3B1-05

ブックマークを用いた人の繋がりの発見手法

The discovery method for human relationship by using WWW bookmarks

濱崎 雅弘[*1] 武田 英明[*1 *2] 河野 恭之[*1] 木戸出 正継[*1]
Masahiro HAMASAKI Hideaki TAKEDA Yasuyuki KONO Masatsugu KIDODE

[*1]奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

[*2]国立情報学研究所
National Institute of Informatics

In this paper, we show how human network can be found and used on WWW. We proposed a system called *kMedia* that can assist users to form knowledge for community by showing shared topics networks(STN) among them. kMedia uses bookmarks as users' knowledge because the structure of bookmarks and web pages registered in bookmarks reflect user's interest. We conducted an experiment to know how kMedia can support users. One result is that folder recommendation is more effective than page recommendation. The other is that recommendation is more effective for people belonging to the same real communities than those to different communities. According to these result, we propose a new measurement called "category resemblance" that is recommendation measurement based on resemblance of folder structures.

1. はじめに

近年の WWW に代表される情報伝達技術の進歩により，我々は膨大な量の情報を利用する事が可能になった．しかし，同時に行き交う情報が増えた事によって，自分の欲しい情報を見失ってしまう事がたびたび起こるようになった．

このような状況を改善するために，多くの情報収集支援技術が開発された．その一つに情報推薦がある．これはユーザの関心に基づいて，システムが情報を選別・推薦する技術である．情報推薦にはユーザプロファイル作成が必要であり，これは一般に，ユーザにとって負荷がかかるものである．

そこで我々はブラウザのブックマークが，所有者の WWW に関する関心を構造的に示している知識であると考え，プロファイルとして利用する情報推薦システム kMedia[Takeda 00] を作成した．本システムでは，ブックマークの階層構造を利用する事により，単純に Web ページの推薦を行うよりも効果的に推薦をする事ができる．本稿では，被験者を用いた実験とそこから得られた知見について報告する．

2. kMedia システム

2.1 システム概要

kMedia では，ブックマークからユーザが興味を持っている話題を調べ，次にユーザ同士の話題の共通性を調べる．ユーザがどんな話題に興味を持っているかを知るために，ブックマークのフォルダを利用する．

ブックマークのフォルダとは，WWW ブラウザのブックマークに登録する Web ページを階層構造で整理するためものである．ユーザは任意にフォルダを作り，その中に記録しておきたい Web ページを登録できる．その Web ページが登録されたフォルダを一つの話題とみなし，フォルダに登録している Web ページはそのフォルダ（話題）を示す内容であるとする．これらの関係を図 1 に示す．

ユーザ同士の話題の共通性は，その話題に含まれる情報がどれだけ類似しているかによって決める．kMeda ではブックマークフォルダに含まれる情報，すなわちフォルダ内に登録された Web ページの類似性から話題の共通性を求める．つまり，Web ページの類似性が図 2 における左図のように発見された場合，話題の共通性は右図のようになるとする．

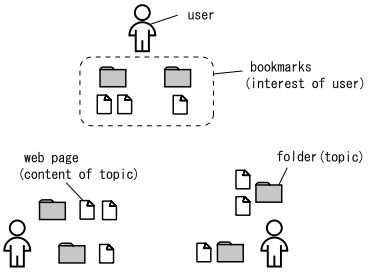

図 1: 話題とブックマークの関係

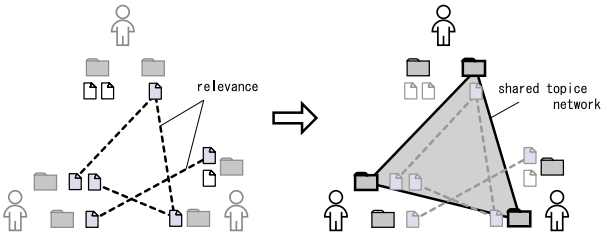

図 2: Web ページの類似例

2.2 ページ関連度

Web ページの類似性はページ中の重要な単語（キーワード）の一致によって決める．まず Web ページ中の出現頻度が多い単語の中から上位幾つかをそのページのキーワードとし，互いのページで一致したキーワード数が多いほど，そのページ間には類似性があるとしている．

kMedia は指定されたブックマークファイルを解析して，ブックマーク先の Web ページを読み込み，そのページの文章から単語を切り出す．単語の切り出しには，茶筌 [松本 99] などによる形態素解析ではなく，字種に注目した単語切り出し [片岡 97] を用いた．

連絡先: 奈良先端科学技術大学院大学 情報科学研究科，〒 630-0101 奈良県生駒市高山町 8916-5, Tel:0743-72-5085, Fax:0743-72-5269, E-Mail:masa-ha@is.aist-nara.ac.jp

字種に注目した単語切り出しとは，解析する文章を先頭から一字ずつ読み取り，連続する漢字，カタカナ，全角英数文字を単語として切り出す方法である．単語と見なす最短文字長は調整可能であるが，今回は 3 文字とした．

字種に注目した単語切り出しは，形態素解析に比べて厳密さには欠けるが，特にカタカナの新語，略語に対して柔軟な切り出しができるため，WWW の情報のように，雑多でやわらかい文章の解析に適している．

切り出された単語のうち，頻度順に，頻度が同じ場合は文字長の長い順に最大 10 単語をそのページのキーワードとしている．ただし,「ホームページ」や「カウント」など，あまりその文書を特徴付けないような頻出単語はキーワードとして除外した．

このようにして選ばれた各 Web ページのキーワードの一致から，Web ページ間の類似を調べる．今回はキーワードが 2 語以上一致し，一致したキーワードが両 Web ページで合計 10 回以上現れた場合に Web ページ間に類似性があったと判断し，Web ページの推薦を行う．なお，このキーワードの出現回数のことを，以後，ページ関連度と呼ぶ．

2.3 フォルダ関連度

話題の共通性は，話題を示すブックマークフォルダ中の Web ページの類似によって決める．フォルダ間で互いに類似性がある Web ページの数をフォルダ関連度と呼ぶ．今回は，フォルダ関連度が 3 以上，つまりフォルダ中の Web ページが 3 つ以上，互いに類似性がある場合に，そのフォルダ（話題）間に共通性があるとし，フォルダの推薦を行う．

なお，話題の共通性を調べる際にはフォルダの階層は 2 階層以上は考慮していない．つまり，2 階層以上深いフォルダ内の Web ページは，全て親フォルダ内の Web ページと見なした．

3. 実験概要

4 つの異なるコミュニティから 3 人ずつを被験者として集めた．ここで言う同じコミュニティに属するということは，同じ研究室に所属している事を指す．

この実験では，kMedia を通常の利用法とは異なる方法を用いた．まず，被験者から集めたブックマークを実験者が kMedia システムに与える．そして，その結果を HTML 形式に加工し，フォームを用いて評価してもらった．

同じコミュニティに所属する人同士と，異なるコミュニティに所属する人同士とで，ユーザの評価がどの程度変わるかを調べるために，互いに異なるコミュニティに属する者だけのグループを，同じ 12 人の被験者を用いて 4 グループ作った．

各被験者は，まず同じコミュニティに所属する人同士で kMedia を使った場合の推薦結果に対する評価を行い，次に異なるコミュニティに所属する人同士での結果に対する評価を行った．被験者には推薦相手の名前は伏せており，推薦された情報が同じコミュニティに所属している人からのものか，そうでないかは判断できない状況で評価を行ってもらった．

評価は，推薦されたページに対しての評価と推薦されたフォルダに対しての評価，そしてページやフォルダを推薦した推薦者に対しての合計３種類５項目，それぞれ５段階（5:大変そう思う，4:そう思う，3:どちらとも言えない，2:そう思わない，1:全く思わない）で行った．評価項目は以下のとおりである．

- あるページに対してのページ推薦が妥当であるか
- 推薦されたフォルダは似ているか
- 推薦されたフォルダは役立つか
- 相手と連絡を取りたいか
- 相手と会いたいか

4. 実験結果

4.1 フォルダ推薦の有効性

本システムの特徴として，Web ページではなくフォルダを推薦するというものがある．この推薦方法は，Web ページを推薦するよりも良い結果が得られた．図 3 は，ページ推薦とフォルダ推薦の評価結果を示したものである（フォルダ推薦については 2 種類の評価をしてもらったが，両方とも同じ傾向であったため,「似ているか」の評価結果を示す）．

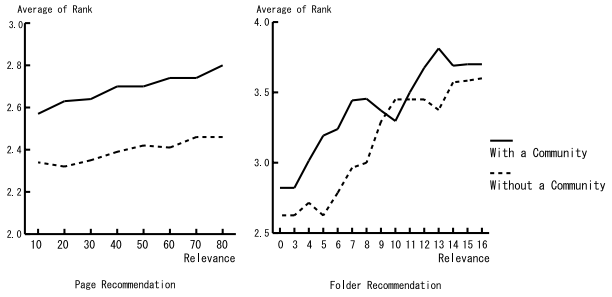

図 3: フォルダ推薦の有効性

それぞれ縦軸が平均評価値，横軸が推薦を行うための関連度の閾値である．ページ推薦の平均評価値が，閾値を高く設定しても高々2.7 程度であるのに対して，フォルダ推薦では閾値を高く設定すれば，3.7 程度にまでなっている．これは，フォルダが概念的に Web ページよりも抽象的であるため，ユーザにとって受け入れ易かったのではないかと考えられる．

4.2 人への評価

推薦結果を評価した後で，その推薦相手に対して「連絡したいか」「会いたいか」という２つの評価を行った．この評価結果から，人の繋がりがどういった要素に左右されるのかを調べた．表 1 は，システムによって求めたそれぞれの値と，被験者による推薦者への評価との相関である．

表 1: 推薦者の評価との相関 (1)

	連絡したい	会いたい
推薦ページ数	0.42	0.30
平均ページ関連度	-0.13	-0.19
推薦フォルダ数	0.45	0.35
平均フォルダ関連度	0.38	0.30

「連絡したい」「会いたい」という２つの評価は，比較すると，前者は情報交換で済む繋がり，後者は単なる情報交換だけではない繋がりと言える．推薦数や関連度などの，お互いが持っている情報が似ているかどうかを示すだけの値では，情報交換で済む繋がりと相関が高いという結果が出た．

表 2 は，推薦されたページやフォルダに対する被験者の評価と，その推薦者に対する被験者の評価との相関を示したものである．平均フォルダ評価 1，2 はそれぞれ「似ているか」「役立つか」という評価である．被験者の評価は，システムが求めた値とは「連絡したい」と「会いたい」とで，相関の大小関係が逆転している．

表 2: 推薦者の評価との相関 (2)

	連絡したい	会いたい
平均ページ評価	0.29	0.40
平均フォルダ評価 1	0.28	0.32
平均フォルダ評価 2	0.09	0.20

システムが求めた値は「会いたい」よりも「連絡したい」という気持ちとの相関の方が高い．対して，人が行った評価は，「連絡したい」よりもむしろ「会いたい」という強い繋がりと相関が高い．これは人の繋がりと単純な情報交換のための繋がりとは，微妙に異なることを示している．

また,「会いたい」という繋がりは，システムが求めた値よりも被験者による評価値の方と相関が高い．この事から，人が情報交換だけではない繋がりを求める相手は，単純に所有する情報の近似では駄目なことがわかる．

5. 人の繋がり

前節の結果より，人の繋がりの指標として，単に共有する情報の量や質以外のものがあるのではないかと考えた．そこで，まず実際に存在する人の繋がりである，実社会における同一コミュニティ内に所属する者同士の特徴を調べる事によって，適切な指標を見つけ出す事が出来るのではないかと考えた．

5.1 ページ推薦数とフォルダ推薦数

図 4 は，同じコミュニティ間でのページ推薦と異なるコミュニティ間でのページ推薦数，およびフォルダ推薦数を比較したものである．

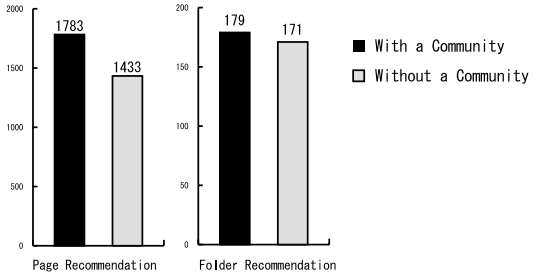

図 4: ページ・フォルダ推薦数

コミュニティの有無で推薦ページ数の差は 300 ほどあるのに対し，推薦フォルダ数の差はわずか 8 であった．つまり，同じコミュニティ間での推薦の方が，ページ推薦に対してフォルダ推薦の数が多いという傾向がある事がわかる．

5.2 ページ関連度とフォルダ関連度

図 5 は，ページ関連度による閾値を変化させた場合，推薦されるページのフォルダ関連度がどのように変化するかを示したものである．推薦ページのフォルダ関連度とは，類似性のあるページ組の，それぞれが登録されているフォルダ間のフォルダ関連度を示す．

ページ関連度による閾値をを高くすると，同じコミュニティに所属する人同士で推薦されるページは，フォルダ関連度の高いものが多くなっている．対して，異なるコミュニティでは，そのような傾向は現れない．

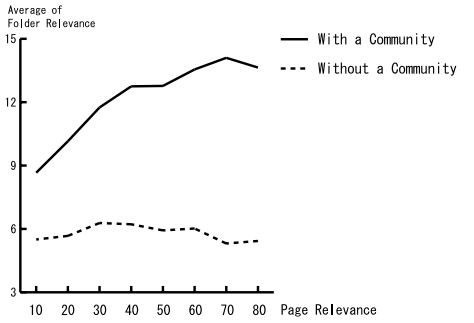

図 5: ページ関連度とフォルダ関連度

5.3 カテゴライズ近似度

推薦数および関連度についての調査から，実社会において同じコミュニティに属している（つまり人の繋がりがある）ことは，そうでない場合と比較して以下のような二つの傾向があるといえる．

- ページ推薦数に対してフォルダ推薦数は少ない
- ページ関連度とフォルダ関連度の相関が高い

この二つの傾向から，人の繋がりの尺度として，Web ページの分類方法の類似性が使えるのではないかと考えられる．図 ??では点線は関連性があるページ組を示しているが，左右両方とも同じ 4 つである．だが，その関連性のあるページの分類の仕方がそれぞれ異なっている．この場合，同じような分類をしている左の方が，右の方と比較して人の繋がりが強いと判断できる．

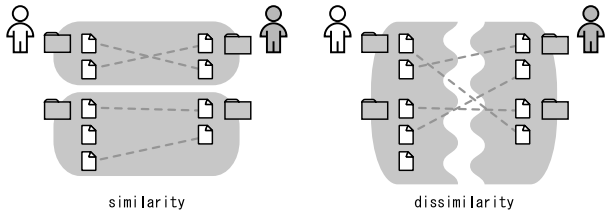

図 6: 分類方法の類似

ここで分類方法が類似性を示すものとして，カテゴライズ近似度という指標を新たに作る．これは，共有している情報の量でユーザ間を評価するのではなく，一致する率で評価するものである．ユーザ i にとってユーザ j のカテゴライズ近似度は，以下の式で求めることができる．

$$C_{ij} = \frac{Nf_{ij} \times Rf_{ij}}{Np_{ij}}$$

C_{ij} : カテゴライズ近似度
Nf_{ij} : 推薦フォルダ数
Rf_{ij} : 平均フォルダ関連度
Np_{ij} : 推薦ページ数

この計算式で求めたカテゴライズ近似度と，被験者による推薦者への評価との相関値を表 3 に示す．カテゴライズ近似度は，システムが求めた他のどの評価値よりも相関が高く，か

つ,「連絡したい」よりも「会いたい」という評価との相関値の方が高い．このことから，カテゴライズ近似度という新しい指標は，単純なシステム評価よりもユーザの声を反映している事を示しているといえる．

表 3: 推薦者の評価との相関 (3)

	連絡したい	会いたい
カテゴライズ近似度	**0.49**	**0.55**

6. 結論

12 人の被験者に kMedia を利用してもらい，推薦結果（推薦されたページ，推薦されたフォルダ，推薦した相手）に対してそれぞれ評価を行ってもらった．

kMedia ではページ推薦とフォルダ推薦を行うが，ページ推薦は良い結果を得られなかった．対してフォルダ推薦は，ページ推薦に用いた関連性発見アルゴリズムに大きく依存する仕組みでありながら，良い結果を得られた．これは，ページよりも概念的に抽象度の高いフォルダが，被推薦者にとって受け入れられ易いためであると考えられる．

次に，人の繋がりを示す要素について検証するため，被験者による推薦者への評価と相関の高い要素はどのようなものかを調べた．結果，システムが計算した値（関連度や推薦数）よりも，人が評価した値（推薦ページ・フォルダへの平均評価値）の方が高い相関を示した．

この結果から，人の繋がりを示すものとしてページやフォルダに対してシステムが計算した値をそのまま使うのは適切でないと判断した．そこで実社会におけるコミュニティの有無がどのような影響を与えているかを調べることによって，適切な人の繋がりを示す指標を探した．推薦数や関連度に対するコミュニティの有無の影響を調べた結果，人の繋がりを示す要素として，共有している情報の量や質よりもむしろ，情報をどのように分類しているかという点が重要であることがわかった．実際の人の評価との相関を調べた結果，この指標は高い相関を示した．

7. 関連研究

ブックマークは人間によってその情報が有益であるかどうかのフィルタリングがかかった利用価値の高い情報である．この情報を元に，効果的な情報推薦を行おうとする研究がある．

PowerBookmarks[Wen 99] は，ユーザのコメントが記入されているブックマークデータをデータベースで管理し，情報の推薦を行う．WebTagger[Keller 00] は，タグ付けを行ったブックマークを用いて，情報共有を効率よく行おうとしている．BookmarkAgent[森 00] は，ユーザのブックマーク情報を持つエージェントが，ユーザが現在見ている Web ページに対して情報推薦を行う．Siteseer[James 97] は，ブックマークのフォルダ構造を利用し，フォルダ内の登録された Web ページの比較によって推薦を行う．

これらのシステムでは，ブックマークに登録された Web ページや，ページに対するコメント情報のみに注目している．本システムとは，ページが入っているフォルダ，ページを登録した人（の分類の仕方）といった，周りの状況に注目している点で異なっている．

8. 今後の展望

WWW の拡大に伴い，ユーザの興味を反映したフィルタリングを可能とする WWW 検索システムの必要性が高まっている．このとき，実世界と同じように，同じ関心を持つ人間同士で情報を交換・共有することが可能ならば，効率的に情報収集することが可能である．

より良い情報を得るために，情報共有を行うユーザ数は多い方が良い．だが，ユーザが多くなると，共有された情報から選択するのが困難になる．特に Web ページのように分類が困難である場合，それは顕著に表れる．ユーザ間での情報共有を効率よく行うためには，まず情報源をある程度分別する必要がある．

本研究では，ブックマークという，ユーザによって作られた階層構造を持った知識を利用する事によって，自動的に話題の繋がりおよび人の繋がりを発見する手法を提案した．膨大な数のユーザによる情報共有が行われた場合，この手法を用いることにより，適切な情報推薦を支援することができると考えられる．

本手法は情報推薦を支援するが，推薦された情報はユーザにとって未知の情報であり，その情報がどの様な繋がりによって自分に推薦されたのかが適切に示されなければ，ユーザはその情報を理解することが出来ない．これでは良い推薦情報を発見したとしても，情報の推薦が適切に行われたことにはならない．そこで今後は，発見された推薦情報をどのように提示すればユーザが負担無く受け入れられるか，という点について考察を深めたい．

参考文献

[Takeda 00] Hideaki Takeda, Takeshi Matsuzuka, Yuichiro Taniguchi: Discovery of shared topics networks among people — a simple approach to find community knowledge from www bookmarks. Proceedings of the PRICAI 00 (2000)

[松本 99] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸: 日本語形態素解析システム『茶筅』version 2.2.0 使用説明書. NAIST Technical Report, NAIST-IS-TR99012 (1999)

[片岡 97] 片岡充照, 今中武, 水谷研治, 若見昇. テキスト情報を対象としたキーワード抽出と関連情報収集システム. 日本ファジイ学会誌, Vol.9, No.5, pp.710-717 (1997)

[Wen 99] Wen-Syan Li, Quoc Vu, Divakant Agrawal, Yoshinori Hara, and Hajime Takano: PowerBookmarks: A System for Personalizable Web Information Organization, Sharing, and Management. In Proceedings of 8th International World Wide Web Conference(WWW8), pp297-311, 1999

[Keller 00] Keller, H., Selman, B., and Shah, M. A Bookmarking Service for Organizing and Sharing URLs. Computer Network and ISDN Systems, Vol.29, pp.1103-1114 (2000)

[森 00] 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, Vol.J83-D-I, No.5, pp.487-494 (2000)

[James 97] James Rucker and Marcos J. Polanco. Siteseer: Personalized navigation for the web. Communications of the ACM, Vol.40, No.3, pp.73-75 (1997)