i)} \sum_{j=i}^{N} \frac{j}{x(O_i)} (1 + \sum_{k=1}^{j-1} x(O_k)) \tag{13}$$

4.4　実 験 結 果

ここでは，3 章で提案した，ブックマークの階層構造を利用し協調フィルタリングに基づく Web ページ推薦を行う手法の有効性を判断するために実験を行った．

実験の項目としては，次の四つを用いた．

（1）　Web ページの間の類似度のみ
（2）　(1) の類似度とフォルダの類似度を評価統合したもの
（3）　(1) の類似度と構造の類似度を評価統合したもの
（4）　(1) の類似度，フォルダの類似度と構造の類似度を評価統合したもの

各項目の再現率－適合率グラフを図 7 に，平均適合率を表 2 に示す．この結果より，提案したフォルダの類似度と構造の類似度を考慮した推薦手法より，構造の類似度を考慮した推薦手法がよりよい精度で推薦を行うことがわかる．この理由としては，フォルダの類似度が嗜好の類似した利用者をうまく特定できず，推薦精度の低下することが挙げられる．以下にフォルダの類似度の問題点を述べる．

- フォルダの類似度の問題点

フォルダの類似度が高く評価されたフォルダの上位に，フォルダに含まれる Web ページ数が少ないフォルダが出現し，推薦精度の低下を招いている．この原因としては Web ページが増えるほど，Web ページ群に含まれる単語が多くなり，単語一つ一つの重みが小さくなる．その結果，情報量が少なく，単語の一つ一つの重みが大きいフォルダに含まれる Web ページ数の少ないフォルダが上位に推薦される．これではフォルダの特徴を十分抽出しているとは言えない．したがって，フォルダの特徴づけ手法を検討すべきと考える．

しかし，単語の出現頻度に基づいた Web ページの類似度の

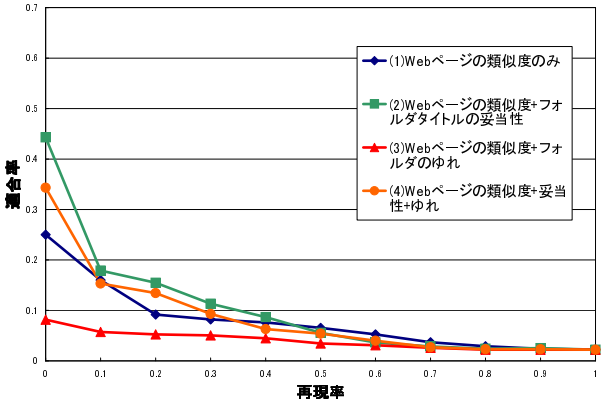

図 7　実験結果の再現率－適合率グラフ
Fig. 7　recall-precision graph

表 2　実験結果の平均適合率
Table 2　average precision

Web ページの類似度のみ	0.080900714
フォルダの類似度	0.068731703
構造の類似度	0.109932284
フォルダ＋構造の類似度	0.093167443

みを用いる場合に比べて，ブックマークの階層構造を利用した，フォルダの類似度と構造の類似度を考慮した推薦手法の方が，利用者の嗜好の類似したブックマーク提供者からの推薦を受けることが可能であることがわかった．

また，Web ページの類似度と構造の類似度にも改善の余地があることもわかった．以下にその問題点を述べる．

- Web ページの類似度の問題点

推薦結果の上位に，適合していない情報量の少ない Web ページが出力される問題が確認された．情報量が少ないと単語一つ一つの重みが大きくなり，推薦結果の上位に出現することが多くなる．ここで，情報量が少ない Web ページとは，他の Web ページにサイトが移転し，移転することだけが書かれた Web ページや，閉鎖されたことだけが書かれた Web ページ，さらに Flash などのアプリケーションで作成され，テキスト情報が少ない Web ページなどがあげられる．したがって，情報量の少ない Web ページに対して，基準を設けて，これらの Web ページを推薦結果から取り除く機構が必要である．

- 構造の類似度の問題点

構造の類似度があまり高くないブックマーク提供者のブックマークに多くの適合 Web ページが含まれ，推薦の精度の低下を招いている．これは追加ブックマークページ α に類似した Web ページ数が推薦情報享受者とくらべて多く，edit-distance にコストがかかり，類似度が低くなっている．このように必ずしも，適合 Web ページ数を多く持つブックマーク提供者の類似度が高くならないことがわかった．このように，現在の手法では，類似 Web ページ数に依存することが大きいため，より構造の分類のされ方の嗜好を判定するような提案が必要となる．

5. ま と め

本稿ではブックマークの階層構造を利用した協調フィルタリングに基づく Web ページ推薦手法について提案を行った．評価実験を行った結果，構造の類似度を加味した，場合よい結果が得られた．本手法の利点は以下の通りである．

- 利用者が独自に作成しているブックマークからの推薦ができる．
- 推薦情報享受者が現在興味を持っている Web ページの情報を基に，これまで利用者がブックマークしてきた情報を考慮して推薦を行うため，推薦情報享受者の嗜好の変化に対応できる．

また，本論文で行った実験で以下のような問題点があることがわかった．

- 情報量の少ない Web ページを推薦結果から排除する機構
- 新たなフォルダの特徴づけ手法の提案
- より構造のされ方の違いを評価できる構造の類似度の提案

さらに，今後の以下のような検討を行う必要があると考える．

- 本提案手法の有効性を判断するには Jung ら [8] の手法など，関連研究との比較が必須であり，早急に比較実験を行う必要がある．
- ブックマークはプライバシーにかかわる情報であるため，公開を拒む利用者も多く存在する．そこで，利用者のプライバシーを考慮したブックマークの利用方法について考える．

謝　　辞

本研究の一部は，日本学術振興会および文部科学省科学研究費補助金 (課題番号はそれぞれ 14780325，15017243) の支援によるものである．ここに記して謝意を示す．

文　　献

[1] Kehoe, C. and Pitkow, J.: Surveying the Territory: GVU's Five WWW User Surveys, *The World Wide Web Journal*, Vol. 1 (1996).

[2] Keller, R. M., Wolfe, S., Chen, J. R., Rabinowitz, J. L. and Mathe, N.: A Bookmarking Service for Organizing and Sharing URLs, *Proc. of the 6th International World-Wide Web Conference*, Vol. 29(8-13) (1997).

[3] Ruckerand, J. and Polanco, M.: Siteseer: personalized navigation for the Web, *Communications of the ACM*, Vol. 40, pp. 73–74 (1997).

[4] Li, W.-S., Vu, Q., Chang, E., Agrawal, D., Hara, Y. and Takano, H.: PowerBookmarks: A System for Personalizable Web Information Organization, *Proc. of the 8th International World-Wide Web Conference*, Vol. Computer Networks 31(11-16) (1999).

[5] 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.SIG5-4 (TOD14) (2002).

[6] 森幹彦, 山田誠二: ブックマークエージェント: ブックマークの共有による情報検索の支援, 電子情報通信学会論文誌, J-83-D-I, pp. 487–494 (2000).

[7] 濱崎雅弘, 武田英明, 松塚健, 谷口雄一郎, 河野恭之, 木戸出正継: Bookmark からの共通話題ネットワークの発見手法の提案とその評価, 人工知能学会論文誌, 17 巻 3 号 SP-D, pp. 276–284 (2002).

[8] Jung, J., Yoon, J. and Jo, G.: Collaborative Information Filtering by Using Categorized Bookmarks on the Web, *Proc.of the 14th International Conference on Applications of Prolog, 20-22* (2001).

[9] Jensen, F.: *Bayesian Networks and Decision Graphs*, Springer-Verlag (2001).

[10] 松本裕治, 北内啓, 山下達雄, 平野義隆, 松田寛, 高岡一馬, 浅原正幸: 形態素解析システム『茶筌』version2.2.6 使用説明書 (2001).

[11] Witten, L., Moffat, A. and Bell, T.: *Managing Gigabytes*, Morgan kaufmann publishers (1999).

[12] Salton, G. and McGill, M. J.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).

[13] Cho, K. and Kim, J.: Automatic Text Categorization on Hierarchical Category Structure by using ICF(Inverted Category Frequency)Weighting, *Proc. of the KISS Conference*, pp. 507–510 (1997).

[14] *http://specialist.nlm.nih.gov/GSpell.html*.

[15] 西尾章治郎, 大田友一, 横田一正, 西田豊明, 佐藤哲司: 情報の共有と統合, 岩波書店 (1999).

[16] Cluet, S., Delobel, C., Siméon, J. and Smaga, K.: Your mediators need data conversion!, *SIGMOD Conference*, pp. 177–188 (1998).

[17] Buneman, P., Davidson, S. B., Hart, K., Overton, C. and Wong, L.: A Data Transformation System for Biological Data Sources, *Proceedings of the Twenty-first International Conference on Very Large Databases*, Zurich, Switzerland, VLDB Endowment, Saratoga, Calif. (1995).

[18] *XML Data Mediator:http://www.alphaworks.ibm.com/tech/*.

[19] 鈴木優, 波多野賢治, 吉川正俊, 植村俊亮: 複数のメディアで構成された電子文書の検索手法, 情報処理学会論文誌：データベース, 第 42 巻, 第 SIG10(TOD11) 号 (2001).

[20] Montague, M. and Aslam, J.: Relevance score normalization for metasearch, *Proc. of the ACM Tenth International Conference on Information and Knowledge Management (CIKM)* (2001).

[21] *National Institute of Standards and Technology: Text REtrieval Conference (TREC), http://trec.nist.gov/*.

[22] R.Baeza-Yatas and B.Ribeiro-Neto: *Modern Information Retrieval*, ACM Press (1999).

[23] 岸田和明: 検索実験における評価指標としての平均精度の性質, 情報処理学会論文誌：データベース, 第 43 巻, 第 SIG2(TOD13) 号 (2002).