kMedia: ユーザ間の共通話題ネットワークの発見

松塚健, 谷口雄一郎, 武田英明

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

あ ら ま し　情報収集やコミュニティ形成には、人どうしの繋がりを知ることが重要である。kMedia は、各人の興味ある話題の関連性が視覚的に確認できる、共通話題ネットワークを発見するシステムである。共通話題ネットワークとは、ブラウザのブックマークフォルダを各人が興味を持つ話題とし、類似するフォルダを無向グラフで表示したものである。フォルダの類似はフォルダの中にある web ページの類似から求め、web ページの類似は web ページから抽出したキーワードから求める。kMedia はクライアント・サーバ型のシステムで、共通話題ネットワークや web ページ間の類似による web ページの推薦情報をグラフィカルに表示する。

キーワード　情報統合, 内容に基づくフィルタリング, 協調フィルタリング, ネットワークコミュニティ, WWW

kMedia: Discovery of Shared Topic Networks among Users

Takeshi Matsuzuka, Yuichiro Taniguchi, Hideaki Takeda

Graduate School of Infomation Science, Nara Institute of Science and Technology

8916-5 Takayama, Ikoma-City, Nara 630-0101

Abstract　We propose a technique to discover shared topics network (STN) among users that can show common interest for two or more persons, and describe kMedia system based on this technique. We interpret folders in browser's bookmark as topics, and STN is realized as the undirected graph of related folders. The relationship among folders is computed by the similarity among web pages included in folders. The similarity among web pages is computed by measuring shared keywords including these web pages. kMedia is a server-client system to show the STN and recommend information computed from page similarity graphically.

Key words　information integration, content-based filtering, collaborative filtering, network community, WWW

1　はじめに

近年の WWW をはじめとす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ことが可能となった。しかし、行き交う情報が増えたことによって、自分の欲しい情報を見失ってしまうという事態が度々起こっている。

このような状況を打破するために、情報収集のみでなく分類・抽出など情報の利用まで含めて考える、いわゆる知的情報統合の必要性が指摘されている [1] 。この情報統合を進めていくための手法のひとつとして、人どうしの繋がりを発見するという手法が考えられている。

われわれは日常、人とのコミュニケーションを通して情報を獲得したり、人の情報分類知識を用いて情報の整理を行っている。人どうしの繋がりを発見することで、これらの活動を実社会およびネットワーク社会において活発化させようという試みがこの手法の目的である。また、このことによりコミュニティの形成を支援していくという側面もある 。

本論文では、人が興味をもつ話題の共通性によって人どうしの繋がりを発見するシステム kMedia を紹介する。kMedia はブラウザのブックマークデータを用いて、人が興味をもつ話題の共通性を共通話題ネットワークとして提示する。また、ブックマークしている web ページで類似性のあるものを、互いに推薦情報として提示する機能も有している。

2　kMedia の概要

kMedia のシステム構成図を図 1 に示す。kMedia はクライアント・サーバ型のシステム構成をとっていて、クライアント部は Java2、サーバ部は CGI+Perl5.0 で作成されている。

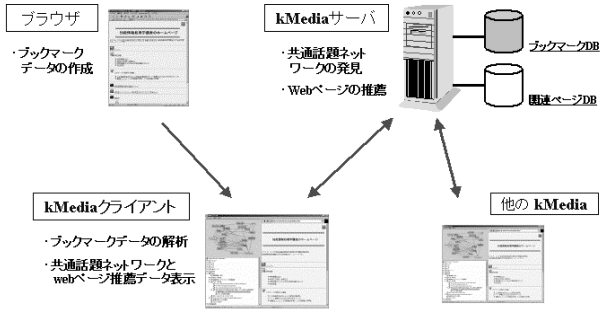

図 1: kMedia 構成図

利用方法は、クライアントを起動させて、設定画面からユーザ名と自分のブックマークファイルのある場所、サーバのある場所を設定して処理を実行させる。すると、クライアントはブックマークファイルの解析と処理を行い、サーバにその結果を送信する。そしてサーバでの処理が終わり、計算結果が返送されると、クライアントのユーザインターフェイスに共通話題ネットワークと web ページの推薦情報がそれぞれ表示される。

なお、ブックマークデータは Netscape のものを用いているが、InternetExplorer のブックマークデータでも InternetExplorer から Netscape 形式にブックマークデータを変換しておくことで利用可能である。

3　kMedia のユーザインターフェイス

kMedia のユーザインターフェイス（図 2 ）は、主に画面左上の共通話題ウィンドウ、画面左下のブックマークウィンドウ、画面右部のブラウザウィンドウ、画面下側のステータスウィンドウの 4 つのウィンドウから構成されている。

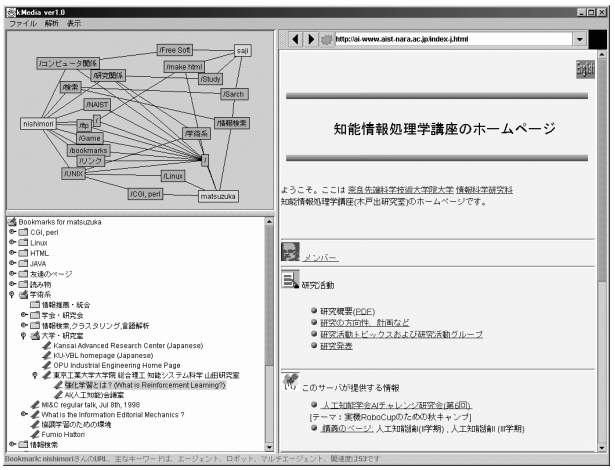

図 2: kMedia のユーザインターフェイス

共通話題ウィンドウ（図 3）には共通話題ネットワークが表示される。共通話題ネットワークとは、それぞれのユーザが興味を持つ話題のうち、互いに共通性のあるものを無向グラフ表示したものであり、ユーザを示す黄色のノード（この図では薄い灰色）と話題を示す青色のノード（この図では濃い灰色）によって、それぞれのユーザの話題の共通性が示されている。

例として、図 3には 3 人のユーザの実際のブックマークデータから kMedia が生成した共通話題ネットワークが表示されているが、このウィンドウからユーザ nishimori の話題「研究関係」に対してユーザ saji の話題「study」とユーザ matsuzuka の話題「学術系」に共通性があることや、ユーザ nishimori,saji,matsuzuka の話題「検索」,「serch」,「情報検索」に互いに共通性があることが見て取れる。

ブックマークウィンドウ（図 4）には、自分のブックマークデータに、他のユーザのデータで類似したものが

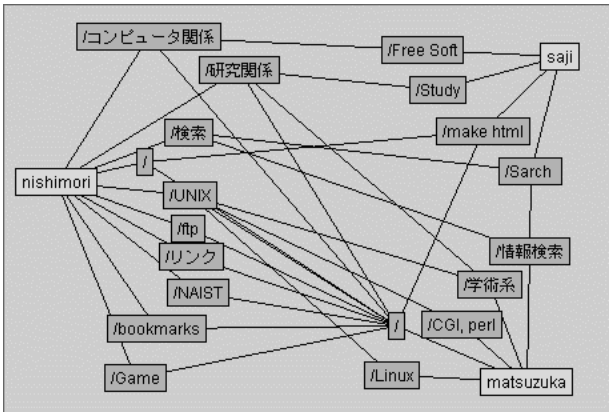

図 3: 共通話題ウィンドウ

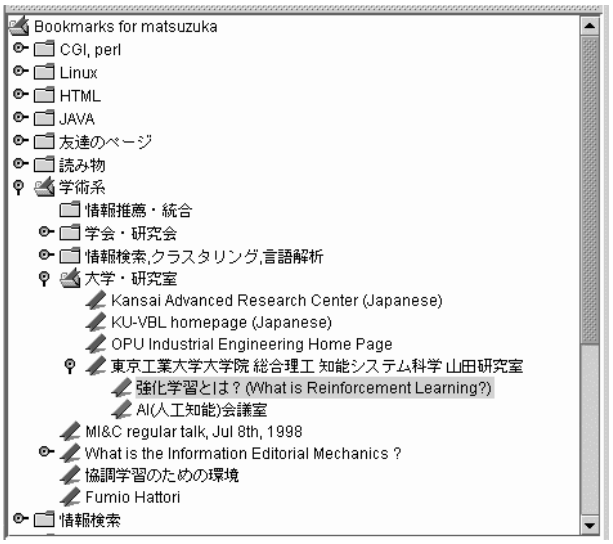

図 4: ブックマークウィンドウ

推薦情報として付加されて表示される。自分のブックマークデータはフォルダがフォルダアイコンで、ブックマークした web ページは緑色のファイルアイコンで表示される。そしてこのデータの下に推薦された web ページが赤色のファイルアイコンで表示される。

例として、図 4では、16 行目の「東京工業大学...」というタイトルの web ページの下に、17 行目の「強化学習とは...」と 18 行目の「AI（人工知能）会議室」というタイトルの web ページが推薦情報として付加されている。

ブラウザウィンドウ（図 5）は簡易なブラウザの役割を果たす。ブックマークウィンドウのブックマークデータをクリックすることでブラウジングが可能である。また、通常のブラウザ同様ウィンドウ上部の URL 表示部に URL を入力することで、直接 web ページを閲覧することも可能である。

ステータスウィンドウ（図 6）には今現在のシステムの

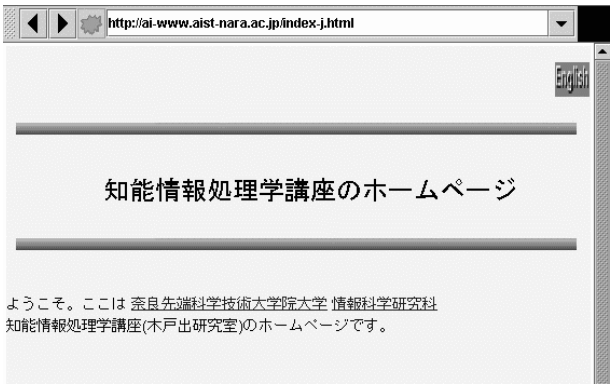

図 5: ブラウザウィンドウ

図 6: ステータスウィンドウ

状況が表示される。また、共通話題ウィンドウやブックマークウィンドウのデータににポインタをあわせると、そのデータの詳細データ（データの所有者や一致したキーワード等）が表示される。

例として、図 6は図 4の「強化学習とは...」にポインタを合わせたときに表示されたものである。ここから、このブックマークデータはユーザ nishimori のデータからの推薦情報で、このページと推薦元のページは「エージェント」,「ロボット」,「マルチエージェント」というキーワードで一致しており、これらのページ間の関連度は 53 であるということが読み取れる。

4 共通話題ネットワークの発見

共通話題ネットワークとは、ユーザが興味を持つ話題で互いに共通性のあるものを無向グラフ表示したものである。しかし、ここで重要なことは、話題の共通性は単純に話題名の比較によって求めることはできないということである。ある情報が与えられたとき、その情報がどういう話題に分類されるかは基本的に各人の判断に委ねられている。そのため、話題の共通性はその話題に含まれる情報までを調べてはじめて判断できる。以下、この共通話題ネットワークの発見手法について述べる。

4.1 共通話題ネットワークの発見手法

共通話題ネットワークを発見するために、kMedia は大きく分けて次の二つの手順を踏んでいる。

1. ユーザが興味を持っている話題を知る
2. ユーザどうしの話題の共通性を調べる

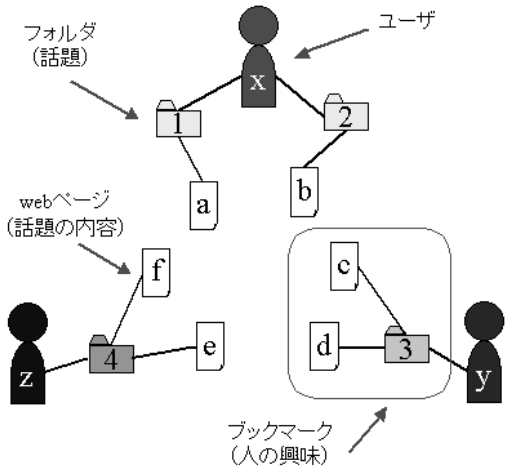

図 7: 話題とブックマークの関係

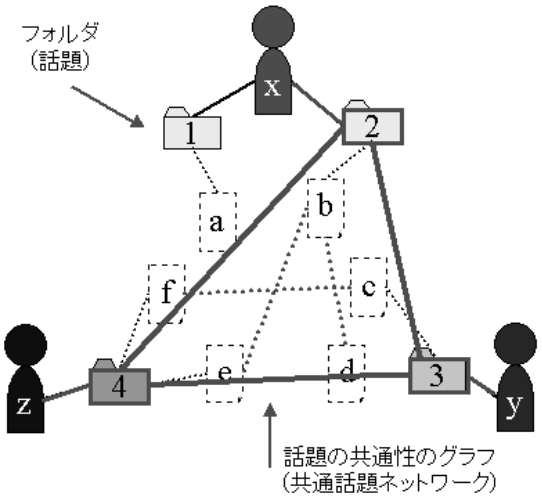

図 9: 発見された話題の共通性

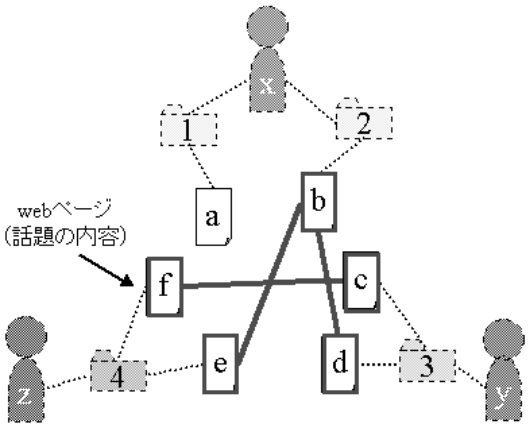

図 8: web ページの類似例

要素	kMedia での定義
話題	ブックマークのフォルダ
話題に関する情報	フォルダ中の web ページ
話題の共通性	フォルダ中の web ページの類似性
web ページの類似性	web ページ中のキーワードの一致

表 1: 各要素の kMedia での定義

まず、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、kMedia はブックマークのフォルダを利用している。ブックマークのフォルダとはブラウザのブックマークデータを整理するために用意されているものであり、ユーザは任意にフォルダを作って、その中にはブックマークした web ページやさらにフォルダを入れることができる。

この各フォルダにはユーザが任意の名前を付けることができるのであるが、そのフォルダ名には一般的に、その中にある web ページの話題は何であるかが記述されている。そこで、kMedia ではブックマークのフォルダに付けられた名前は、その中にある web ページの話題を表しているとしている。また、フォルダ中にある web ページは、その話題に関する情報であるとしている。これらの関係を図 7に示す。

ユーザどうしの話題の共通性は、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まると考えられる。そこで kMedia では、話題の共通性はブックマークフォルダに含まれる情報、すなわちフォルダ中に含まれる web ページの類似性から求めている。つまり、web ページの類似性が図 8のように示された場合、話題の共通性は図 9のように示される。

さらに、その 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ている。web ページ中の出現頻度の多い単語の上位いくつかをそのページのキーワードとし、互いのページのキーワードが一致すればするほど、そのページ間には類似性があるとしている。

また、ここで求められた web ページの類似性を元に、kMedia はブックマークデータの推薦を行っている。すなわち、各ユーザが持っているブックマークの web ページで、キーワードの一致する web ページを互いに推薦情報として提示しあう。

以上で定義した各要素の kMedia での定義の一覧を表 1に示す。

4.2 web ページ間の類似

kMedia は次の手順で web ページ間の類似を求めている。

1. web ページから単語切り出し
2. 単語からキーワードを選定
3. キーワードの一致から web ページ間の類似を決定

まず、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 ページを読み込む。そしてそのページの文章から単語を切り出していくのであるが、この単語の切り出しは茶筅 [2] などによる形態素解析でなく、字種に注目した単語切り出し [3] によって行っている。

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見ていって、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。単語とみなす最短文字長は調整できるが、今回は 3 文字以上としている。例として、以下のような文章があったとする。

奈良先端大学のホームページへようこそ！
ここが入口です。

この文章からは「奈良先端大学」,「ホームページ」の二つが単語として切り出される。「入口」は文字長が 2 のため単語として扱われない。このような単語の切り出し手法は、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, 略語などに対して柔軟な切り出しができるため、web の情報のように雑多で柔らかい文章の解析に適している。

そして、ここから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」,「カウント」など、基本的にあまり意味が無くかつ出現頻度の多い単語は、キーワードとして扱うと不具合が生じるため除外している。今回 kMedia では以下の単語をキーワードから除外している。

アクセス, アドレス, アンケート, カウンタ, カウンター, クリック, ゴシック, コンテンツ, サーバ, サイト, サポート, システム, ダウンロード, チャット, ディレクトリ, ドキュメント, バイト, バナー, ファイル, フォント, ブラウザ, フリー, フレーム, ページ, ホームページ, メーリングリスト, メール, リンク, 掲示板

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ている。なお、今回 kMedia ではキーワードが二つ以上一致するとそれらのページ間に類似性があるとしている。

4.3 話題の共通性

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ている。今回 kMedia ではフォルダ中の web ページが 3 ページ以上互いに類似性があると、そのフォルダ間、すなわち話題間に共通性があるとしている。

また、話題の共通性を調べる際、今回はフォルダの階層について 2 階層以降を区別していない。つまり、「学術系」の中にある「大学・研究室」というフォルダの web ページは、「学術系」というフォルダにあるものとして話題の共通性を調べている。なお、これは表示の煩雑さを考慮したための仕様で、若干の変更で 2 階層以降の話題の区別して扱うことが可能である。

5 評価実験

kMedia が提示した話題および web ページの共通性が、ユーザにとっても認められるものであるかどうかを評価実験で確かめた。kMedia を無作為に選んだ被験者 3 人に使ってもらい、kMedia が提示した 3 人の話題および web ページの共通性を被験者によって評価してもらった。被験者は全員、日常的に web ブラウジングを行い、ブックマーク作成に対する知識を持っている人達であった。評価は非常に類似性が認められるものを 5, 全く類似性が認められないものを 1 とする 5 段階評価で、評価項目は推薦された web ページの類似性と kMedia が提示した話題の共通性についての 2 項目である。

実験に使用した各ユーザのブックマークデータは表 2 のようなデータであった。

	ユーザ A	ユーザ B	ユーザ C
全ページ数	376	278	297
解析ページ数	263	185	240
全話題数	13	17	5

表 2: 各ユーザのブックマークデータ

解析ページ数とは、ブックマークの全ページ中、システムがキーワード抽出に成功したページの総数である。kMedia は以下のようなページは解析を行わない。

- ページが現在存在しない（デッドリンク）
- ページが読み込めない（タイムアウト等）
- ページの URL が ftp:で始まるもの
- ページから日本語単語が抽出できない

また、話題数とはブックマークの第 1 階層にあるフォルダの総数である。

kMedia が提示した各被験者の話題の共通性および推薦情報は表 3 のようになった。なお、推薦ページ数はのべ数、すなわち一つのページが複数回推薦されたものを複数回数えている。また、共通話題数とは自分の話題のうち、他のユーザのいずれかの話題と共通性が示されたものの数である。

	ユーザ A	**ユーザ B**	**ユーザ C**
推薦ページ数	345	513	454
共通話題数	10	10	3

表 3: kMedia が提示したデータ

推薦された web ページが、推薦元となる自分の web ページとどれだけ類似性があるか評価してもらった結果は表 4 のようになった。また、示された自分と他のユーザの話題の共通性について、自分の話題との共通性がどれだけ適切であるか評価してもらった結果は表 5 のようになった。

評価	**5**	**4**	**3**	**2**	**1**
ユーザ A	29	30	27	77	182
ユーザ B	94	86	73	185	75
ユーザ C	66	90	88	88	122
合計	189	206	186	350	379

表 4: 推薦されたページの類似性の評価（単位:ページ数）

評価	**5**	**4**	**3**	**2**	**1**
ユーザ A	3	3	2	0	2
ユーザ B	3	3	2	2	0
ユーザ C	2	0	1	0	0
合計	8	6	5	2	2

表 5: 提示された話題の共通性の評価（単位:話題数）

結果、ユーザによる評価は web ページの類似性より抽象度の高い話題の共通性でより良くなっている。これは、個々の情報に関して荒い理解をすることで、情報の類似性という細かい情報は荒いものとなるが、話題の共通性という抽象度の高い情報は人間にとって意味のあるものとなるという kMedia の性質が客観的にも示された結果となっている。

しかし、やはり現時点では個々の情報、すなわち web ページの推薦手法としての価値はさほど高いものではないということも明らかになった。細かい情報の類似性も kMedia で正確に扱うためには、本手法の改良もしくは他の手法の併用を考える必要がある。

6 関連研究

kMedia と関連する研究として、次のような項目を目的とした研究がある。

- 人どうしのネットワークの発見
- ブックマークを用いた情報推薦
- 話題による情報検索、情報要約

これらについて、以下で順に説明する。

6.1 人どうしのネットワークの発見

情報獲得やコミュニティ形成促進のため、人と人の関係、人と話題の関係を発見するということをテーマとした研究がいくつか行われている。

Refferal Web [4] は、論文の書誌データや各ユーザの自分のホームページの記述から人と人の関係を提示するというシステムである。吉田らは、この関係発見に用いるユーザのプロファイルを、ユーザからのフィードバックを用いて学習させることでより正確な関係を導き出そうとしている [5] 。また、CoMeMo-Community [6] では各人が連想表現で表現した知識から人と人の関係を提示している。

これらの研究と kMedia の違いは、Refferal Web や CoMeMo-Community などでは、ユーザプロファイルとなるデータの直接の結びつきからユーザ間の関係を導き出しているのに対し、kMedia は話題間の関係という 1 つ抽象化した関係からユーザ間の関係を求めていることである。また、これらのシステムは直接ユーザに情報を推薦していないが、kMedia はシステムが直接情報推薦を行っているところも異なる。

6.2 ブックマークを用いた情報推薦

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ている、利用価値の高い情報である。この情報を基に有益な情報推薦を行おうとする研究がある。

PowerBookmarks [7] 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、情報の推薦と組織化を行おうとしている。ブックマークエージェント [8] は、ユーザのブックマーク情報を持つエージェントが互いに、ユーザが現在見ている web ページに対する情報推薦を行うシステムである。Siteseer [9] は、同じ web ページを複数のユーザがブックマークしているとき、その web ページが入っているブックマークフォルダ中の情報を互いに推薦するというシステムである。

これらのシステムと kMedia の違いは、これらのシステムはブックマークデータの情報推薦に焦点を当てていて、自分以外のユーザの区別をあまりしていないところである。kMedia は、どのユーザがどのような話題に興味があるかという点に注目してユーザ間の話題の共通性を発見するとともに、web ページの推薦を行っている。

6.3 話題による情報検索、情報要約

kMedia では、その情報がどういった話題を扱っているかという判断はブックマークを作成したユーザに委ねられているが、これを機械的に判断して情報検索や要約に役立てようとする研究がある。

仲川らは、文章を話題（この研究では分類観点と呼ぶ）によって分類しておき、情報検索に役立てようとしている [10]。 また、福原らは、単語出現頻度分布の歪度と尖度を用いて文章の話題特定を行い、その話題から複数テキストの要約を行っている [11]。

7 おわりに

本研究では、人の興味ある話題を特定し人どうしの話題の共通性を発見することで情報の流通と整理がすみやかになると考え、各人の興味ある話題の共通性が視覚的に確認できる共通話題ネットワークを発見するシステム kMedia を開発した。kMedia では

- 話題特定にはブックマークデータを用いる
- 話題の共通性は web ページの類似関係から求める

というアプローチで共通話題ネットワークを発見した。また、共通話題ネットワークを発見する際求める web ページの類似関係から、各ユーザのブックマークデータを互いに推薦させた。

評価実験の結果、kMedia が提示した話題の共通性で一定の評価が得られた。これは、kMedia の有効性が示されたものといえる。

参考文献

[1] 武田 英明. ネットワークを利用した知的情報統合. 人工知能学会誌, Vol.10, No.5, pp.680-688, 1996.

[2] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 浅原 正幸. 日本語形態素解析システム『茶筌』version 2.0 使用説明書 第二版NAIST Technical Report, NAIST-IS-TR99012, December 1999.

[3] 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. 日本ファジィ学会誌, Vol.9, No.5, pp.710-717, 1997.

[4] Henry Kautz, Bart Selman, and Mehul Shah. The hidden web. AI Magazine, Vol.18, No.2, pp.27-36, 1997.

[5] 吉田 仙, 亀井 剛次, 横尾 真, 大黒 毅, 船越 要, 服部 文夫. 潜在的なコミュニティの可視化. 第 6 回マルチ・エージェントと協調計算ワークショップ オンライン予稿集, 1997.

[6] 平田 高志, 村上 晴美, 西田 豊明. 連想表現を用いたコミュニティにおける知識の視覚化とその評価実験. システム制御情報学会論文誌, 12(7), pp.428-436, 1999.

[7] 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), pp.297-311.

[8] 森 幹彦, 山田 誠二. ブックマークエージェント: WWW における協調的情報フィルタリング. 第 12 回人工知能全国大会資料, pp.646-648, 1998.

[9] James Rucker, and Marcos J. Polanco. Personalized Navigation for the Web. Communications of the ACM Vol.40, No.3, pp.73-75, 1997.

[10] 仲川こころ, 高田喜朗, 関浩之. 検索目的を反映したカテゴリ構造に基づく WWW 検索支援. 情報処理学会研究報告, HI82, pp.59-64, 1999.

[11] 福原知宏, 武田英明, 西田豊明. 統計情報を用いた話題特定と文脈の再構築による複数テキスト要約. 人工知能学会全国大会 (第 13 回) 論文集, 1999.